SOBRAPO

PODes
Revista Eletrônica Pesquisa Operacional para o Desenvolvimento

Rio de Janeiro, v.3, n.3, p. 186-200, setembro a dezembro de 2011

APLICAÇÃO DE UM MÉTODO DE GERAÇÃO DE AGRUPAMENTOS, BASEADO NO PROBLEMA DO CAIXEIRO VIAJANTE, PARA SEGMENTAÇÃO DE MERCADO

Camilo Brandão de Resende
Instituto Tecnológico de Aeronáutica (ITA)
camilo@ita.br

Rodrigo Arnaldo Scarpel
Instituto Tecnológico de Aeronáutica (ITA)
rodrigo@ita.br

Resumo

Os métodos aplicados a geração de agrupamentos são usados em diversas áreas do conhecimento para categorizar entidades em grupos homogêneos. Alguns exemplos aparecem em mineração de dados, onde a organização de grandes grupos de dados torna a análise estatística mais fácil e eficiente e em reconhecimento de padrões. O objetivo deste trabalho é aplicar um método de geração de agrupamentos, baseado no problema do caixeiro viajante, em que o número ideal de agrupamentos é determinado de forma automática, não necessitando, por parte do analista, nenhum critério subjetivo para sua determinação, conforme verificado nos métodos tradicionais, na criação de agrupamentos de respondentes para segmentação de mercado.

Palavras-chave: geração de agrupamentos, problema do caixeiro viajante, número ideal de agrupamentos.

Abstract

The clustering methods are commonly used in many knowledge areas to categorize entities into homogeneous groups. Some examples appear in data mining, where the organization of big data sets make the statistical analysis easier and more efficient and in pattern recognition. The objective of this work is to propose a new clustering method, based on the traveling salesperson problem, where the ideal number of clusters is automatically determined, avoiding any subjective criterion to determine it as happens with other clustering methods, used in clustering respondents on market segmentation.

Keywords: clustering, traveling salesperson problem, ideal number of clusters.

Versão inicial submetida em 17/01/2011. Versão final recebida em 24/05/2011.

1. Introdução

Os métodos aplicados a geração de agrupamentos tiveram seu início nas ciências biológicas, onde foram desenvolvidos a fim de prover taxonomias de espécies de animais e vegetais (WEDEL; KAMAKURA, 2000). O uso e desenvolvimento desses métodos se espalharam por um grande número de disciplinas científicas, como medicina, psicologia, sociologia, economia, pesquisa de mercado e outras. Os métodos de geração de agrupamentos são conhecidos por diferentes nomes, dependendo da sua área de aplicação, como taxonomia numérica, análise-Q, reconhecimento de padrões não-supervisionado, métodos de geração de segmentos, métodos de geração de *clusters* e métodos de geração de conglomerados.

Esses métodos são comumente usados em diversas áreas do conhecimento para categorizar entidades (objetos, animais, indivíduos etc.) em grupos que são homogêneos ao longo de uma série de características observadas (WEDEL; KAMAKURA, 2000). Alguns exemplos aparecem em mineração de dados, onde a organização de grandes grupos de dados torna a análise estatística mais fácil e eficiente e na identificação de variáveis que são mais importantes para descrever um fenômeno (MINGOTI; LIMA, 2006).

Segundo Webb (2002), os métodos de geração de agrupamentos são importantes ferramentas na área de reconhecimento de padrões, sendo também amplamente utilizados como uma etapa que precede o desenvolvimento de modelos preditivos. Por esses métodos agrupam-se observações de um conjunto de dados formando subgrupos, nos quais as observações são mais similares entre si do que as observações contidas em outros subgrupos (DUDA *et al.*, 2001).

Jain *et al.* (1999) afirmam que a análise de agrupamentos é útil em diversas situações para análise exploratória de padrões, agrupamento de observações, tomada de decisão e para aprendizado de máquina, mas em vários desses problemas, existe pouca informação *a priori* (por exemplo, modelos estatísticos) disponível sobre os dados, e o tomador de decisão deve fazer o mínimo possível de suposições a respeito dos dados. Os autores dizem que é sob essas restrições que métodos de geração de agrupamentos são particulamente apropriados para a exploração de interrelações entre os dados a fim de avaliar sua estrutura.

Segundo Duda *et al.* (2001), a utilização de métodos de geração de agrupamentos é normalmente justificada por ser de fácil aplicação e, muitas vezes, produzir resultados interessantes que podem orientar a aplicação de procedimentos mais rigorosos.

Uma grande questão na utilização dos métodos de geração de agrupamentos é a determinação do número ideal de agrupamentos (WEDEL; KAMAKURA, 2000). Assim sendo, a eliminação da subjetividade na determinação do número ideal de agrupamentos é de grande interesse acadêmico.

Este trabalho tem por objetivo aplicar um método de geração de agrupamentos, baseado no problema do caixeiro viajante, em que o número ideal de agrupamentos é determinado de forma automática, não necessitando, por parte do analista, nenhum critério subjetivo para sua determinação, conforme verificado nos métodos tradicionais, na criação de agrupamentos de respondentes para segmentação de mercado.

Em relação aos procedimentos metodológicos, este trabalho teve uma abordagem quantitativa na resolução do problema e quanto aos procedimentos técnicos, ele trabalho baseou-se em uma revisão bibliográfica dos métodos de formação de agrupamentos para, então, aplicar um método baseado no problema do caixeiro viajante em segmentação de mercado.

Na Seção 2 é feita uma revisão dos métodos de geração de agrupamentos. A Seção 3 é dedicada a descrever o método de geração de agrupamentos utilizado. Na Seção 4 o método foi empregado em segmentação de mercado e na Seção 5 foi feita a conclusão do trabalho, bem como, propostas de trabalhos futuros.

