"Raio de Influência": um método de agrupamento alternativo para Análise de *Cluster*

Bruno Monte de Castro, *bruno.monte19@gmail.com*, DEMA, UFC
Silvia Maria de Freitas, *silvia@ufc.br*, DEMA, UFC
Bruno de Athayde Prata, *baprata@yahoo.com.br*, DEMA, UFC
George Leitão Evangelista, *geogle1@gmail.com,* Geslog, UFC

Resumo: A análise multivariada, de uma maneira geral, refere-se a todos os métodos estatísticos que, de forma simultânea, analisam múltiplas variáveis em relação aos objetos em investigação. Dentre esses métodos, destaca-se a análise de *cluster* ou agrupamento, que se aplica em diversas áreas. As técnicas de análise de *cluster* têm a função de organizar, em grupos disjuntos, os objetos em estudo, de forma que os mesmos apresentem semelhanças entre si – dentro de cada grupo. Essas técnicas dividem-se em hierárquicas e não-hierárquicas, sendo que não existe uma técnica "ótima", pois ambas apresentam vantagens e desvantagens. Com o interesse em suprir essa falha, é proposto nesse trabalho uma nova abordagem para agrupamentos, unindo-se as características das duas técnicas na forma de um algoritmo híbrido chamado de *Raio de Influência*. Um exemplo clássico usado na literatura foi testado, verificando-se e comparando-se os seus resultados com os outros métodos já conhecidos. As comparações feitas são expostas na forma de um gráfico, chamado Dendograma que mostra o *layout* do agrupamento.

Palavras-chave: Análise de Cluster, Medidas de similaridade, Método híbrido

1 Introdução

O grande sábio grego Aristóteles disse: "O homem vive classificando tudo o que vê". Classificar significa agrupar, tendo por base aspectos de semelhança entre os elementos classificados. Ao classificar moedas, por exemplo, leva-se em conta critérios de semelhança como o país de origem, o ano em que a moeda foi criada, etc. Um dos maiores problemas encontrados em várias áreas de pesquisa é realizar uma maneira de agrupar as informações para um melhor entendimento dos dados e assim obter resultados significativos. O agrupamento é realizado de forma a minimizar as diferenças entre os objetos em estudo dentro do agrupamento (*cluster*) e maximizar as diferenças entre os elementos de agrupamentos diferentes. A análise de *cluster* constitui-se de métodos multivariados cujo interesse é a apresentação de uma estrutura de classificação dos elementos em grupos, com base nas semelhanças obtidas pelas características (variáveis) em estudo. O objetivo desse trabalho é comparar e relatar as vantagens e desvantagens das técnicas conhecidas como: hierárquicas, não-hierárquicas e um método híbrido, proposto por Freitas & Prata (2007), conhecido como *Raio de Influência*. Nesse estudo foi utilizado um conjunto de dados descrito na literatura e, considerando-se as medidas de distância Euclidiana, Euclidiana ao quadrado e Mahalanobis, com alguns métodos hierárquicos e não-Hierárquicos, para comparar a performance e eficiência com o método *Raio de Influência.*

2 Dados

Os dados considerados nessa análise são provenientes de um estudo da ONU (2002), extraídos de Mingoti (2005), que apresenta os seguintes índices: expectativa de vida, educação, renda (PIB) e estabilidade política relativos a um conjunto de 21 países. Esses índices foram constituídos por uma metodologia proposta pela ONU e, quanto maiores seus valores, melhor caracterizado seria o país. É desejável se agrupar países com índices cujos valores são próximos, pois indicam um padrão de desenvolvimento semelhante.

3 Metodologia

Considere uma amostra aleatória multivariada (isto é, p-variada) em n elementos, dada por:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$$

sendo $\mathbf{X}_j$ o vetor $p \times 1$ das variáveis mensuradas no elemento j da amostra.

Deve-se definir uma medida de similaridade a ser utilizada para decidir se dois elementos da amostra são semelhantes ou não. As medidas de distâncias são úteis na comparação dos n elementos da amostra através das medidas realizadas em p-variáveis, onde serão agrupados aqueles elementos que possuírem menor distância, pois indicam uma maior semelhança. Para esse conjunto de dados foram usadas três medidas de similaridade: distância Euclidiana, distância Euclidiana ao quadrado e distância de Mahalanobis. Foram comparadas as seguintes técnicas:

i) Hierárquicas, que por sua vez podem ser classificadas como aglomerativas ou divisivas;
ii) Técnicas não-hierárquicas;
iii) *Método do Raio de Influência*, que é uma combinação das duas técnicas citadas acima.

