O PROBLEMA DE *CLUSTERING* HETEROGÊNEO

Éverton Santi
Escola de Ciências e Tecnologia
Universidade Federal do Rio Grande do Norte, Natal-RN, Brasil, 59072-970
esanti@ect.ufrn.br

Daniel Aloise
Departamento de Engenharia de Computação e Automação
Universidade Federal do Rio Grande do Norte, Natal-RN, Brasil, 59072-970
aloise@dca.ufrn.br

Simon J. Blanchard
McDonough School of Business
Georgetown University, Washington, DC 20057, Estados Unidos da América
simon.blanchard@georgetown.edu

RESUMO

Este trabalho apresenta a definição do Problema de Clustering Heterogêneo (PCH). Diferentemente de modelos de *clustering* tradicionais, este novo elimina a necessidade de que haja a agregação das matrizes de dissimilaridades quando diferentes indivíduos estão envolvidos no processo de julgamento dos objetos a serem particionados. Adicionalmente, este modelo permite identificar segmentos de indivíduos que compartilham opiniões similares, relacionando-se uma estrutura de categorias a cada um destes segmentos. São apresentadas também duas formulações alternativas para o problema, considerando-se sua convexificação. A partir destas formulações, limitantes inferiores e soluções ótimas são mostradas, considerando-se um conjunto de instâncias geradas a partir da literatura.

PALAVRAS CHAVE. Clustering. Heterogeneidade. Programação Inteira Mista.

Área Principal: Programação Matemática. Otimização Combinatória. Outras Aplicações em P. O.

ABSTRACT

This work presents the Heterogeneous Clustering Problem (HCP). Differently from the classical clustering models, this new model eliminates the need for aggregating the dissimilarities matrices when several individuals are considered in the judgment of a set of objects to be partitioned. Additionally, this model allow us to identify segments of individuals who share similar opinions, relating a category structure to each one of these segments. Two alternative formulations are given for the problem, considering its convexification. From these formulations, lower bounds and optimal solutions were obtained for a set of instances generated with basis on the literture.

KEYWORDS. Clustering. Heterogeneity. Mixed Integer Programming.

Main Area: Mathematical Programming. Combinatorial Optimization. Other Applications in O.R.

1. Introdução

Clustering, ou análise por *clusters*, pode ser entendida como a tarefa de agrupar um conjunto de objetos de forma que os objetos mais similares deste conjunto sejam colocados em um mesmo grupo, assim como objetos diferentes sejam colocados em *clusters* distintos. Do ponto de vista *de cima para baixo* ou *top-down*, considera-se que *clustering* consiste na segmentação de uma população heterogênea em subgrupos menos heterogêneos (Aldenderfer e Blashfield, 1984). Do ponto de vista *de baixo para cima* ou *bottom-up*, define-se *clustering* como a tarefa de encontrar grupos em um conjunto de dados considerando algum critério de similaridade (Duda e Hart, 1973).

Em um contexto prático, considerando-se a aplicação de métodos de *clustering*, pode-se objetivar a identificação de categorias ocultas (ou não-observadas) em relação a um conjunto de objetos reais ou mesmo artificiais. Busca-se, a partir disto, a obtenção de informações relevantes ao tomador de decisão, seja este de qualquer área de atuação (Blanchard e DeSarbo, 2013; Blanchard *et al.*, 2012). Aplicações de métodos de *clustering* podem ser vistas em ciências naturais, engenharia, psicologia, medicina, marketing e economia, como por exemplo, categorização de alimentos (Kohn *et al.*, 2010; Ross e Murphy, 1999), animais (Kelter *et al.*, 1977), elementos léxicos (Miller, 1969), bens duráveis (Urban *et al.*, 1993), bens de consumo (Griffin e Hauser, 1993), frases (Perkins, 1993), dentre outros.

No entanto, ao se utilizar métodos de *clustering* tradicionais, como por exemplo o k-means (Lloyd, 1957), assume-se que a percepção das categorias não-observadas é homogênea. Isto se deve ao fato de que o principal dado de entrada para estes métodos consiste em uma única matriz de dissimilaridades (ou similaridades) entre os objetos considerados (Daws, 1996). Esta abordagem pode não ser apropriada em certos casos, pois na realidade diferentes indivíduos podem julgar um mesmo conjunto de objetos com base em percepções diferenciadas, podendo gerar arranjos distintos entre estes objetos (Blanchard *et al.*, 2012). Isto ocorre em razão de que muitos fatores podem influenciar os indivíduos no momento de seu julgamento, como por exemplo, seu nível de conhecimento, sua idade, a finalidade para a qual o modelo está sendo empregado e até mesmo seu humor (Isen, 2012; Ross e Murphy, 1999; John e Sujan, 1990; Sujan e Dekleva, 1987; Medin e Schaffer, 1978).

De forma a exemplificar tal situação, considera-se o caso em que dois indivíduos devem categorizar três marcas de chocolates (Twix, KitKat e Snickers) de acordo com o critério que acharem mais adequado por meio de um Processo de Triagem (*Sorting Task*). Um possível resultado para tal categorização é mostrado na Figura 1. Para este resultado, pode-se supor que o Indivíduo 1, por exemplo, classificou os chocolates Twix e KitKat na mesma categoria por considerar que ambos são crocantes. O chocolate Snickers, por sua vez, está em uma categoria diferente, por este indivíduo considerar que este chocolate não é crocante.