2. Métodos de geração de agrupamentos

O problema de geração de agrupamentos foi descrito por Everitt (1992) conforme se segue: dada uma coleção de N objetos, cada qual descrito por uma série de p variáveis, deseja-se encontrar grupos que sejam internamente homogêneos e externamente heterogêneos. Tanto o número de grupos quanto as suas propriedades em termos das p variáveis devem ser determinadas.

Segundo Wedel e Kamakura (2000), três grandes categorias de métodos de geração de agrupamentos podem ser identificadas : métodos sem sobreposição; métodos com sobreposição e métodos fuzzy.

No caso dos métodos sem sobreposição, uma entidade pertence a um, e somente um, agrupamento. Dois diferentes tipos de métodos de geração de agrupamentos sem sobreposição são comumente distinguidos: métodos hierárquicos e não-hierárquicos. Os métodos hierárquicos não identificam um conjunto de agrupamentos diretamente. Esses métodos identificam relações hierárquicas entre os N objetos utilizando alguma medida de similaridade entre os mesmos. Alguns exemplos de métodos hierárquicos são os métodos da ligação simples, da ligação completa, da ligação média, do centróide e o método de Ward. Métodos não-hierárquicos derivam agrupamentos da amostra diretamente de uma matriz de dados, normalmente através da otimização de uma função objetivo. Os métodos k-médias e k-medóides são exemplos de métodos não hierárquicos onde uma função quadrática é minimizada.

A hipótese de isolação externa é relaxada nos métodos de geração de agrupamentos com sobreposição e fuzzy. Em agrupamentos com sobreposição, uma entidade pode pertencer a mais de um agrupamento. Segundo Banerjee *et al.* (2005) existe uma variedade de importantes aplicações para métodos de geração de agrupamentos onde é mais apropriado permitir que as observações pertençam simultaneamente a mais de um agrupamento. Em biologia, por exemplo, alguns genes possuem mais do que uma função ao codificar proteínas que participam de múltiplas funções metabólicas. Assim sendo, ao agrupar uma série de genes, é apropriado alocá-los em múltiplos agrupamentos sobrepostos, conforme feito por Segal *et al.* (2003).

No caso de agrupamentos fuzzy, entidades pertencem parcialmente a mais de uma agrupamento. Dois diferentes tipos de métodos de geração de agrupamentos fuzzy podem ser distinguidos: os procedimentos baseados na teoria de conjuntos fuzzy e os processos de mistura. Os processos fuzzy e de mistura são conceitualmente diferentes. Processos de mistura assumem que os agrupamentos são não-sobrepostos, porém devido à informação limitada presente nos dados, entidades são relacionadas aos agrupamentos com incertezas, refletidas em probabilidades de pertinência a cada agrupamento, enquanto procedimentos fuzzy assumem que as entidades realmente pertencem parcialmente a diferentes agrupamentos (WEDEL; KAMAKURA, 2000).

Um esquema da classificação de categorias de métodos de geração de agrupamentos apresentadas por Wedel e Kamakura (2000) é mostrado na Figura 1.

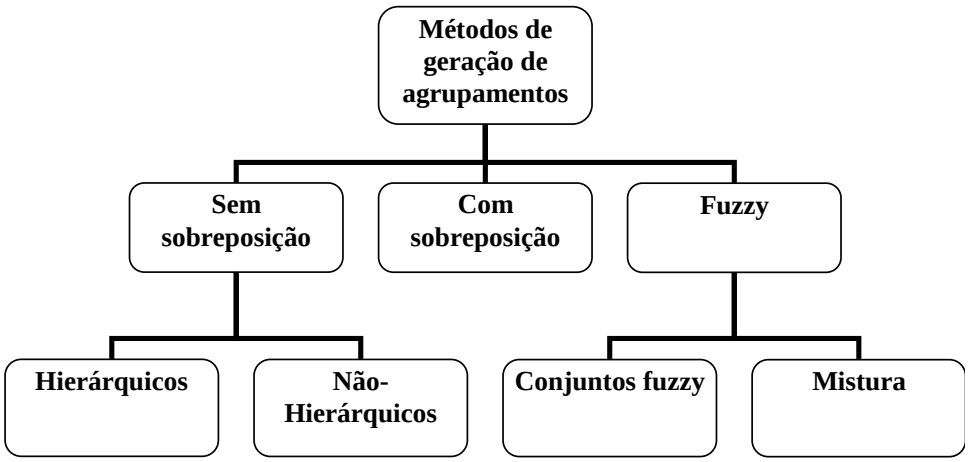

Figura 1 – Classificação dos métodos de geração de agrupamentos

2.1. Métodos hierárquicos de geração de agrupamentos

Métodos hierárquicos de geração de agrupamentos foram amplamente aplicados em marketing a fim de identificar a estrutura hierárquica de mercados de produtos e ainda com o intuito de encontrar segmentos de mercado (WEDEL; KAMAKURA, 2000).

Classificações hierárquicas tipicamente resultam em um dendograma, uma árvore estruturada que representa as relações hierárquicas entre todos os objetos que estão sendo agrupados. Segundo Wedel e Kamakura (2000), agrupamentos não são encontrados diretamente pelos métodos hierárquicos, e um pesquisador que procure uma solução com um determinado número de agrupamentos precisará decidir como retirar esses agrupamentos da árvore estruturada produzida. Um exemplo hipotético dendograma, que é o resultado da aplicação de um método hierárquico de formação de agrupamentos, é mostrado na Figura 2.

Métodos hierárquicos de geração de agrupamentos operam utilizando uma base de similaridade/dissimilaridade relativa aos objetos que estão sendo agrupados. Uma variedade de medidas de similaridade, dissimilaridade e distância podem ser utilizadas para se efetuar a análise hierárquica de geração de agrupamentos. Essas medidas determinam o quão forte é a relação entre os objetos agrupados e são derivadas de variáveis conhecidas dos objetos. Segundo Wedel e Kamakura (2000), o tipo de medida de similaridade/dissimilaridade utilizada deve ser escolhido pelo pesquisador, dependendo das características do problema em questão. Os tipos de medidas de dissimilaridade mais comumente utilizados para dados métricos são apresentados na Tabela 1.