3.1 Medidas de Similaridade

Todas as medidas de similaridade têm vantagens e desvantagens e nesse trabalho abordou-se somente as medidas de distâncias. As principais medidas de distâncias descritas por Cormack (1971) são:

1. Distância Euclidiana: A distância entre dois elementos i e i′ é dada por:

$$d_{ii'} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

ou na forma matricial

$$d_{ii'} = [(\mathbf{X}_i - \mathbf{X}_{i'})^\top (\mathbf{X}_i - \mathbf{X}_{i'})]^{1/2}$$

em que $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^\top$ e $\mathbf{X}_{i'} = (X_{i'1}, \cdots, X_{i'p})^\top$.

2. Distância Generalizada ou Ponderada: A distância Generalizada entre dois elementos i e i' é dada por:

$$d_{ii'} = [(\mathbf{X}_i - \mathbf{X}_{i'})^\top \mathbf{A} (\mathbf{X}_i - \mathbf{X}_{i'})]^{1/2}$$

em que $A_{p\times p}$ é a matriz de ponderação, sendo a mesma positiva definida. A escolha dessa matriz diz o nível de ponderação que pode ser adotado. Desse método obtém-se algumas formas particulares, tais quais, se A é a matriz identidade tem-se a distância euclidiana, se A é igual a S^{-1} tem-se a distância de Mahalanobis.

3. Distância Euclidiana ao Quadrado: A distância entre dois elementos i e i' é dada por:

$$d_{ii'} = [(\mathbf{X}_i - \mathbf{X}_{i'})^\top (\mathbf{X}_i - \mathbf{X}_{i'})]$$

3.2 Técnicas Hierárquicas

Com a escolha da medida de distância deve-se agora selecionar o critério de construção dos grupos. A técnica inicia com o cálculo de uma matriz se similaridade/dissimilaridade entre os elementos, baseado nas medidas de distâncias vistas na seção anterior, e termina com um dendograma (diagrama da árvore), mostrando as junções sucessivas dos indivíduos até formar um único grupo. Existem vários algoritmos de agrupamento, mas neste trabalho será abordado somente cinco técnicas, consideradas mais importantes, que são:

i) Método da Ligação Simples (Single Linkage) ou também conhecido como método do vizinho mais próximo;
ii) Método da Ligação Completa (Complete Linkage) ou método do vizinho mais distante;
iii) Método da Média das Distâncias (Average Linkage);
iv) Método de Ward (Ward's Method).

Essas técnicas satisfazem a propriedade de hierarquia, na qual a cada iteração, um novo grupo é formado a partir dos anteriores. Se dois elementos aparecem juntos um dado agrupamento, eles permanecem juntos até o final do processo. Devido a essa propriedade é possível construir dendogramas.

3.3 Técnicas não-Hierárquicas

Essas técnicas têm por objetivo encontrar diretamente uma partição de n elementos em g grupos pré-especificados pelo pesquisador. Além da especificação inicial do número de grupos, a novidade é que novos agrupamentos podem ser feitos a partir de

outros já formados, isto é, se dois elementos estão juntos em um *cluster*, não necessariamente eles estarão unidos até o final do processo, como consequência, não se pode construir Dendogramas. O método mais utilizado é conhecido como K-Means, na qual se escolhe k centróide como sementes iniciais e cada elemento do conjunto de dados é comparado a cada centróide inicial. O elemento é alocado ao grupo cuja distância é a menor e após repetir esse processo para cada elemento, recalcula-se os valores dos centróides para cada novo grupo formado, e repete-se o procedimento até que todos os elementos amostrais estejam "bem alocados"em seus grupos, não necessitando de uma nova iteração. Deve-se ter cuidado com os valores da semente inicial, pois podem influenciar o resultado final de forma considerável. As possíveis formas são:

i) Vetor de média de cada grupo nas técnicas hierárquicas;
ii) Escolha aleatória;
iii) Primeiros valores do banco de dados.