O Indivíduo 2, diferentemente, determinou que os chocolates Twix e Snickers podem estar em uma mesma categoria por considerar que ambos possuem recheio de caramelo. O chocolate KitKat, no entanto, foi atribuído a outra categoria em razão de que este segundo indivíduo considera que este chocolate contém *wafer* em seu interior. Caso um método de *clustering* tradicional (k-means ou p-medianas, por exemplo) fosse empregado para obter quais categorias poderiam representar melhor a opinião dos dois indivíduos de forma conjunta, primeiramente uma matriz de dissimilaridades deveria ser calculada para cada uma das classificações obtidas. Ou seja, obtém-se uma matriz de dissimilaridades para representar o julgamento de cada indivíduo separadamente. Após este processo, torna-se necessária a agregação destas matrizes, para que então se possa utilizar tal método.

Para casos como o descrito neste texto, o processo de agregação tende a gerar a perda da informação contida nos dados, uma vez que torna-se inviável observar diferentes estruturas de categorias entre os objetos considerados, situação que de fato pode ocorrer. Adicionalmente, relacionar estas categorias a diferentes perfis (ou grupos) de indivíduos pode vir a ser uma tarefa não trivial. Neste sentido, apresenta-se neste trabalho um novo modelo de *clustering*, o Problema

Figura 1: Exemplo de categorização heterogênea

de *Clustering* Heterogêneo (PCH), no qual elimina-se a necessidade de agregação das matrizes de dissimilaridades (ou similaridades).

Este modelo permite identificar de forma direta diferentes grupos de indivíduos. Para cada um destes grupos, por sua vez, uma estrutura de categorias entre os objetos é recuperada, permitindo-se observar de que forma estes indivíduos percebem as relações entre os diferentes objetos. Adicionalmente, apresentam-se duas formulações convexas para este novo problema, a partir das quais limitantes inferiores foram obtidos para um conjunto de instâncias geradas com base na literatura. Estes, por sua vez, poderão auxiliar na análise de resultados a serem obtidos em estudos futuros, considerando-se o desenvolvimento de metaheurísticas para o PCH.

O restante deste texto está organizado como segue: a Seção 2 apresenta a formalização do PCH. A Seção 3 apresenta duas formulações alternativas para este problema. Considera-se nesta seção a linearização do modelo original, bem como a adição de cortes a estas formulações de forma a viabilizar a obtenção de limitantes inferiores para o problema original, sem afetar sua solução ótima. A Seção 4 apresenta resultados computacionais para ambas as formulações, considerando-se dados gerados a partir de um estudo baseado em Simulação de Monte Carlo, descrito em Blanchard *et al.* (2012). Por fim, a Seção 5 apresenta as considerações finais para este trabalho.

2. O Problema de *Clustering* Heterogêneo

De forma a definir o PCH, considera-se que m indivíduos devem classificar n objetos, obtendo-se uma matriz de dissimilaridades $D^k = (d^k_{ij})$ para cada um destes k indivíduos. Isto é, para todo $k = 1, \ldots, m$. Para cada uma destas matrizes, $d^k_{ij} \in [0, 1]$ representa o grau de dissimilaridade entre cada par de objetos i e j, segundo a opinião do indivíduo k. Considera-se também que c^k, para todo $k = 1, \ldots, m$, representa o número de categorias que o indivíduo k considera apropriado para classificar todos os n objetos. O método utilizado para a aquisição destes dados, considerando-se a definição do PCH, poderá ser escolhido pelo usuário do modelo, não havendo restrições.

O Problema de *Clustering* Heterogêneo, a partir dos parâmetros descritos, consiste em identificar $G(g = 1, \ldots, G)$ *segmentos*, dentre os quais os m indivíduos serão designados. Busca-se alocar em um mesmo segmento g aqueles indivíduos cujas opiniões são similares em relação aos objetos considerados durante o processo de classificação. Adicionalmente, para cada segmento g, busca-se identificar uma estrutura de categorias que melhor represente a percepção destes indivíduos em relação à classificação dos objetos e suas relações entre si. Para o problema descrito, G é um parâmetro que deverá ser informado pelo usuário do modelo.

A categorização dos objetos, por sua vez, é realizada considerando-se o conceito de mediana. Isto significa que, ao se definir um conjunto de partições entre os n objetos, o centro de cada partição será um de seus próprios elementos, e não um elemento artificial (centroide). No contexto do PCH, este objeto é aquele que melhor representa a categoria na qual está inserido. A utilização de medianas como centros de *cluster* pode ser vantajosa em certos tipos de aplicação, como por exemplo, a segmentação de consumidores, permitindo-se uma melhor representação de marcas e

produtos.