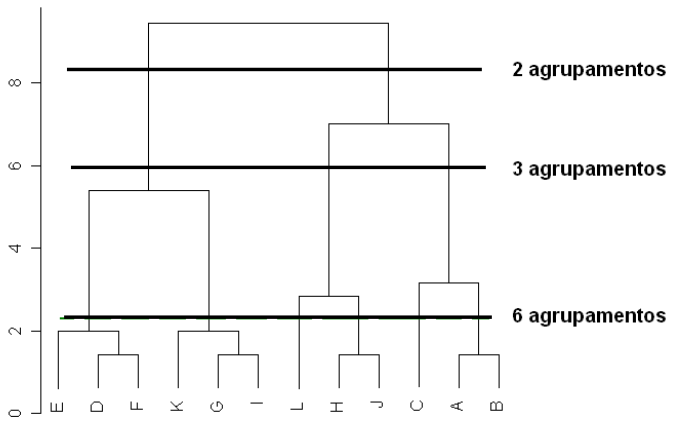

Figura 2 – Exemplo hipotético de dendograma

Tabela 1 – Medidas de dissimilaridade mais utilizadas

Distância Euclidiana	$(\sum_k (y_{nk} - y_{mk})^2)^{1/2}$
Coeficiente de Correlação	$1 - \sum_k (y_{nk} - y_n)(y_{mk} - y_m) / \sigma_n \sigma_m$
Distância "*City Block*"	$\sum_k \lvert y_{nk} - y_{mk} \rvert$
Distância de Mahalanobis	$(y_n - y_m)' \sum^{-1} (y_n - y_m)$
Distância de Minkowski	$(\sum_k w_k (y_{nk} - y_{mk})^r)^{1/r}$
Distância Angular	$\sum_k y_{nk} y_{mk} / (\sum_k y_{nk}^2 y_{mk}^2)^{1/2}$
Distância de Camberra	$\sum_k \lvert y_{nk} - y_{mk} \rvert / (y_{nk} + y_{mk})$

Os métodos hierárquicos mais comumente utilizados são os métodos hierárquicos aglomerativos, que funcionam basicamente da seguinte forma (MINGOTI; LIMA, 2006): no primeiro estágio cada um dos N objetos a serem agrupados são considerados como sendo um agrupamento distinto. Os objetos são então comparados entre si utilizando alguma medida de dissimilaridade como, por exemplo, distância euclidiana. Os dois agrupamentos que são mais similares são unidos. O mesmo procedimento é repetido até atingir o número desejado de agrupamentos. Somente dois agrupamentos podem ser unidos a cada estágio e, uma vez que esses tenham sido unidos, eles não podem mais serem separados. Uma relação de dissimilaridade entre agrupamentos é utilizada para comparar os agrupamentos em cada estágio e para decidir quais deles devem ser os próximos a serem unidos.

A Tabela 2 apresenta as definições de dissimilaridade entre agrupamentos para alguns dos métodos mais comumente utilizados. Segundo Johnson e Wichern (2002) *apud* Mingoti e Lima (2006), os métodos da ligação simples, da ligação completa e da ligação média podem ser utilizados para variáveis quantitativas e qualitativas, enquanto os métodos de Ward e centróide são apropriados somente para variáveis quantitativas.

Tabela 2 – Relação de dissimilaridade entre agrupamentos mais utilizadas

Algoritmo	**Relação de dissimilaridade entre agrupamentos**
Ligação Simples	Menor distância entre membros dos dois agrupamentos
Ligação Completa	Maior distância entre membros dos dois agrupamentos
Ligação Média	Distância média entre membros dos dois agrupamentos
Centróide	Distância entre os centróides dos dois agrupamentos
Ward	Mínimo incremento na soma total de quadrados

2.2. Métodos não hierárquicos de geração de agrupamentos

Nos métodos não hierárquicos de geração de agrupamentos, o número desejado de agrupamentos k deve ser predefinido. A proposta então é agrupar as N observações em k agrupamentos homogêneos internamente e heterogêneos externamente (MINGOTI; LIMA, 2006).

O método k-médias é um algoritmo não-hierárquico de geração de agrupamentos, sendo provavelmente o mais bem conhecido (JOHNSON; WICHERN 2002 *apud* MINGOTI; LIMA 2006).

Se o objetivo é particionar N observações, no espaço P-dimensional, ou seja, com P característica, em k agrupamentos, a formulação, por programação matemática, do método k-médias é mostrada nas equações **1-3**:

Minimizar

$$\sum_{i=1}^{N}\sum_{c=1}^{k} z_{ic}\left[\sum_{p=1}^{P}(\omega_{ip}-m_{cp})^2\right]^{\frac{1}{2}} \quad \textbf{(1)}$$

Sujeito a

$$\sum_{c=1}^{k} z_{ic} = 1 \qquad i = 1,\ldots, N \quad \textbf{(2)}$$

em que

$$z_{ic} = \begin{cases} 1, \text{se a observação i pertencer ao agrupamento c} \\ 0, \text{caso contrário} \end{cases}$$

$$m_{cp} = \frac{\sum_{i=1}^{N} z_{ic}\ \omega_{ip}}{\sum_{i=1}^{N} z_{ic}} \quad \textbf{(3)}$$

sendo ω_{ip} o valor da i-ésima observação na p-ésima dimensão (i=1,...,N e p=1,...,P).

Segundo Mangasarian (1997), o k-médias pode ser resolvido por um procedimento de otimização iterativa em dois passos. No primeiro passo é feita a atribuição dos pontos aos agrupamentos e no segundo passo os centróides dos agrupamentos são atualizados, levando-se em consideração as alocações correntes. Isso conduz ao seguinte algoritmo:

Dado os k centróides $\mathbf{m}_1^t$, $\mathbf{m}_2^t$, …, $\mathbf{m}_k^t$ dos agrupamentos na iteração t, calcula-se $\mathbf{m}_1^{t+1}$, $\mathbf{m}_2^{t+1}$, …, $\mathbf{m}_k^{t+1}$ pelos seguintes passos:

1. Alocação dos pontos aos agrupamentos: Para cada i = 1, …, N, aloca-se a obervação i ao agrupamento c de forma que $\mathbf{m}_{c(i)}^t$ seja o centróide mais próximo da observação i tomando como distância a Euclideana.

