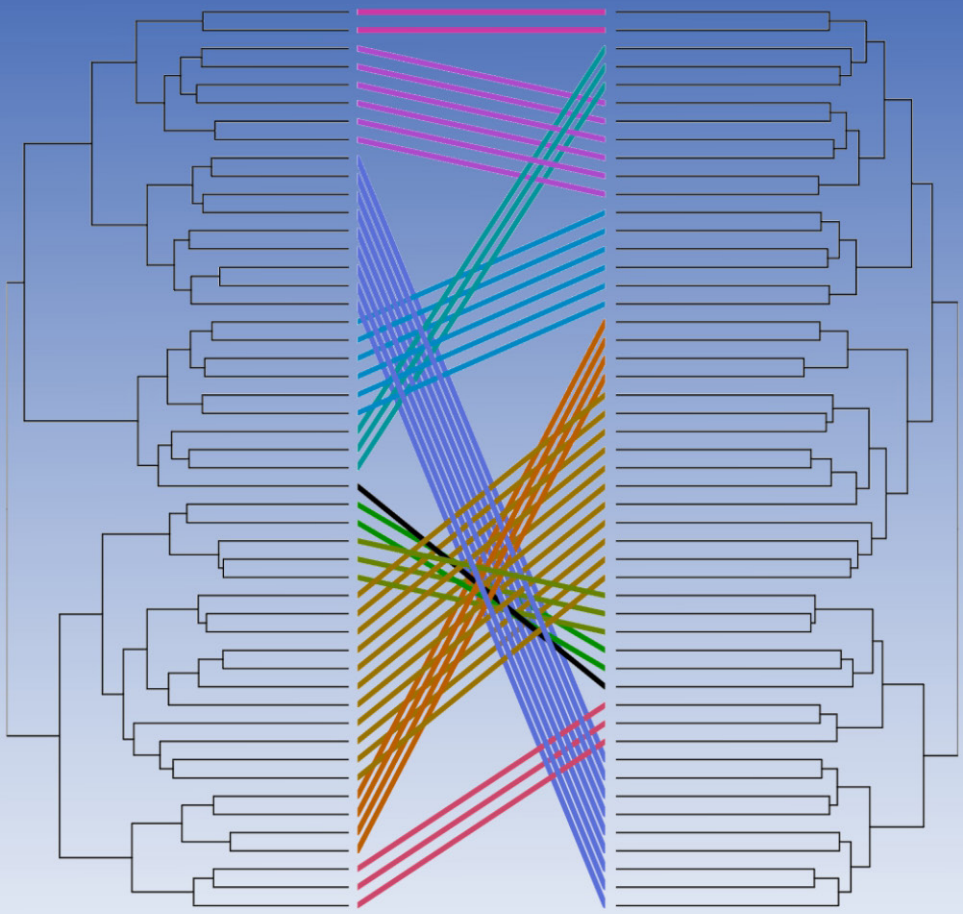
Mácio Augusto de Albuquerque
Kleber Napoleão N. O. Barros

INTRODUÇÃO À ANÁLISE DE AGRUPAMENTO: TEORIA E PRÁTICA COM APLICAÇÕES EM R



eduepb

MÁCIO AUGUSTO DE ALBUQUERQUE
KLEBER NAPOLEÃO NUNES DE OLIVEIRA BARROS

Introdução à Análise de Agrupamento: teoria e prática com aplicações em R

eduepb

Campina Grande-PB
2020

EDITORA DA UNIVERSIDADE ESTADUAL DA PARAÍBA

Diretor
Luciano do Nascimento Silva

Design Gráfico e Editoração
Erick Ferreira Cabral
Jefferson Ricardo Lima Araujo Nunes
Leonardo Ramos Araujo

Revisão Linguística
Elizete Amaral de Medeiros
Antonio de Brito Freire

Divulgação
Danielle Correia Gomes

Depósito legal na Biblioteca Nacional, confome Lei nº 10.994, de 14 de dezembro de 2004.
FICHA CATALOGRÁFICA ELABORADA POR HELIANE MARIA IDALINO DA SILVA - CRB-15ª/368

A345i Albuquerque, Mácio Augusto de.

Introdução à Análise de Agrupamento: teoria e prática com aplicações em R [Livro eletrônico]./Mácio Augusto de Albuquerque, Kleber Napoleão Nunes de Oliveira Barros.– Campina Grande: EDUEPB, 2020.

2.476 Kb. - 174 p. il. color.

ISBN 978-65-86221-01-5 (E-book)

ISBN 978-65-86221-00-8w (Impresso)

1. Análise Multivariada. 2.Medidas de distância. 3. Técnicas estatísticas multivariadas. I.Barros, Kleber Napoleão Nunes de Oliveira. II. Título.

