Uma nova abordagem para a análise de agrupamento com uma aplicação em agronomia

Bruno de Athayde Prata[1]**; Sílvia Maria de Freitas**[2]**.**

Resumo

A Estatística Multivariada, por avaliar múltiplas variáveis em uma observação de uma amostra, destaca-se por ter diversas aplicações, tanto no campo científico quanto nas atividades empresariais. Dentre a família de técnicas multivariadas, a análise de agrupamento (*cluster*) é uma das mais utilizadas, dada sua utilidade pragmática. Os métodos de análise de agrupamento dividem-se em hierárquicos e não-hierárquicos, ambos com suas vantagens e desvantagens. Este artigo tem o objetivo de reportar uma nova abordagem para a análise de *cluster*, com um algoritmo de agrupamento que combine características dos métodos hierárquicos e não-hierárquicos. Um conjunto de dados sobre densidade de vegetação foi utilizado para a avaliação do método, que mostrou-se eficiente, cujos resultados esperados de sua aplicação mostraram-se promissores. Como consideração final, sugere-se a utilização de outras medidas de dissimilaridade, com vistas a avaliar o desempenho do algoritmo em diversas circunstâncias.

Palavras-chave: Análise Multivariada, Análise de Agrupamento, *Cluster*, Raio de Influência.

1. Introdução

A análise multivariada refere-se a todas as técnicas estatísticas que avaliam simultaneamente múltiplas aferições em cada indivíduo de uma amostra. As aplicações da estatística multivariada se estendem sobre diversos campos do conhecimento, dentre os quais podem ser ressaltados as ciências exatas, as engenharias, as ciências da terra, a medicina, a psicologia e a administração.

Dentre as diversas técnicas que compõem o arcabouço da análise multivariada, a análise de agrupamento (*cluster*) é uma das que mais se destaca dada sua potencial aplicação nos ramos de atuação supramencionados. A análise de *cluster* destina-se a segmentar indivíduos de uma amostra de modo a formar conjuntos mutuamente excludentes que apresentem similaridades entre seus elementos.

[1] Mestrando em Logística e Pesquisa Operacional pela Universidade Federal do Ceará. E-mail: bprata@det.ufc.br.

[2] Profa. Dra. do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará. E-mail: silvia@ufc.br.

Conforme Hair Jr. *et al.* (1998), a análise de agrupamento é o nome do grupo de técnicas multivariadas que tem, por objetivo primordial, agrupar indivíduos de uma amostra de acordo com as suas características.

A análise de *cluster* não é uma técnica essencialmente estatística, dado o seu caráter não inferencial, tratando-se assim de um método descritivo. Os métodos de agrupamento, geralmente, não fornecem soluções únicas, pois diversos resultados podem ser obtidos a partir de uma mesma massa de dados. Logo, percebe-se o impacto da subjetividade do pesquisador quando estiver estudando o agrupamento das observações de uma amostra.

As medidas de similaridade são fundamentais para a análise de *cluster*, pois permitem avaliar o grau de semelhança entre os indivíduos que compõem uma amostra, subsidiando um posterior processo de particionamento (ou agrupamento). Duas medidas de similaridade bastante conhecidas são as distâncias Euclidiana (JOHNSON e WICHERN, 1998) e de Mahalanobis (MARDIA *et al.*, 1979).

Este trabalho destina-se a apresentar uma nova abordagem para a análise de agrupamento, que combine vantagens dos métodos hierárquicos e não-hierárquicos. O artigo reporta uma aplicação na área de Agronomia e Engenharia Florestal.

2. Materiais e métodos

A proposta desse método tem, como interesse principal, agregar a filosofia de agrupamento dos métodos hierárquicos e não-hierárquicos, com o intuito de: (*i*) evitar a subjetividade inerente ao pesquisador, comumente utilizada; e (*ii*) reduzir o número de interações utilizadas até a convergência para o agrupamento desejado. A seguir, no Quadro 1, é apresentado o algoritmo do método denotado ***raio de influência***.

Quadro 2 – Método do ***raio de influência***.

Passo 1: Determinar, para cada observação do conjunto de dados analisado, o somatário das distâncias Euclidianas a todos os demais pontos do conjunto. Ordenar as observações em ordem crescente numa lista DMIN. **Passo 2:** Determinar o raio de influência de cada observação. O raio de influência é dado pelo somatório das distâncias Euclidianas de cada ponto aos demais, dividido pelo número de observações. **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*. **Passo 4:** Repetir, para os elementos subseqüentes 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.

Uma sucinta descrição do método é feita a seguir. O nó semente é aquele cujo somatório das distâncias entre os demais pontos seja mínimo. Determina-se, então, o raio de influência do nó semente, que equivale à distância Euclidiana média do ponto supracitado às demais observações. Em seguida, deve-se repetir o procedimento para todos os pontos fora do raio de influência do nó semente da iteração anterior. Caso um novo ponto semente esteja mais próximo de uma observação que já compõe um *cluster*, esta deve migrar para o novo *cluster*.