2. Atualização dos centróides: Para c = 1,…,k faz-se $\mathbf{m}_c^{t+1}$ ser a média de todas as observações alocados a $\mathbf{m}_c^t$.

Para-se quando $\mathbf{m}_c^t = \mathbf{m}_c^{t+1}$, para c = 1,…,k.

Em relação ao número de iterações necessárias para a convergência, Duda *et al.* (2001) afirmam que é muito menor do que o número de pontos existentes. Esses autores posicionam o método k-médias em uma categoria de procedimentos iterativos de otimização, pois os valores dos centróides tendem a se mover de forma a minimizar uma função de erro quadrática, podendo, então, ser vista como uma forma de se obter estimativas de máxima verossimilhança da média.

Outro método de geração de agrupamentos não hierárquico é o método k-medóides. Esse método seleciona uma observação para ser o centro de cada agrupamento e o seu algoritmo básico é como se segue (ZHANG *et al.*, 2006): inicialmente k observações são selecionadas aleatoriamente para representar cada um dos k agrupamentos a serem encontrados e os demais objetos da amostra são alocados no agrupamento mais próximo, de acordo com sua distância relativa a cada um dos k objetos representantes de seus agrupamentos. Para melhorar a qualidade dos agrupamentos, os objetos representantes são mudados repetidamente. A qualidade dos agrupamentos é estimada por uma função objetivo, como, por exemplo, a função erro quadrático **4**.

Se o objetivo é particionar N observações, no espaço P-dimensional, ou seja, com P característica, em k agrupamentos, a formulação, por programação matemática, do algoritmo k-medoides é mostrada nas equações **4-5**:

Minimizar

$$\sum_{i=1}^{N}\sum_{c=1}^{k} z_{ic}\left[\sum_{p=1}^{t}(\omega_{ip}-\mu_{cp})^2\right]^{\frac{1}{2}} \quad \textbf{(4)}$$

Sujeito a

$$\sum_{c=1}^{k} z_{ic} = 1 \qquad i = 1,\ldots, N \quad \textbf{(5)}$$

em que

$$z_{ic} = \begin{cases} 1, \text{se a observação i pertencer ao agrupamento c} \\ 0, \text{caso contrário} \end{cases}$$

sendo ω_{ip} o valor da i-ésima observação na p-ésima dimensão (i=1,...,N e p=1,...,P) e μ_{cp} são os valores da observação representante do agrupamento c na p-ésima dimensão (c=1,...,k e p=1,...,P).

2.3. Determinação do número ideal de agrupamentos

Uma das grandes dificuldades na análise de agrupamentos é a determinação do número ideal de agrupamentos. Métodos não-hierárquicos são todos condicionais a um número assumido de agrupamentos. Entretanto, o número real de agrupamentos presentes nos dados é desconhecido na maioria dos casos e precisa ser determinado (WEDEL; KAMAKURA, 2000).

Alguns métodos foram propostos para determinar o número de agrupamentos. Uma alternativa é a técnica proposta por Duda *et al.* (2001). Segundo os autores, quando os agrupamentos são formados pela otimização de uma função critério, ou função objetivo, uma abordagem comum é repetir o procedimento de geração de agrupamentos para k = 1, k = 2, k = 3, ..., k = n, verificando como essa função critério varia em função do número de agrupamentos k. No caso do método k-médias verifica-se que a função critério **1** decresce com o aumento de k. Ainda segundo Duda *et al.* (2001), se as N observações estão agrupadas naturalmente em $\hat{k}$ agrupamentos bem separados, espera-se que haja um decrescimento rápido da função critério até que k = $\hat{k}$ e um decrescimento mais lento da função critério a partir desse ponto, até chegar a zero quando k = N.

Esse procedimento é bastante subjetivo. Mais que isso, segundo Wedel e Kamakura (2000), em um extenso estudo de simulação essa abordagem não obteve bons resultados: ela recuperou o número real de agrupamentos somente em aproximadamente 28% das vezes com diversas variações de condições.

3. Proposta de um novo método para lidar com o problema de geração de agrupamentos

Nesta seção, um método para tratar o problema de geração de agrupamentos baseado no problema do caixeiro viajante é proposto. É sabido, conforme foi discutido na seção 2.3, que a determinação do número ideal de agrupamentos é uma das grandes questões na análise de agrupamentos. A vantagem do método proposto é que o número de agrupamentos é determinado automaticamente, sem necessitar nenhum critério subjetivo por parte do analista.

Nesse trabalho é proposto um método de geração de agrupamentos baseado no problema do caixeiro viajante (TSP) dado por:

Seja um caixeiro viajante que necessita visitar cada uma de N cidades antes de retornar à sua casa. Qual ordem de cidades visitadas minimiza a distância total que o caixeiro viajante deve percorrer antes de retornar?

Matematicamente, o problema do caixeiro viajante consiste em encontrar o circuito hamiltoniano que tenha o menor comprimento em um grafo completo. Segundo Schrijver (2005), essa questão foi inicialmente estudada por Kirkman e Hamilton, em 1856, sendo também abordada por Kowalewski, em 1917. O tamanho do espaço de soluções do TSP é $\frac{1}{2}(N-1)!$ para N maior que 2, onde N é o número de cidades. Esse é o número de circuitos hamiltonianos em um grafo completo de N nós, isto é, circuitos fechados que passam por cada nó exatamente uma vez.