21. ed.CDD 519.535

DEDICATÓRIA

Dedico esse trabalho as pessoas que me ofereceram incentivo, carinho e apoio por toda a vida: meus pais, meus irmãos, minha esposa Edna e filhos Tarsyla e Tárcio que fizeram cada dia da minha vida valer a pena.

Mácio Augusto de Albuquerque

Dedico esse trabalho à minha esposa Patrícia e filhos Khalel e Katharine. Obrigado pela paciência, incentivo e carinho.

Kleber Napoleão N. O. Barros

Sumário

PREFÁCIO

As técnicas estatísticas multivariadas têm sido amplamente empregadas em estudos envolvendo simultaneamente variáveis de clima, solo, relevo, vegetação, economia e geologia no agrupamento. Essas técnicas são utilizadas com objetivos básicos de ordenamento, visando determinar a influência de fatores do meio na composição e produtividade de objetos, e de agrupamento, com o propósito de classificação.

A denominação "Análise Multivariada" corresponde a um conjunto de métodos e técnicas que analisam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados. O primeiro passo para a utilização da análise multivariada é saber o que se pretende afirmar a respeito dos dados. A técnica e o método estatístico ideal para a aplicação devem ser escolhidos de acordo com o objetivo da pesquisa. Há diversas técnicas para a análise multivariada e cada uma delas se aplica a um objetivo de pesquisa específico.

Análise de Agrupamento engloba uma variedade de técnicas e algoritmos cujo objetivo é encontrar e separar objetos em grupos similares. Essa atividade pode ser observada, por exemplo, numa criança brincando com blocos coloridos de diferentes formas, cores e tamanhos. É comum ela separá-los em pilhas segundo uma de suas características, cor por exemplo. Ela está praticando Análise de Agrupamentos. Usar mais de uma característica para formar pilhas torna-se uma atividade mais trabalhosa, exigindo conceitos mais sofisticados de semelhança e procedimentos mais "científicos" para empilhá-las. É sobre este procedimento multidimensional que este livro irá abordar.

Este livro, cujo conteúdo foi elaborado para servir como guia prático para estudantes universitários de diferentes áreas, apresenta a estrutura conceitual

e metodológica do instrumental estatístico da Análise de Agrupamento e suas aplicações computacionais.

Uma vez que o recurso a meios informáticos e nomeadamente a software apropriado é atualizado fundamentalmente no tratamento de dados, introduz-se a vertente computacional através da linguagem R (www.r.projec.org), cuja utilização se tem vindo a alargar rapidamente devido não só à sua simplicidade, mas também ao fato de ser muito vasta e abrangente e de estar disponível gratuitamente na internet.

Sem a pretensão de constituir um livro-texto detalhado, os seus Capítulos convergem para um meio termo entre a teoria e prática dos principais métodos de Análise Agrupamento.

Exemplos resolvidos e exercícios são a base da prática computacional, bem como alguns desafios para aqueles usuários mais interessados em desenvolver suas habilidades em R.

Os data sets utilizados são, em sua grande maioria, disponibilizados na distribuição dos seus próprios pacotes e, portanto, contam com documentação a respeito de origem e descrição dos dados. A seleção do conteúdo é baseada na minha experiência acadêmica com a Análise de Agrupamento e na análise de dados utilizando o R.

Desse modo, no Capítulo 1 apresenta-se o princípio básico de Análise de Agrupamento., através de um exemplo simples e usando critérios intuitivos para distância e técnica de aglomeração.

O Capítulo 2 aborda a difícil questão de escolher o critério de parecença entre elementos. Embora não seja usual, optou-se pelo termo parecença para indicar e chamar a atenção de que está representando os dois conceitos: similaridade e dissimilaridade.

No Capítulo 3, descrevem-se os algoritmos mais conhecidos para a produção de agrupamentos, procurando ressaltar as suas diferenças.

As particularidades para uma aplicação adequada de Análise de Agrupamento, são tratadas no Capítulo 4.

Já no Capítulo 5 descreve-se os principais programas aplicativos para utilização de Análise de Agrupamento bem como ilustra-se a aplicação através de um exemplo mais elaborado.

E finalmente no Capítulo 6 apresentamos alguns comandos e pacotes.

Assim é que, sem negar o meu próprio viés de classificação literária, considero-o como um guia do usuário e, não obstante, assumimos os erros que porventura surgirão. Por fim, deixamos os nossos agradecimentos, a priori, aos leitores que relatarem críticas e sugestões.

CAPÍTULO 1

Análise de Agrupamentos

> Todos nós acreditamos que qualquer população é composta de segmentos distintos. Se trabalhamos com as variáveis, os indivíduos adequados, a Análise de Agrupamentos nos ajudará a ver se existem grupos que são mais semelhantes entre si do que com objetos de outros grupos.

Introdução

Análise de Agrupamento é um nome dado ao conjunto de técnicas utilizadas na identificação de padrões de comportamento em bancos de dados através de formação de grupos homogêneos de casos. Essas técnicas têm aplicabilidades em várias áreas do conhecimento.