Formalmente, o Problema de *Clustering* Heterogêneo é definido como:

$$\min \sum_{k=1}^{m} \sum_{g=1}^{G} z^{kg} \left[\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^{k} e_{ij}^{g} \right] \tag{1}$$

sujeito a

$$\sum_{j=1}^{n} e_{ij}^{g} = 1, \ \ \forall g = 1, \ldots, G, \ i = 1, \ldots, n \tag{2}$$

$$e_{ij}^{g} \leq e_{jj}^{g}, \ \ \forall g = 1, \ldots, G, \ i, j = 1, \ldots, n \tag{3}$$

$$\sum_{g=1}^{G} z^{kg} = 1, \ \ \forall k = 1, \ldots, m \tag{4}$$

$$\sum_{k=1}^{m} z^{kg} \geq 1, \ \ \forall g = 1, \ldots, G \tag{5}$$

$$\sum_{j=1}^{n} e_{jj}^{g} = \left\lfloor \frac{\sum_{k=1}^{m} c^{k} z^{kg}}{\sum_{k=1}^{m} z^{kg}} \right\rfloor, \ \ \forall g = 1, \ldots, G \tag{6}$$

$$e_{ij}^{g} \in \{0, 1\}, \ \ \forall g = 1, \ldots, G, \ i, j = 1, \ldots, n \tag{7}$$

$$z^{kg} \in \{0, 1\}, \ \ \forall g = 1, \ldots, G, \ k = 1, \ldots, m \tag{8}$$

O modelo (1-8) tem como variáveis de decisão z^{kg} e e_{ij}^{g}. As variáveis z^{kg} representam a designação de cada um dos indivíduos em relação aos G segmentos, ou seja, z^{kg} deve ser igual a 1 se o indivíduo k é alocado ao segmento g, e 0 caso-contrário. As variáveis e_{ij}^{g}, por sua vez, devem ter valor igual a 1 se o objeto i está associado ao objeto j dentro do segmento g, e 0 caso-contrário. Neste contexto, quando um objeto i é associado a um objeto j em um segmento g, o objeto j será, obrigatoriamente, uma mediana deste segmento.

A função objetivo em (1) minimiza a soma das dissimilaridades entre cada objeto e a mediana da categoria à qual este objeto está sendo associado, condicional à pertinência dos indivíduos em relação aos segmentos. As restrições em (2) e (3) garantem que cada objeto i será associado a exatamente uma mediana em cada segmento g, sendo que esta associação só será válida se j for uma mediana dentro do segmento em questão. As restrições em (4) impõem que cada indivíduo deverá ser associado a exatamente um segmento, enquanto que as restrições em (5) garantirão que nenhum segmento poderá ser vazio.

As restrições em (6) impõem que o número de medianas para cada segmento g deverá ser igual ao piso do número médio de medianas esperado pelos indivíduos daquele segmento. Esta definição busca se adequar a definições de autores da literatura comportamental, que sugerem que os indivíduos tendem a seguir regras simples quando buscam exprimir preferências e percepções (Bettman *et al.*, 1998; Bettman e Park, 1980; Shugan, 1980; Simon, 1955). Se os julgamentos de diferentes indivíduos forem similares em termos de estruturas de categorias, então admite-se que estas categorias serão similares não só em termos de composição, mas também em termos de quantidade. Por fim, tem-se as restrições de integralidade sobre as variáveis e_{ij}^{g} e z^{kg} em (7) e (8), respectivamente.

Para as restrições em (6), ainda, o processo de otimização irá garantir que $\sum_{j=1}^{n} e_{jj}^{g}$ seja sempre igual ao maior valor possível (viável), pois quanto maior o número de medianas em um segmento menor será o custo da função objetivo, uma vez que se considere que o custo em se

associar uma mediana a ela mesmo é zero. Desta forma, pode-se substituir o conjunto de restrições dado em (6) pelas seguintes desigualdades:

$$\sum_{j=1}^{n} e_{jj}^{g} \leq \frac{\sum_{k=1}^{m} c^k z^{kg}}{\sum_{k=1}^{m} z^{kg}}, \forall g = 1, \ldots, G, \tag{9}$$

sem que a solução ótima seja afetada. Caso se faça necessário, adicionalmente, estas restrições podem ser modificadas de forma a se adequar à preferência do usuário. Pode-se, por exemplo, considerar que o número de medianas para cada um dos segmentos g seja igual ao inteiro mais próximo de $\frac{\sum_{k=1}^{m} c^k z^{kg}}{\sum_{k=1}^{m} z^{kg}}$, ao invés do piso deste valor. Para tal, basta que se adicione 0.5 ao lado direito da restrição em (9).

Para maior clareza e compreensão do texto, destaca-se que o termo *segmentos* será sempre empregado com referência às partições estabelecidas entre os indivíduos, enquanto o termo *categoria*, ou *estrutura de categorias*, será empregado empregado com referência às partições estabelecidas entre os objetos, considerando-se que a cada segmento g está associada uma estrutura de categorias (e^g).

3. Formulações alternativas para o Problema de *Clustering* Heterogêneo

O PCH permite que se considere a heterogeneidade presente nos dados, pois não há necessidade de agregação das matrizes de distâncias a serem consideradas como entrada. No entanto, o modelo dado em (1-8) é um modelo de programação não-linear inteira mista, sendo de complexa resolução mesmo para instâncias pequenas. Testes computacionais foram realizados a fim de se verificar tal afirmação. Utilizou-se, desta forma, o *solver* Couenne (Couenne, 2015), instalado em um computador com 62GB de memória e 12 processadores Intel Xeon(R) CPU X5650 de 2.67GHz. Para um limite de 24 horas de execução, mostrou-se impraticável a obtenção de limitantes inferiores para o problema, mesmo considerando-se uma instância gerada de forma aleatória com $m = 5$ indivíduos, $n = 5$ objetos e $G = 2$ segmentos.