O TSP pode ser formulado e resolvido utilizando programação linear inteira. Winston (2004) o formulou como se segue: suponha que o TSP consista das cidades 1, 2, 3, ..., N. Para i ≠ j seja d_{ij} = distância entre as cidades i e j e seja d_{ii} = M, onde M é um número muito grande relativamente às demais distâncias presentes no problema. Definindo d_{ii} = M assegura-se que a cidade i não será visitada imediatamente após deixar a cidade i. Definindo as variáveis de decisão do problema como

$$x_{ij} = \begin{cases} 1, & \text{se a solução do TSP vai da cidade i para a cidade j} \\ 0, & \text{caso contrário} \end{cases}$$

Então a formulação para o TSP é mostrada nas equações **6-11.**

Minimizar $z = \sum_i \sum_j d_{ij} x_{ij}$ **(6)**

Sujeito a

$\sum_{i=1}^{N} x_{ij} = 1$ para j = 1, 2, ..., N **(7)**

$\sum_{j=1}^{N} x_{ij} = 1$ para i = 1, 2, ..., N **(8)**

$u_i - u_j + N x_{ij} \leq N - 1$ para i ≠ j; i ,j = 2 , …, N **(9)**

$Todo\ x_{ij} \in \{0, 1\}$ **(10)**

$Todo\ u_j \geq 0$ **(11)**

A função objetivo **(6)** fornece o total de comprimentos de arcos incluídos em uma volta. As restrições em **(7)** asseguram que se chegue somente uma vez a cada cidade. As restrições em **(8)** asseguram que se saia somente uma vez de cada cidade. As restrições em **(9)** são a chave para a formulação. Elas asseguram que qualquer combinação de x_{ij}'s contendo uma sub-rota será inviável, isto é, viola **(10)** e ainda que qualquer combinação de x_{ij}'s que forme uma rota completa será viável, ou seja, existirá uma combinação de u_j's que satisfaça **(10)**.

É importante notar, entretanto, que a resolução do TSP usando programação inteira se torna ineficiente e inviável para grandes problemas, sendo usualmente utilizados métodos heurísticos (WINSTON, 2004).

Entretanto, enquanto no problema do caixeiro viajante a formação de sub-rotas é proibida, no método proposto a presença de sub-rotas sugere a presença de subgrupos com coesão interna, ou seja, agrupamentos.

O método de geração de agrupamentos é formulado como se segue: suponha que a amostra consista das observações 1, 2, 3, ..., N; descritas por um conjunto de variáveis. Para i ≠ j seja d_{ij} = uma medida de dissimilaridade entre as observações i e j, por exemplo, a distância

euclidiana entre as observações, e seja d_{ii} = M, onde M é um número muito grande relativamente às medidas de dissimilaridade presentes no problema. Definindo d_{ii} = M assegura-se que a observação i deve se conectar a outra observação, ou seja, não existe nenhum agrupamento que contenha apenas uma observação. Ainda definindo as variáveis de decisão do problema como

$$x_{ij} = \begin{cases} 1, & \text{caso as observações i e j pertençam ao mesmo agrupamento} \\ 0, & \text{caso contrário} \end{cases}$$

Então a formulação do método de geração de agrupamentos é **12-16**.

Minimizar $z = \sum_i \sum_j d_{ij} x_{ij}$ **(12)**

Sujeito a

$$\sum_{i=1}^{N} x_{ij} = 1 \qquad \text{(para j = 1, 2, ..., N)} \qquad \textbf{(13)}$$

$$\sum_{j=1}^{N} x_{ij} = 1 \qquad \text{(para i = 1, 2, ..., N)} \qquad \textbf{(14)}$$

$$x_{ij} + x_{ji} \leq 1 \qquad \text{(para i > j; i = 1, …, N; j = 2, N - 1)} \qquad \textbf{(15)}$$

$$Todo\ x_{ij} \in \{0, 1\} \qquad \textbf{(16)}$$

A função objetivo **(12)** fornece o comprimento total de arcos incluídos em todos os subgrupos (sub-rotas). As restrições em **(13)** asseguram que toda observação j deve ter uma, e somente uma, outra observação que se ligue a ela. Isso significa que as duas observações pertencem ao mesmo agrupamento. As restrições em **(14)** asseguram que toda observação i deve se ligar a uma, e somente uma, outra observação.

As restrições em **(15)** são essenciais à formulação. Elas asseguram que qualquer combinação de x_{ij}'s contendo um subgrupo com 2 observações será inviável e ainda que qualquer combinação de x_{ij}'s que formar subgrupos com 3 ou mais observações será viável.

As restrições **(15)** são necessárias porque caso permitamos a formação de subgrupos com duas observações, é possível que todos os agrupamentos encontrados com um número par de componentes tenham somente dois componentes. Isto ocorre porque agrupamentos com um número par de componentes e maior do que dois, sempre podem ser divididos em outros que tenham somente dois componentes, resultando em um menor ou igual comprimento total de arcos. A Figura 3 mostra um exemplo de agrupamento com quatro componentes.

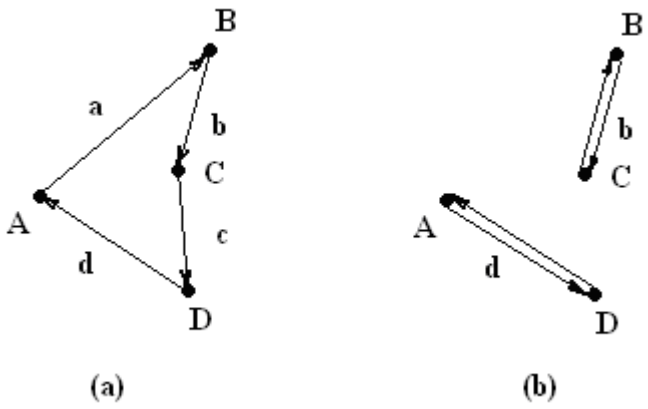

Figura 3 - Exemplo de um agrupamento com quatro componentes

Inicialmente compare (a+c) com (b+d). Suponha, sem perda de generalidade, que (b+d) é menor ou igual a (a+c). Então podemos formar dois novos agrupamentos: um formado por A e D, e o outro formado por B e C. A nova distância total de arcos é 2(b+d) e é menor ou igual à distância total de arcos original (b+d+a+c).