3.4 Método do *Raio de Influência*

Os métodos hierárquicos constroem agrupamentos de maneira simples, pecando por enumerarem os agrupamentos de maneira exaustiva. Os métodos não-hierárquicos convergem rapidamente, contudo, carecem de subjetividade na escolha dos agrupamentos iniciais. Para não ficar limitado a somente um determinado método propõe-se o uso o método do Raio de Influência, que combina as técnicas de agrupamento hierárquicas e as não hierárquicas. O algoritmo mostrou-se eficiente e os resultados esperados de sua aplicação mostraram-se consideráveis. Para realizar o método do Raio de Influência, deve-se seguir os seguintes passos:

i) Passo 1: Determinar, para cada observação do conjunto de dados analisado, o somatório das distâncias especificada pelo pesquisador a todos os demais pontos do conjunto. Ordenar as observações em ordem crescente numa lista DMIN.

ii) Passo 2: Determinar o raio de influência de cada observação. O raio de influência é dado pelo somatório das distâncias de cada ponto aos demais, dividido pelo número de observações.

iv) Passo 3: Avaliar, para o primeiro elemento de DMIN (primeiro nó semente), quais as observações estão contidas dentro do seu raio de influência, compondo, então, um cluster.

v) Passo 4: Repetir, para os elementos subsequentes de DMIN que encontram-se fora dos raios de influência dos seus antecessores, o Passo 3. Caso uma observação que já compõe um cluster esteja mais próxima de outro candidato a nó semente, ele deve sair do agrupamento inicial e compor um novo cluster com esse novo candidato.

4 Resultados e Conclusões

Na análise de agrupamento não se tem um método ótimo de grupos, tanto as técnicas hierárquicas, não-hierárquicas como as híbridas têm suas vantagens e

desvantagens. As vantagens das técnicas hierárquicas são: a simplicidade, uso de diferentes medidas de similaridades e a rapidez. As desvantagens são: redução do impacto dos *outliers*, não são boas para grandes conjunto de dados.

Nas técnicas não-hierárquicas as vantagens são: os resultados não são tão afetados por *outliers* e podem ser utilizados para grandes conjuntos de dados. As desvantagens são: o uso aleatório de centróides iniciais faz com que o método seja inferior ao hierárquico e mesmo não sendo as sementes aleatórias, a técnica não garante uma solução ótima. O método não é aconselhável em situações onde existem muitos agrupamentos.

A Figura 1 apresenta os resultados dos agrupamentos para a técnica não-hierárquica, considerando-se a distância Euclidiana respectivamente.