A partir destes resultados, buscou-se reformular o problema para uma forma convexa, visando-se não só obter limitantes inferiores, como também soluções ótimas para algumas instâncias. Como resultado, obtiveram-se duas formulações alternativas. A primeira delas, a PCH1 baseia-se no método do Grande M e nas desigualdades de Fortet (Fortet, 1960; Fortet, 1959), linearizando-se função objetivo e restrições. A segunda formulação, a PHC2, por sua vez, baseia-se também no processo de convexificação por meio das desigualdades de Fortet (Fortet, 1959; Fortet, 1960), porém de forma a permitir a adição de cortes à formulação. A formulação PCH1 e PCH2 são apresentadas nas subseções que seguem, bem como uma análise acerca de suas limitações.

3.1. A formulação PCH1

Considerando-se o método do Grande M, pode-se definir que custo de associar um objeto i a outro objeto j em um segmento g é dado pela soma das dissimilaridades atribuídas a este par de objetos por todos os indivíduos k que estão alocados neste segmento. Logo, o custo de tal ligação pode ser atribuído a uma variável $c_{ij}^{g} \in [0, m - G + 1]$. Para tal, adiciona-se ao modelo (1–8) o conjunto de restrições

$$\sum_{k=1}^{m} z^{kg} d_{ij}^{k} \leq c_{ij}^{g} + (1 - e_{ij}^{g})(m - G + 1), \;\; \forall\, g = 1\ldots., G, \quad i, j = 1, \ldots, n, \tag{10}$$

Desta forma, quando a variável e_{ij}^{g} assumir valor 1, isto é, o objeto i for associado ao objeto j dentro do segmento g, o segundo termo do lado direito desta desigualdade será anulado. Esta condição implica que para que a solução do problema seja viável, c_{ij}^{g} deverá ser igual à soma das dissimilaridades relacionada a este par de objetos, considerando-se os valores de d_{ij}^{k} para todos os indivíduos k alocados ao segmento g. Caso contrário, quando e_{ij}^{g} tem valor 0, o segundo termo

desta desigualdade corresponderá a um valor suficientemente grande (a saber, $m - G + 1$), permitindo ao *solver* atribuir um valor zero a c_{ij}^g, dado o processo de minimização. Desta forma, o custo de associar um objeto i a outro objeto j dentro do segmento g será nulo.

A partir da introdução desta nova variável ao modelo original do problema, podemos modificar a sua função objetivo dada em (1) para:

$$\min \sum_{g=1}^{G} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^g, \tag{11}$$

eliminando-se, portanto, o produto das variáveis binárias z e e.

Apesar da modificação da função objetivo, o modelo continua não-linear, dado o conjunto de restrições definidas em (9). De forma a tornar linear este conjunto de restrições, pode-se passar o denominador do lado direito para o lado esquerdo desta desigualdade, obtendo-se então um novo conjunto de restrições dado por:

$$\sum_{k=1}^{m} \sum_{j=1}^{n} z^{kg} e_{jj}^g \leq \sum_{k=1}^{m} c^k z^{kg}, \ \ \forall\, g = 1, \ldots, G. \tag{12}$$

Obtém-se, novamente, o produto das variáveis binárias z e e, para todo $i = j$. No entanto, o produto $z^{kg} \times e_{jj}^g$ somente terá valor igual a 1 se ambas variáveis tiverem valor 1. Caso contrário, o valor deste produto será zero. Logo, pode-se aplicar técnica de reformulação dada em Fortet (1960, 1959). Para tal, substitui-se $e_{jj}^g \times z^{kg}$ por w_j^{kg} ($w_j^{kg} \in [0, 1]$), devendo-se adicionar ao modelo três conjuntos de restrições que garantam que $\max\{0, z^{kg} + e_{jj}^g - 1\} \leq w_j^{kg}$. São eles:

$$w_j^{kg} \leq e_{jj}^g, \ \ \forall\, g = 1, \ldots, G, \quad k = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{13}$$

$$w_j^{kg} \leq z^{kg}, \ \ \forall\, g = 1, \ldots, G, \quad k = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{14}$$

$$w_j^{kg} \geq z^{kg} + e_{jj}^g - 1, \ \ \forall\, g = 1, \ldots, G, \quad k = 1, \ldots, m, \quad j = 1, \ldots, n. \tag{15}$$

Pode-se, a partir da adição das variáveis w ao modelo e do novo conjunto de restrições (13-15), reescrever as restrições em (12) como:

$$\sum_{k=1}^{m} \sum_{j=1}^{n} w_j^{kg} \leq \sum_{k=1}^{m} c^k z^{kg}, \ \ \forall\, g = 1, \ldots, G. \tag{16}$$

Embora se tenha obtido uma formulação linear a partir das manipulações apresentadas, destaca-se que todas as variáveis de decisão e_{ij}^g deverão possuir valor integral na solução ótima do problema, a fim de satisfazer as restrições de integralidade em (7). Porém, o processo de otimização sempre buscará definir valores fracionários a estas variáveis, uma vez que estes implicam em custos menores, ou mesmo nulos, para cada associação de objetos i e j em todos os segmentos g. Isto é, as restrições dadas por (10) poderão diminuir consideravelmente o desempenho de algoritmos de *branch-and-bound* na solução destes problema, dado o número total de nós de ramificação que serão necessários para o algoritmo gerar uma solução que satisfaça a (6) para todas estas variáveis.