Agora, suponha que tenhamos um agrupamento com 2n componentes, onde n é um número inteiro, conforme mostra a Figura 4.

Inicialmente compare $\sum_{i=1}^{n} a_{2i}$ e $\sum_{i=1}^{n} a_{2i-1}$. Suponha, sem perda de generalidade, que $\sum_{i=1}^{n} a_{2i}$ é menor ou igual a $\sum_{i=1}^{n} a_{2i-1}$. Então podemos formar n novos agrupamentos: agrupamento 1 = {A_2 ;A_3}, agrupamento 2 = {A_4,A_5}, …, agrupamento n = {A_{2n},A_1}. A nova distância total de arcos é $2\sum_{i=1}^{n} a_{2i}$ e é menor ou igual à distância total original ($\sum_{i=1}^{2n} a_i = \sum_{i=1}^{n} a_{2i} + \sum_{i=1}^{n} a_{2i-1}$).

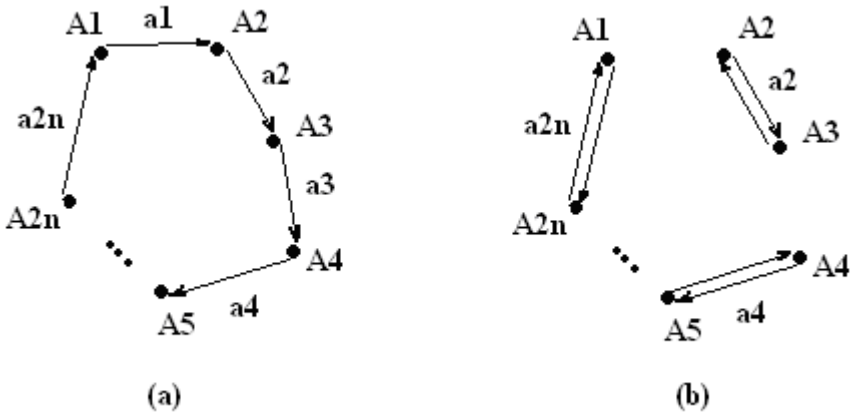

Figura 4 - Exemplo de um agrupamento com um número par de componentes

4. Utilização do método proposto em segmentação de mercado

A fim de ilustração, o método proposto foi aplicado na criação de agrupamentos de respondentes, em uma pesquisa de preferências por veículos automotores, objetivando agrupar os respondentes com preferêrncias similares. Este procedimento é comumente realizado em segmentação de mercado. Neste trabalho, tomou-se as preferências dos respondentes, em relação às características dos produtos, estimada utilizando a técnica de análise conjunta.

Os dados utilizados nesta ilustração foram coletados em uma pesquisa com 44 respondentes. Para estimar suas preferências foram oferecidos dez modelos de veículos (combinações de atributos), todos do segmento sedan-compacto, descritos por suas característas, e pediu-se que os mesmos fossem ordenados de 1 a 10, sendo atribuído nível 10 ao preferido e nível 1 ao menos desejado. As características consideradas e os modelos oferecidos são mostrados nas Tabelas 3 e 4, respectivamente.

Utilizando os dados coletados, foram então estimados os coeficientes a_{iml}'s da equação

$$R_{ij} = \sum_{m=1}^{M} \sum_{l=1}^{L_m} a_{iml} x_{jml} + e_{ij} \qquad \textbf{(17)}$$

para cada um dos respondentes, em que R_{ij} é a classificação do respondente i para o produto j, a_{iml} é o peso dado ao l-ésimo nível do m-ésimo atributo pelo respondente i, L_m é o número de níveis do atributo m, M é o número de atributos, e_{ij} é o termo de erro, assumido como tendo uma distribuição normal de média zero e variância σ^2 para todo i e j e

$$x_{jml} = \begin{cases} 1, \text{se o l - ésimo nível do m - ésimo atributo está presente no produto j} \\ 0, \text{caso contrário} \end{cases}$$

A estimação dos pesos a_{iml} é feita para cada um dos respondentes utilizando o método dos mínimos quadrados. De forma alternativa, uma vez que foi utilizada uma combinação ortogonal de atributos no experimento, os pesos podem ser estimados resolvendo um sistema linear com 14 incógnitas e 14 equações, em que as 14 incógnitas representam os pesos dos 13 possíveis níveis de atributos, mais uma constante e as 14 equações são dadas pelas 10 observações obtidas através do formulário, mais 4 equações que correspondem a igualar um nível de cada uma das 4 características a zero. Esse procedimento é equivalente a resolução de um problema de regressão linear pelo método dos mínimos quadrados no caso em que as variáveis independentes são ortogonais e a soma dos pesos estimados é uma constante.

Tabela 3 – Características consideradas e seus respectivos níveis

Marca					**Preço (10^3 R$)**			**Porta-Malas (l)**			**Potência (cv)**	
F	R	D	C	V	28	33	38	350	425	500	75	105

Tabela 4 – Modelos (combinações de características) oferecidos aos respondentes

Modelo	**Marca**	**Preço (1000 R$)**	**Porta-Malas (l)**	**Potência (cv)**
1	F	33	500	105
2	R	38	500	105
3	D	38	350	105
4	C	38	425	75
5	V	38	350	75
6	F	28	425	105
7	D	28	500	75
8	V	33	425	105
9	C	28	350	105
10	R	33	350	75

Após a resolução dos sistemas lineares para cada respondente, a fim de facilitar a interpretação, os valores dos pesos a_{iml}'s obtidos foram normalizados, sendo atribuído o valor zero ao peso do nível menos desejado de cada atributo e o valor 100 à combinação de atributos que produz o perfil de produto mais desejado por esse respondente, conforme sugerido por Lilien et al. (2007) e foram calculadas as importâncias relativas de cada característica na determinação das preferências. A importância relativa de uma característica foi considerada como sendo o peso dado ao maior nível daquela característica dividido pela soma dos maiores pesos de cada uma das características. Os resultados são mostrados nas Tabelas 5 e 6.