Com relação à eficiência do algoritmo, podem ser expostos os seguintes comentários. O algoritmo refina o seu processo de busca, construindo melhores soluções a cada iteração, ou, na pior hipótese, corroborando soluções já obtidas. O algoritmo converge rapidamente, pois apresenta, no pior caso, um número de iterações igual ao número de observações. É pertinente ressaltar que o algoritmo ofertará sempre os mesmos resultados, não sofrendo influência da subjetividade do analista.

Para a avaliação da utilidade do método do raio de influência na área de Agronomia, foram utilizados dados de um levantamento da vegetação da Mata da Silvicultura (vide Tabela 1), cuja fonte é Albuquerque *et al.* (2006). Cada espécie ilustrada na Tabela 1 foi representada por um número, de 1 a 17, conforme sua ordem de aparição na mesma.

Tabela 1 – Densidade de 17 espécies da Mata da Silvicultura, em parcelas de 20 X 50 m, Universidade Federal de Viçosa, Viçosa, MG, 1993.

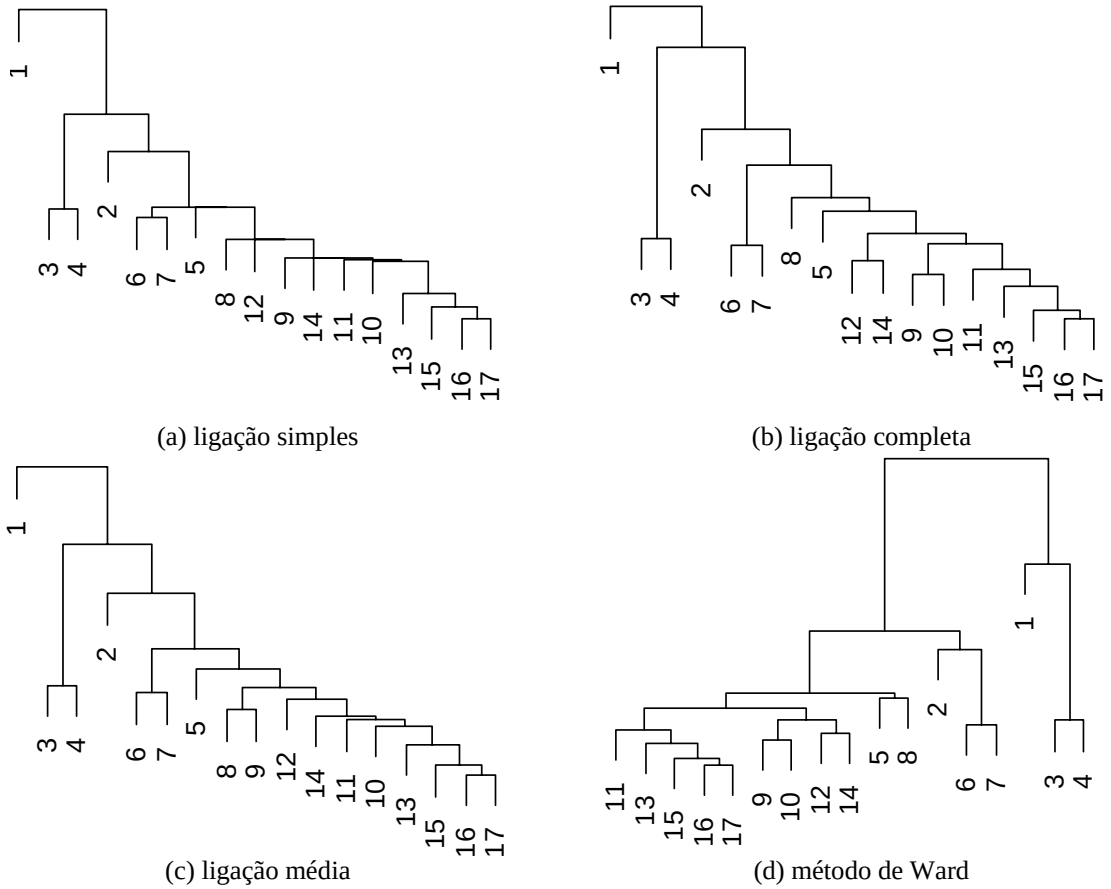
Espécies	Parcelas										
	1	2	3	4	5	6	7	8	9	10	11
Casearia decandra Jacq.	8	1	27	0	1	9	2	3	22	15	7
Anadenanthera peregrina Spreg.	0	0	0	0	0	0	12	1	17	1	9
Apuleia leiocarpa (Vog.) Macbr.	3	9	4	6	22	9	5	2	7	4	4
Mabea fistulifera Mart.	6	3	3	4	29	12	0	4	4	4	4
Anadenanthera macrocarpa (Benth.) Brenan.	0	12	0	1	0	0	1	0	2	0	0
Platypodium elegans Vog.	0	0	1	1	9	1	0	0	5	11	1
Machaerium floridum (Benth.) Ducke	0	0	10	1	9	2	1	0	0	11	5
Copaifera langsdorffii Desf.	1	1	0	2	1	13	0	0	0	3	1
Ocotea pretiosa Mez.	2	0	2	2	2	6	0	5	0	2	2
Cabralea cangerana Saldanha	1	0	0	2	0	0	1	6	2	3	1
Piptadenia gonoacantha Macbr.	0	0	0	0	0	0	6	0	1	0	5
Dalbergia nigra Allem. *ex* Benth.	5	0	7	0	5	0	0	0	0	1	0
Luehea divaricata Mart.	0	0	1	0	0	0	2	0	0	5	2
Cecropia hololeuca Miq.	7	0	0	0	0	1	0	1	0	0	0
Melanoxylon brauna Schott.	0	0	0	0	0	0	0	0	0	2	1
Cedrela fissilis Vell.	0	0	0	0	0	0	1	0	0	0	0
Croton floribundus Spreng.	0	0	1	0	0	0	0	0	0	0	0