3.2. A formulação PCH2

De forma a obter uma segunda formulação convexa para o PCH, definido em (1-8), pode-se substituir todos os produtos $z^{kg} \times e_{ij}^g$ por variáveis w_{ij}^{kg}, de forma similar à utilizada para a formulação PCH1. Considerando-se novamente o método das Desigualdades de Fortet (Fortet, 1960; Fortet, 1959), tem-se que o produto $e_{ij}^g \times z^{kg}$ somente terá valor igual a 1 se ambas variáveis tiverem valor 1. Caso contrário, o valor deste produto será zero. Para tal, garante-se que $w_{ij}^{kg} =$

$z^{kg} \times e^g_{ij}$ (com $w^{kg}_{ij} \in [0,1]$) por meio da adição de três conjuntos de restrições que garantam que $\max\{0, z^{kg} + e^g_{ij} - 1\} \leq w^{kg}_{ij}$. São elas:

$$w^{kg}_{ij} \leq e^g_{ij}, \forall g = 1, \ldots, G,\ k = 1, \ldots, m,\ i, j = 1, \ldots, n, \tag{17}$$

$$w^{kg}_{ij} \leq z^{kg}, \forall g = 1, \ldots, G,\ k = 1, \ldots, m,\ i, j = 1, \ldots, n, \tag{18}$$

$$w^{kg}_{ij} \geq e^g_{ij} + z^{kg} - 1, \forall g = 1, \ldots, G,\ k = 1, \ldots, m,\ i, j = 1, \ldots, n. \tag{19}$$

A introdução deste novo conjunto de restrições ao modelo, bem como das variáveis w^{kg}_{ij}, permite que se reescreva a função objetivo dada em (1) como:

$$\min \sum_{g=1}^{G} \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} d^k_{ij} w^{kg}_{ij}. \tag{20}$$

Pode-se, também, reescrever o conjunto de restrições em (9), considerando-se passar o denominador do lado direito da desigualdade para o lado esquerdo e substituindo-se os produtos $z^{kg} \times e^{kg}_{jj}$ por w^{kg}_{jj}, como:

$$\sum_{k=1}^{m} \sum_{j=1}^{n} w^{kg}_{jj} \leq \sum_{k=1}^{m} c^k z^{kg},\ \forall g = 1, \ldots, G. \tag{21}$$

Pode-se, ainda, acelerar o processo de otimização desta nova formulação, tornando-a mais robusta por meio da adição de cortes sem que a solução ótima do problema original seja afetada. Baseando-se na Técnica de Reformulação e Linearização (TRL) descrita em Sherali e Desai (2005) e Sherali e Alameddine (1992), um novo conjunto de cortes é obtido multiplicando-se as $n \times G$ restrições em (2) por z^{kg}, para todo $k = 1, \ldots, m$, e então substituindo-se os produtos $e^g_{ij} \times z^{kg}$ por w^{kg}_{ij}. A partir destas manipulações são obtidas as seguintes restrições:

$$\sum_{j=1}^{n} w^{kg}_{ij} = z^{kg}, \forall g = 1, \ldots, G,\ i = 1, \ldots, m,\ i = 1, \ldots, n. \tag{22}$$

Ambas as formulações, PCH1 e PCH2, embora convexas, são de difícil resolução, pois para $G = 1$ ambas são equivalentes ao problema das p-medianas em sua formulação clássica, o qual é NP-árduo (Kariv e Hakimi, 1979). Além disso, a formulação PCH2, possui uma grande quantidade de variáveis de decisão e restrições, o que poderá afetar sua resolução dada a grande quantidade de memória e processamento que poderão ser necessários.

4. Resultados computacionais

Compara-se, nesta Seção, a performance de cada uma das formulações apresentadas, a PCH1 e PCH2. Para tal, utilizou-se o *solver* CPLEX, versão 12.5 (IBM, 2015) e o computador com a configuração descrita anteriormente. Cada uma das formulações foi executada durante 24 horas em modo paralelo, isto é, 12 *threads* paralelas, para cada uma das 27 instâncias utilizadas. Estas, por sua vez, foram geradas de acordo com o padrão descrito por Blanchard *et al.* (2012) para um estudo com base em simulação de Monte Carlo, para o qual se considerou utilizar um conjunto instâncias em que a estrutura de categorias e segmentos são conhecidos. Desta forma, é possível avaliar a qualidade das soluções obtidas em termos de recuperação da informação.