Tabela 5 – Pesos dos níveis de atributos estimados para os respondnetes

Atributo	Nível	Peso médio	Desvio padrão
Marca	F	14,86	19,81
	R	19,18	18,51
	D	17,52	18,28
	C	24,25	20,52
	V	21,68	17,47
Preço	28000	20,54	18,98
	33000	17,38	13,33
	38000	4,95	8,66
Porta-Malas	350	4,14	6,43
	425	8,89	10,80
	500	8,01	10,87
Potência	75	2,76	8,29
	105	18,72	16,89

Analisando as Tabelas 5 e 6, percebe-se que tanto o peso dado a cada nível de atributo como à importância relativa de cada atributo considerado possuem altos desvios padrão, o que sugere grande heterogeneidade de preferências.

Em relação à importância relativa dos atributos, na média, a marca é o fator mais importante na determinação das preferências, seguido por preço e potência e o atributo menos importante é o volume do porta-malas. Porém, como há grande heterogeneidade de preferências entre os respondentes, é possível que hajam agrupamentos homogêneos de respondentes com importâncias relativas diferentes da média.

Tabela 6 – Importâncias relativas estimadas para os respondentes

Atributo	Importância relativa média (%)	Desvio padrão (%)
Marca	37,9	19,8
Preço	26,0	15,9
Porta-Malas	14,5	11,6
Potência	21,5	15,8

No que diz respeito às marcas, na média dos respondentes, para a categoria considerada, a C é a preferida, seguida pela V, depois pela R, D e a de menor preferência é a F. Quanto ao preço, na média, ocorre o esperado, ou seja, a maior preferência é pelo menor valor (R$28.000,00) e a menor preferência é pelo maior valor (R$38.000,00). Em relação a potência do motor, a maior preferência é pelo motor mais potente (105 CV) e para o volume do porta-malas é pelo valor intermediário (425 litros).

Objetivando formar agrupamentos homogêneos de respondentes, os métodos k-médias e o proposto foram utilizados.

4.1. Utilização do método k-médias para geração de agrupamentos

No primeiro caso ilustrado, foi utilizado o método k-médias para agrupar os respondentes. A primeira decisão a ser tomada é em relação ao número de agrupamentos a ser considerado.

Na Figura 5 é apresentada a evolução da função objetivo do problema (*Within-Cluster sum of Squares*) variando o número k de agrupamentos. Pela Figura 5, não fica claro que exista um valor $\hat{k}$, tal que haja um decrescimento rápido da função objetivo até que k = $\hat{k}$ e um decrescimento mais lento da função a partir desse ponto. Assim sendo, torna-se difícil e

subjetiva a utilização do critério sugerido por Duda et al. (2001) para a determinação do número de agrupamentos.

Uma vez que a utilização do critério sugerido por Duda et al. (2001) não indicou a presença de agrupamentos bem separados, foi arbitrado o agrupamento dos respondentes em dois segmentos distintos, a fim de se encontrar dois novos produtos, sendo cada um deles o ideal para cada um dos segmentos obtidos. Os resultados são apresentados nas Tabelas 7 a 9.

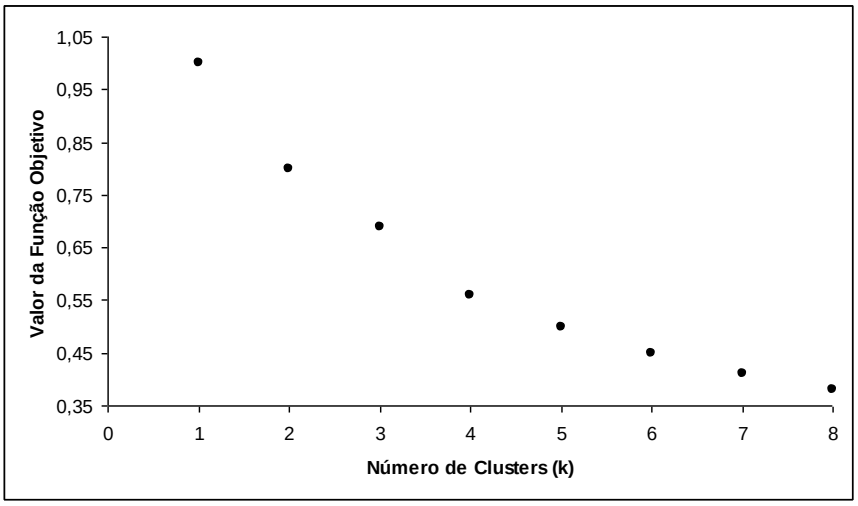

Figura 5 - Evolução do valor da função objetivo para k = 1, 2, ..., 8

Analisando as Tabelas 7 a 9, verifica-se a partir dos valores médios dos pesos encontrados para cada agrupamento juntamente com a importância relativa média de cada atributo que o agrupamento 1 é basicamente composto por respondentes que dão grande importância à marca do veículo, apresentando preferência pela marca C e que são pouco sensíveis ao preço, representando 29,5% dos respondentes. Já o agrupamento 2 representa 70,5% dos respondentes e é composto por respondentes que são mais sensíveis ao preço do que aqueles do agrupamento 1 e que apresentam, também, grande preferência por veículos mais potentes.

Tabela 7 – Freqüência relativa dos respondentes em cada agrupamento

Agrupamento	Frequência relativa (%)
1	29,5
2	70,5

1.1.1.

4.2. Utilização do método proposto para geração de agrupamentos

No segundo caso ilustrado, após a estimação das preferências dos respondentes, foi utilizado o método de geração de agrupamentos proposto para agrupar os respondentes em função de semelhantes preferências. Obteve-se como solução a formação de dois agrupamentos. Os resultados são apresentados nas Tabelas 10 a 12.