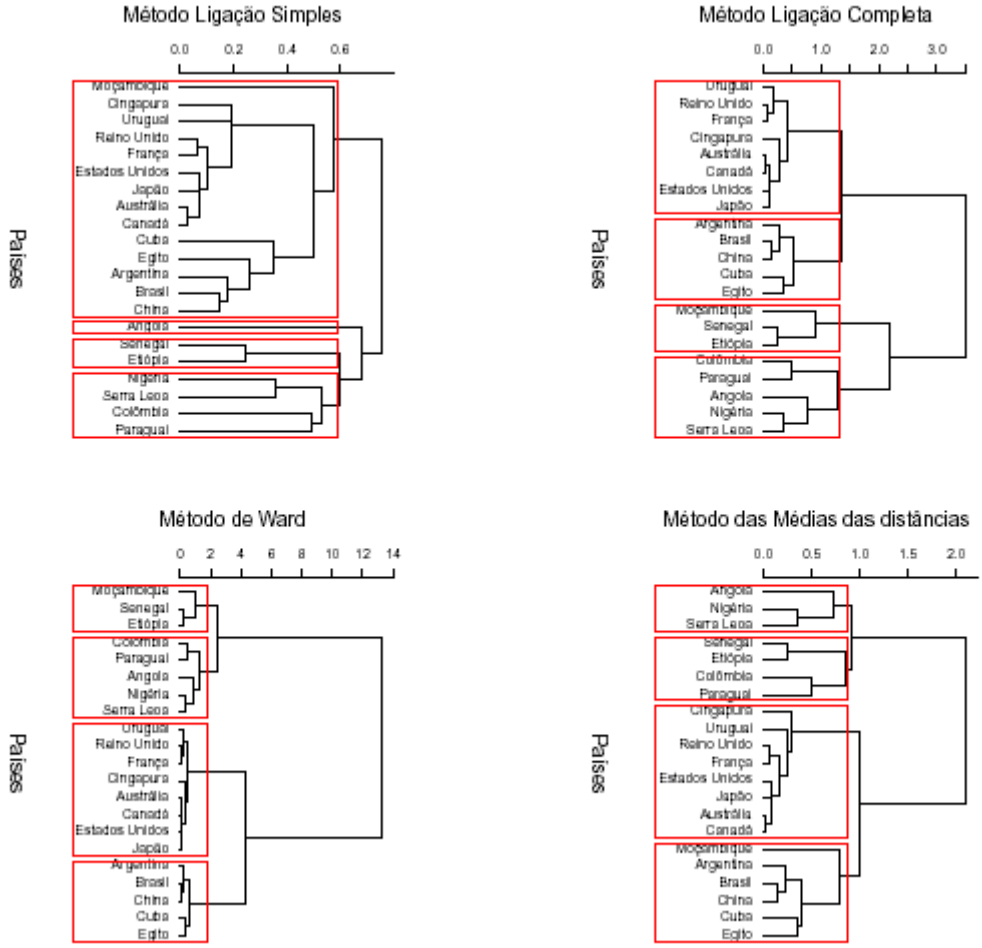

Figura 1. Agrupamento usando a técnica hierárquica – Distância Euclidiana.

Usando a Distância Euclidiana como medida de similaridade e o método “Raio de Influência” obtém-se:

Agrupamento 1 : Reino Unido
Agrupamento 2 : EUA,Japão, Canadá,Moçambique,Austrália
Agrupamento 3 : Brasil, China
Agrupamento 4 : Argentina,Egito, Cuba

Agrupamento 5 : Uruguai, França
Agrupamento 6 : Senegal, Etiópia, Cingapura, Paraguai
Agrupamento 7 : Colômbia,Serra Leoa, Nigéria
Agrupamento 8 : Angola

Aplicando o método de *K-Means* no R, usando o pacote Rcmdr, e considerando-se 4 grupos para a técnica, obtém-se:

Agrupamento 1 : Paraguai, Etiópia, Senegal
Agrupamento 2 : Reino Unido, Austrália, Canadá, Estados Unidos, Japão, França, Cingapura, Uruguai
Agrupamento 3 : Brasil, China, Moçambique, Argentina, Cuba, Egito
Agrupamento 4 : Serra Leoa, Angola, Colômbia, Nigéria.

As técnicas hierárquicas e não-hierárquicas dependem muito da subjetividade do pesquisador, diferente do método do *Raio de Influência* que é desprovido desse fato. Os procedimentos aqui descritos, ainda precisam ser avaliados para outras medidas de distância, para então ser possível um estudo mais criterioso da eficiência entre os métodos. O método do *Raio de Influência* além de ter solução única, converge rapidamente, pois o número de iterações é no máximo o número de observações. Como limitações podem ser citadas: o método é influenciado por valores extremos e pode-se dizer que o método é “cauteloso”, pois só forma agrupamento com os elementos bastantes semelhantes. Pela própria definição da medida DMIN do método do *Raio de Influência*, em função de uma média de distâncias, esta sofre influência dos valores aberrantes, dessa forma, estão sendo avaliadas variações nessa medida, como, por exemplo, uma medida de mediana ou uma média ponderada, para que seja possível um estudo mais aprofundado do comportamento do método.

5 Agradecimentos

Os autores Bruno Monte Castro e George Leitão Evangelista agradecem ao CNPq/PIBIC e à Universidade Federal do Ceará pelo apoio financeiro que contribuiu para a elaboração desse trabalho.

Referências

[1] Everitt, B. S & Hothorn, T. (2010). A Handbook of Statistical Analysis using R. Chapman & Hall. New York.
[2] Freitas, S. M & Prata, B. de A. (2007). Uma nova abordagem para a análise de agrupamento com uma aplicação em agronomia. 12o Seagro.
[3] Hair Jr., J.F.; Anderson, R.E.; Tatham, R.L.; Black, W.C. (1998). Multivariatedata analysis. New Jersey: Prentice Hall.
[4] Mingoti, S. A. (2005). Análise de dados através de métodos de estatística. UFMG. Minas Gerais.
[5] Reis, E. (2001). Estatística Multivariada Aplicada. Edições Silabo.