No entanto, para este trabalho, avalia-se apenas o desempenho das formulações obtidas em termos de limitantes inferiores e soluções ótimas, sendo a capacidade de recuperação da informação do PCH avaliada em trabalhos futuros. Uma descrição detalhada dos fatores utilizados

na geração das 27 instâncias é apresentada na Tabela 1. Dentre estes fatores, tem-se o número total de indivíduos ($m = 150, 300, 450$), o número total de segmentos ($G = 2, 6, 10$), o número total de objetos ($n = 18, 20, 30$), a variância no número total de categorias (medianas) definidos para os segmentos, erros aleatórios introduzidos nas matrizes de distâncias (utilizando-se $N(0, 0.5)$ ou $N(0, 0.1)$) e erros aleatórios introduzidos no número de medianas definido por cada indivíduo (utilizando-se $N(0, 0.5)$ ou $N(0, 0.1)$).

Seguindo-se definições de alguns autores, como Blanchard *et al.* (2012), Blanchard e DeSarbo (2013), Brusco e Cradit (2001) e outros, assumiu-se que o número de segmentos G é conhecido, porém sua composição não. Todas as instâncias utilizadas neste trabalho podem ser obtidas contactando-se o primeiro autor. Os resultados computacionais obtidos a partir da resolução das instâncias descritas são apresentados na Tabela 2.

Tabela 1: Fatores considerados na geração das instâncias

Instância	**Sujeitos** m	**Segmentos** G	**Objetos** n	**Categorias**	**Perturbação Dissimilaridades**	**Perturbação Categorias**
1	150	10	30	50 % 3, 50 % 6	N(0, 0.1)	N(0, 0.5)
2	300	2	18	All 6	N(0, 0.1)	0
3	450	2	18	50 % 3, 50 % 6	N(0, 0.05)	0
4	150	2	18	All 3	N(0, 0.05)	N(0, 0.5)
5	450	10	18	All 6	N(0, 0.05)	N(0, 1)
6	150	10	18	50 % 3, 50 % 6	N(0, 0.05)	0
7	300	2	18	All 6	0	N(0, 0.5)
8	150	10	18	50 % 3, 50 % 6	0	N(0, 1)
9	300	10	30	All 3	N(0, 0.05)	N(0, 0.5)
10	450	6	18	All 3	N(0, 0.1)	N(0, 1)
11	150	6	30	All 6	N(0, 0.1)	0
12	300	10	18	All 3	0	0
13	450	10	18	All 6	N(0, 0.1)	0
14	300	6	18	50 % 3, 50 % 6	0	N(0, 1)
15	300	2	30	All 6	N(0, 0.05)	N(0, 1)
16	450	2	30	50 % 3, 50 % 6	0	N(0, 1)
17	300	6	18	50 % 3, 50 % 6	N(0, 0.1)	N(0, 0.5)
18	300	6	30	50 % 3, 50 % 6	N(0, 0.05)	0
19	150	6	18	All 6	0	N(0, 0.5)
20	450	6	30	All 3	0	0
21	150	2	30	All 3	N(0, 0.1)	N(0, 1)
22	450	2	18	50 % 3, 50 % 6	N(0, 0.1)	N(0, 0.5)
23	450	6	18	All 3	N(0, 0.05)	N(0, 0.5)
24	300	10	18	All 3	N(0, 0.1)	N(0, 1)
25	150	6	18	All 6	N(0, 0.05)	N(0, 1)
26	150	2	18	All 3	0	0
27	450	10	30	All 6	0	N(0, 0.5)

A Tabela 2 apresenta os melhores limitantes inferiores, observados a partir da resolução da formulação PCH2. Estes valores foram utilizados para calcular o GAP para ambas as formulações, uma vez que o *solver* não foi capaz de obter tais valores a partir da resolução da formulação PCH1. Como destacado anteriormente, este problema geralmente ocorre em função da utilização da estratégia de reformulação a partir do método do Grande M. No entanto, a formulação PCH1 mostrou-se vantajosa ao obter soluções integrais de melhor custo para 21 das 27 instâncias. Como indicado na Tabela 2, isto se deve ao fato de que esta formulação, por possuir menos variáveis de decisão, permitiu ao solver explorar um maior número de soluções por meio do algoritmo de

Tabela 2: Resultados computacionais obtidos via CPLEX 12.5: PCH1 e PCH2

Instância	PCH2				PCH1		
	Melhor Lim. Inferior	Melhor Lim. Superior	Nós BB	GAP	Melhor Lim. Superior	Nós BB	GAP
1	2705.93	-	1	-	3507.73	158485	22.86%
2	2336.58	2389.32	8310	2.21%	2391.95	410248	2.31%
3	4398.94	4607.07	1097	4.52%	5104.79	330278	13.83%
4	1823.61	1870.07	41186	2.48%	1871.15	2966838	2.54%
5	3355.22	7353.26	1	54.37%	4900.88	22200	31.54%
6	1382.72	2396.45	276	42.30%	1762.60	82193	21.55%
7	2413.33	2651.01	3432	8.97%	2651.01	1756295	8.97%
8	1470.84	1824.17	8	19.37%	1585.17	84443	7.21%
9	6654.00	8425.28	1	21.02%	7784.45	7443	14.52%
10	4995.30	7267.40	3	31.26%	5889.93	58635	15.19%
11	2500.14	4235.07	1	40.97%	2958.48	585748	15.49%
12	3749.99	4850.00	1	22.68%	3779.99	374311	0.79%
13	3064.39	7330.24	1	58.20%	4678.68	21621	34.50%
14	2905.84	4648.16	15	37.48%	3227.33	354894	9.96%
15	5644.12	6087.38	677	7.28%	5775.77	105515	2.28%
16	9670.02	10752.50	29	10.07%	10204.20	53782	5.23%
17	2671.69	4127.72	1601	35.27%	3386.65	90194	21.11%
18	5958.80	8448.21	1	29.47%	7073.91	14741	15.76%
19	1190.01	1269.01	342	6.23%	1264.67	138893	5.90%
20	10935.00	12645.00	1	13.52%	11435.70	16408	4.38%
21	3405.44	3726.86	111	8.62%	3639.28	8291974	6.43%
22	4233.50	4710.93	1871	10.13%	4906.25	188348	13.71%
23	5312.94	7262.72	5	26.85%	5784.20	46099	8.15%
24	3212.93	4834.31	3	33.54%	3983.77	83792	19.35%
25	1141.32	1460.81	1752	21.87%	1363.56	147203	16.30%
26	1874.99	1874.99	1	0.00%	1874.99	3330090	0.00%
27	8657.60	12690.00	1	31.78%	10103.40	23794	14.31%