Fonte – Albuquerque *et al.* (2006).

3. Resultados e discussões

A aplicação do método do raio de influência forneceu os seguintes resultados: **Agrupamento 1:** 15, 5, 8, 9, 10, 11, 12, 13, 14, 16, 17; **Agrupamento 2:** 6 e 7; **Agrupamento 4:** 2; **Agrupamento 5:** 3; **Agrupamento 6:** 4 e **Agrupamento 7:** 1.

Utilizou-se o mesmo conjunto de dados para avaliar os agrupamentos resultantes das aplicações de quatro métodos correntemente utilizados: ligação simples, ligação completa, ligação média e método de Ward. Para tanto, empregou-se o *software* **R** (**R**, 2006). Os dendogramas dos agrupamentos resultantes estão ilustrados na Figura 1. É pertinente salientar que as abordagens empregadas propiciaram diferentes agrupamentos.



(a) ligação simples (b) ligação completa

(c) ligação média (d) método de Ward

Figura 1 – Dendogramas obtidos por diferentes métodos, para os dados da Tabela 1.

Do ponto de vista estatístico, os resultados obtidos pela aplicação do algoritmo proposto foram condizentes com aqueles obtidos por métodos correntes preconizados pela literatura (vide Figura 1). Do ponto de vista computacional, pode-se destacar que, mesmo para um conjunto com 11 variáveis e 17 observações, o algoritmo convergiu em apenas 5 iterações. Ressalta-se, ainda, que o método é desprovido de qualquer tipo de subjetividade por parte do analista, facilitando sua massiva aplicação por parte de distintos usuários.

4. Conclusões

Este artigo tem como proposta a sugestão de um novo metodo que pode ser utilizado para a análise de agrupamento, denotado método do ***raio de influência***. Foi realizada uma aplicação com um conjunto de dados da área de Agronomia e Engenharia Florestal, cujos resultados se

mostraram bastante satisfatorios. Constatou-se que o método exposto apresenta uma série de benefícios, dentre os quais podem ser destacados: (*i*) o algoritmo converge rapidamente, pois o número máximo de iterações equivale, no pior caso, ao número de observações; (*ii*) a determinação dos pontos sementes independe do analista; portanto, o método não sofre influência de subjetividade; (*iii*) o algoritmo fornece sempre as mesmas soluções, sendo, portanto, determinístico. Assim, para uma mesma massa de dados, o algoritmo só necessita ser executado uma vez; (*iv*) o algoritmo fornece apenas um resultado de agrupamento; e (*v*) sua aplicação em exemplos constantes na literatura apresentou resultados coerentes.

O algoritmo apresenta algumas limitações intrínsecas ou que decorreram de simplificações no escopo da pesquisa, dentre as quais se pode ressaltar que, por trabalhar com a distância Euclidiana média como critério para determinação dos raios de influência dos pontos sementes, a qualidade das soluções geradas é influenciada por *outliers*.

Como sugestões para o aprofundamento do tema pesquisado e aperfeiçoamento do algoritmo proposto, recomenda-se: (*i*) o algoritmo deve ser testado em outros conjuntos de dados de modo a avaliar o desempenho do método em várias situações; tal análise permitirá uma definição mais precisa acerca das vantagens e desvantagens do método; (*ii*) também devem ser empregadas outras medidas de similaridade para avaliar o comportamento do método.

Bibliografia

Albuquerque, M. A. A.; Ferreira, R. L. C.; Silva, J. A. A.; Santos, E. S.; Stosic, B.; Souza, A. L. (2006) Estabilidade em análise de agrupamento: estudo de caso em ciência florestal. Revista Árvore, v.30, p. 257-265.

Hair Jr., J.F.; Anderson, R.E.; Tatham, R.L.; Black, W.C. (1998) *Multivariate data analysis*. New Jersey: Prentice Hall.

Johnson, R. A.; Wichern, D. W. (1998) *Applied multivariate statistical analysis*. New Jersey: Prentice Hall.

Mardia, K. D.; Kent, J. T.; Bibby, J. N. (1979) *Multivariate analysis*. London: Academic Press.

R (2006) *R: A Language and Environment for Statistical Computing - Reference Index*. R Foundation for Statistical Computing.