Tabela 8 – Valores médios dos pesos de cada atributo nos agrupamentos encontrados

Agrup.	Marca					Preço (10^3 R$)			Porta-Malas (l)			Potência (cv)	
	F	R	D	C	V	28	33	38	350	425	500	75	105
1	30,31	33,35	30,37	42,90	34,34	8,55	9,47	8,74	3,50	6,23	7,19	2,19	7,32
2	8,37	13,24	12,14	16,44	16,36	25,57	20,69	3,35	4,41	10,00	8,36	3,00	23,51

Tabela 9 – Importância relativa média de cada atributo para um dos agrupamentos

Agrupamento	Marca	Preço	Porta-Malas	Potência
1	63%	16%	11%	10%
2	27%	30%	16%	27%

Tabela 10 – Freqüência relativa dos respondentes em cada agrupamento

Agrupamento	Frequência relativa (%)
1	50,0
2	50,0

Tabela 11 – Valores médios dos pesos de cada atributo nos agrupamentos encontrados

Agrup.	Marca					Preço (10^3 R$)			Porta-Malas (l)			Potência (cv)	
	F	R	D	C	V	28	33	38	350	425	500	75	105
1	21,32	19,68	17,87	14,02	13,46	24,64	21,25	6,22	5,32	13,49	9,87	1,57	11,73
2	8,39	18,68	17,18	34,49	29,89	16,45	13,51	3,68	2,96	4,29	6,15	3,95	25,71

Tabela 12 – Importância relativa média de cada atributo para um dos agrupamentos

Agrupamento	Marca	Preço	Porta-Malas	Potência
1	36%	31%	20%	13%
2	40%	21%	9%	30%

Analisando as Tabelas 10 a 12, podemos perceber que o agrupamento 1 é caracterizado por respondentes que dão grande importância ao preço de um veículo e ainda apresentam preferência pela marca F. Já o agrupamento 2 é caracterizado por respondentes que dão grande importância à potência de um veículo, sendo menos sensíveis ao preço do que os respondentes do agrupamento 1, e apresentam preferência pela marca C, sendo F a marca menos preferida.

Comparando os métodos empregados para a formação dos agrupamentos, o problema apresentado na utilização do método k-médias foi na determinação do número de agrupamentos, uma vez que não ficou claro na função objetivo do problema o número ideal de agrupamentos, sendo necessário arbitrá-lo. Além disso, o método proposto, mostrou interessante por demandar menos esforço por parte do analista, uma vez que, para se determinar o número de agrupamentos pelo k-médias é necessário processar o método diversas vezes variando o número de agrupamentos.

Em relação aos resultados obtidos aplicando-se os métodos, não é possível tirar conclusões por se tratar apenas de uma ilustração em que os dados de uma amostra não representativa da população.

5. Conclusão

A determinação do número de agrupamentos é uma grande questão em métodos de geração de agrupamentos. Assim sendo, o método de geração de agrupamentos proposto se mostrou interessante pela eliminação da subjetividade na determinação do número de agrupamentos.

O método proposto também se mostrou interessante por demandar menos esforço por parte do analista do que outros métodos tradicionais, uma vez que, para se determinar o número de agrupamentos nos métodos hierárquicos, é necessário fazer um dendograma e o método k-médias precisa ser rodado diversas vezes variando o número de agrupamentos. É importante também salientar que é inviável a utilização de métodos hierárquicos quando o número de observações é muito grande, uma vez que se torna muito difícil a interpretação dos dendogramas obtidos.

O método proposto apresenta como ponto fraco o fato de ser um problema de otimização combinatória, fazendo com que sua resolução usando programação inteira se torne

ineficiente e inviável para grandes problemas, sendo então necessária a utilização de métodos não exatos, como por exemplo metaheurísticas.

Como sugestão para trabalhos futuros, pretende-se aplicar o método proposto para a criação de agrupamentos homogêneos utilizando grandes bases de dados para avaliar a eficiência do método e fazer uso de metaheurísticas que tornem a busca por soluções mais eficiente.

Referências bibliográficas

BANERJEE, A.; KRUMPELMAN, C.; GHOSH, J.; BASU, S.; MOONEY,R. Model-based overlapping clustering. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p.532-537, Chicago, 2005.

DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification. 2.ed. New York: John Wiley & Sons, 2001.

EVERITT, B. S. Cluster analysis. London: Edward Arnold, 1992.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A review. ACM Computing Surveys (CSUR). V.31, n.3, p.264-323, 1999.

LILIEN, G. L.; RANGASWAMY, A.; BRUYN, A. Technical note n.9. Principles of Marketing Engineering, Trafford Publishing, 2007.

MANGASARIAN, O. L. Mathematical programming in data mining. Data Mining and Knowledge Discovery, v.1, n.2, p.183-201, 1997.

MINGOTI, S. A., LIMA, J. O. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research, v.174, n.3, p.1742-1759, 2006.

SCHRIJVER, A. On the History of combinatorial optimization (till 1960). Handbook of Discrete Optimization. Amsterdan: Elsevier, p.1-68, 2005.

SEGAL, E.; BATTLE, A.; KOLLER, D. Decomposing gene expression into cellular processes. Proceedings of the 8th Pacific Symposium on Biocomputing (PSB), 2003.

WEBB, A. Statistical pattern recognition, 2· ed. Chichester District of West Sussex: John Wiley & Sons, 2002.

WEDEL, M.; KAMAKURA, W. A. Market segmentation: conceptual and methodological foundations. 2. ed. Hingham: Kluwer Academic, 2000.

WINSTON, W. L. Operations research: applications and algorithms. 4. ed. Thomson Learning, 2004.

ZHANG, X.; WANG, J.; WU, F. Spatial clustering with obstacles constraints based on genetic algorithms and k-medoids. International Journal of Computer Science and Network security, v.6, n.10, p.109-114, 2006.