branch-and-bound. A Tabela 3 apresenta o número de variáveis de decisão presentes em cada formulação, permitindo-se compará-las sob este aspecto.

Tabela 3: Total de variáveis de decisão para PCH1 e PCH2

Variáveis (tipo)	PCH1 Genérico	PCH2 Genérico	PCH1 Instância 1	PCH2 Instância 1
e (binárias)	$G * n^2$	$G * n$	9000	300
e (reais)	-	$G(n^2 - n)$	0	8700
z (binárias)	$m * G$	$m * G$	4500	4500
w (reais)	$m * G * n$	$m * G * n^2$	135000	4050000
c (reais)	$G * n^2$	-	9000	0
Total Binárias	$G(n^2 + m)$	$G(n + m)$	13500	4800
Total Reais	$G(m * n + n^2)$	$G(n^2 - n + m * n^2)$	144000	4058700
Total	$G(2 * n^2 + m + m * n)$	$G(n + m + n^2 - n + m * n^2)$	157500	4063500

Diferentemente da formulação PCH1, para a PCH2 o *solver* apenas foi capaz de resolver o nó raiz no algoritmo de *branch-and-bound* para grande parte das instâncias, dado o grande número de variáveis e restrições presente nesta formulação. Para a Instância 1, por exemplo, nenhuma solução integral foi retornada. O grande diferencial desta formulação, porém, está em sua aplicação para àquelas instâncias em que se tem $G = 2$ segmentos e $n = 18$ objetos, pois observam-se valores para o GAP relativamente baixos. Para o caso da Instância 26, especialmente, a PCH2 permitiu provar a otimalidade das soluções obtidas por ambas formulações. Destaca-se, neste sentido, o

resultado mostrado para a Instância 12, cujo resultado obtido a partir da formulação PCH1 apresenta um GAP de 0.79%.

Considerando-se avaliar o processo de reformulação, assume-se como satisfatórios os resultados apresentados, uma vez que para o modelo original do PCH, o qual foi testado utilizando-se o *solver* Couenne (Couenne, 2015), não foram observados quaisquer limitantes inferiores. Para as instâncias aqui descritas, baseadas no trabalho de Blanchard *et al.* (2012), não são observados na literatura quaisquer limitantes inferiores. Para o PCH, então, sugere-se que a formulação de métodos aproximados, como por exemplo metaheurísticas, seja apropriado, especialmente para casos em que há a necessidade de resolução de instâncias baseadas em dados reais. Por fim, uma vez que se opte por implementar tais métodos, os resultados computacionais mostrados neste trabalho permitirão validar estes algoritmos em termos de custo da função objetivo.

5. Considerações finais

Apresentou-se neste trabalho um novo problema de *clustering*, o Problema de *Clustering* Heterogêneo. Esta nova bordagem poderá permitir que se represente de forma apropriada a heterogeneidade presente nos dados, uma vez que não é necessária a agregação das matrizes de dissimilaridades consideradas como entrada. Este novo modelo permite ainda que se identifiquem segmentos de indivíduos, associando-se a cada um destes segmentos uma estrutura de categorias que melhor representa a percepção destes indivíduos em relação à forma como os objetos se relacionam.

Visto que a formulação original dada para o PCH representa um problema de programação não-linear inteira mista, apresentou-se também duas formulações alternativas para este problema. A primeira delas, a PCH1, baseia-se no método do Grande M, por meio do qual obteve-se um modelo de programação linear inteira mista. Testes computacionais realizados para esta formulação mostraram que, apesar de linear, a obtenção de limitantes inferiores para o problema é um tanto insatisfatória, dadas as limitações deste método de reformulação.

A formulação PCH2, por sua vez, foi obtida por meio do método das Desigualdades de Fortet (Fortet, 1960; Fortet, 1969). Diferentemente da PCH1, para esta formulação foram observados limitantes inferiores para a maior parte das instâncias consideradas, bem como uma solução ótima (Instância 26). A limitação para a aplicação desta formulação está relacionada ao elevado número de variáveis de decisão e restrições adicionados ao modelo original para sua convexificação.

A partir dos resultados obtidos neste trabalho, pode-se analisar a eficiência de métodos aproximados, como metaheurísticas, permitindo-se validá-los. Adicionalmente, mostrou-se que há a necessidade de que novas estratégias de reformulação sejam exploradas, a fim de que soluções ótimas sejam obtidas para um conjunto maior de instâncias. Como continuidade desta pesquisa, destaca-se a aplicação deste problema a casos reais, bem como a verificação de sua capacidade em recuperar a informação contida nos dados. Para tal, um método aproximado será desenvolvido, a fim de que se possa lidar com instâncias de grande porte a um custo computacional aceitável, bem como será definida uma métrica para a avaliação desta capacidade de recuperação da informação.

Agradecimentos

Esta pesquisa foi parcialmente financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico do Brasil (Edital Universal), prêmio concedido ao segundo autor, e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Referências

Aldenderfer, M. e Blashfield, R., *Cluster Analysis*, Sage Publications, Beverly Hills, 1984.

Bettman, J. R.; Luce, M. F. e Payne, J. W. (1998), Constructive consumer choice processes. *Journal of Consumer Research*, 25(3), 187-217.

Bettman, J. R. e Park, W. (1980), Effects of prior knowledge and experience and phase of the choice process on consumer decision processes: A protocol analysis. *Journal of Consumer Research*, 7, 234-248.

Blanchard, S. J. e DeSarbo, W. S. (2013), A new zero-inflated negative binomial methodology for latent category identification. *Psychometrika*, 78, 322-340.

Blanchard, S. J.; Aloise, D. e DeSarbo, W. S. (2012), The heterogeneous p-median problem for categorization based clustering. *Psychometrika*, 77(4), 741-762.

Brusco, M. J.; Cradit, J. D. (2001), A variable-selection heuristic for k-means clustering. *Psychometrika*, 66(2), 249-270.

Couenne. Couenne, an exact solver for nonconvex minlps. *Relatório técnico IBM e Carnegie Mellon University*. Disponível por WWW em `https://projects.coin-or.org/Couenne`. Acesso em 10 de abril de 2015.

Daws, J. T. (1996), The analysis of free-sorting data: beyond pairwise co-occurrence. *Journal of Classification*, 13(1), 57-80.

Duda, R. e Hart, P., *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

Fortet, R. Lálgèbre de boole et ses applications en recherche opérationelle. *Cahiers du Centre d´ Études de Recherche Opérationelle*, 1(4), 5-36, 1959.

Fortet, R. (1960), Applications de l'algèbre de boole en recherche opérationelle. *Revue Franćaise d´ Informatique et de Recherche Opérationelle*, 4(14), 17-26.

Griffin, A. e Hauser, J. R. (1993), The voice of customer. *Marketing Science*, 12(1), 1-27.

IBM. IBM cplex optimizer. *Relatório técnico IBM*, disponível por WWW em `http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/`. Acesso em 11 de abril de 2015.

Isen, A. M. *Toward understanding the role of affect in cognition*. Lawrence Erlbaum, Hillsdale, 2012.

John, D. R. e Sujan, M. (1990), Age differences in product categorization. *Journal of Consumer Research*, 16, 452-460.

Kariv, O. e Hakimi, S. L. (1979), An algorithmic approach to location problems, part ii: p-medians. *Journal of Applied Mathemathics*, 37, 539-560.

Kelter, S.; Cohen, R.; Engel, D.; List, G. e Stronher, H. (1977), The conceptual structure of aphasic and schizophrenic patients in a nonverbal sorting task. *Journal of Psycholinguistic Research*, 6(4), 279-303.

Kohn, H. F.; D. S. e Brusco, M. J. (2010), *The p-median model as a tool for clustering psychological data.* Psychological Methods, 15(1), 87-95.

Medin, D. L. e Schaffer, M. M. (1978), Context theory of classification learning. *Psychological Review*, 85(3), 207-238.

Miller, G. A. (1969), A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, 6(2), 169-191.

Lloyd, S. P. (1957). Least square quantization in PCM. *Bell Telephone Laboratories Paper*.

Perkins, W. S. (1993), The effects of experience and education on the organization of marketing knowledge. *Psychology & Marketing*, 10(3), 169-183.

Ross, B. H. e Murphy, G. L. (1999), Food for thought: cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 38(4), 495-554.

Sherali, H. D. e Desai, J. (2005), A global optimization rlt-based approach for solving the fuzzy clustering problem. *Journal of Global Optimization*, 33, 597-615.

Sherali, H. D. e Alameddine, A. (1992), A new reformulation-linearization technique for bilinear programming problems. *Journal of Global Optimization*, 2(4), 379-410.

Shugan, S. M. (1980), The cost of thinking. *Journal of Consumer Research*, 7(2), 99-111.

Simon, H. A. (1955), A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1), 99-118.

Sujan, M. e Dekleva, C. (1987), Product categorization and inference making: some implications for comparative advertising. *Journal of Consumer Research*, 14(3), 372-378..

Urban, G. L.; Hulland, J. S. e Weinberg, B. D. (1993), Premarket forecasting for new consumer durable goods: modeling categorization, elimination, and consideration phenomena. *Journal of Marketing*, 57(2), 47-63.