O objetivo da Análise de Agrupamentos é obter grupos homogêneos chamados de *cluster*, em que os objetos no mesmo grupo possuem característica mais semelhantes (homogêneos) e os grupos possuem características heterogêneas entre si.

Vale a pena repetir que o objetivo mais comum da Análise de Agrupamento talvez seja tratado como heterogeneidade nos dados. O resultado esperado é um número de grupos, cada um consistindo em um número de objetos relativamente homogêneos com uma variação dentro do grupo consideravelmente menor do que o total de variação no conjunto completo de dados. Mas há outro objetivo, ligeiramente diferente, que também motiva o uso da análise de aglomerados: encontrar uma modalidade natural dos dados. Nesse caso, usa-se a Análise de Agrupamento para determinar se os dados contêm subconjuntos homogêneos de observações que ocorrem simultaneamente.

Como utilização dessas técnicas de agrupamento, podem-se citar:

- redução de grande massa de dados em grupos, de forma a permitir sua análise;
- descrição de dados;
- formulação de hipóteses sobre estrutura de dados; e
- serve como técnica alternativa à análise estatística clássica.

O processo de realização da técnica de agrupamento envolve basicamente duas etapas, em que a primeira relaciona-se com a estimativa de uma função de agrupamento entre as unidades amostrais e a segunda, com a adoção de uma técnica de agrupamento para formação dos grupos, que são abordados a seguir.

Este Capítulo irá ilustrar as principais etapas do procedimento de Análise de Agrupamento., ressaltando as propriedades comuns à maioria dos métodos. Pretende-se também propor um procedimento "científico" que ajude os usuários dessa técnica a avaliar os seus procedimentos. A estrutura básica da aplicação de técnicas de Análise de Agrupamento pode ser decomposta nas seguintes etapas:

i. Definição de objetivos, critérios, escolha de variáveis e objetos.
ii. Obtenção dos dados.
iii. Tratamento dos dados.
iv. Escolha de critérios de similaridade ou dissimilaridade (parecença).
v. Adoção e execução de um algoritmo de Análise de Agrupamento.
vi. Apresentação dos resultados.
vii. Avaliação e interpretação dos resultados.

Convém observar que essas etapas não são independentes. Às vezes, torna-se necessário voltar à etapas anteriores para corrigir e aprimorar etapas posteriores. Mas com a adoção das etapas acima espera-se providenciar ao usuário de Análise de Agrupamento um procedimento metodológico útil. Em capítulos seguintes serão descritas algumas dessas etapas com maiores detalhes.

As diversas etapas serão apresentadas através de um exemplo hipotético, artificial, cujo único objetivo é ilustrar e apresentar as principais decisões necessárias à aplicação de técnicas de Análise de Agrupamento.

Definição do Problema

Pretende-se investigar, exploratoriamente, o histórico de crescimento da massa corpórea das pessoas. O pesquisador gostaria de escolher representantes "típicos" da população para tentar traçar diferentes históricos através de questionários mais complexos.

Desse modo, seria conveniente classificar a população-alvo em grupos homogêneos segundo alguma característica de interesse. Conseguida essa divisão, poder-se-ia restringir o estudo a um representante de cada grupo, obtendo resultados mais variados e menos custosos. A primeira dificuldade que aparece é a de encontrar um modo rápido de especificar a característica de interesse "massa corpórea". Após investigar o assunto o pesquisador concluiu que as variáveis peso e altura seriam dois indicadores próximos da sua característica de interesse.

Assim, o objetivo operacional passou a ser o de agrupar os indivíduos da população-alvo segundo duas variáveis facilmente mensuráveis: peso e altura.

Esta fase é a mais importante de Análise de Agrupamento, a de fixação dos critérios de homogeneidade. Critérios distintos levam a grupos homogêneos distintos e o tipo de homogeneidade depende dos objetivos a serem alcançados.

Obtenção dos Dados

Como ainda é uma fase exploratória o pesquisador decidiu usar as informações de seis pessoas de seu conhecimento como estudo piloto. A altu-

ra foi medida em centímetros e o peso em quilogramas. Os resultados estão na Tabela 1.1.

Tabela 1.1. Dados Pessoais de Seis Indivíduos do Estudo-Piloto

INDIV.	ALTURA	PESO	IDADE	INSTRUÇÃO	COR	SEXO
A	180	79	30	UNIV.	PRETA	M
B	175	75	28	UNIV.	BRANCA	M
C	170	70	20	SECUND.	BRANCA	F
D	167	63	25	UNIV.	PARDA	F
E	180	71	18	SECUND.	PARDA	M
F	165	60	28	PRIMÁRIO	BRANCA	F

Este é o material básico para a aplicação das técnicas de Análise de Agrupamento, a matriz de dados. Ela indica os valores das características por objetos de interesse. Convencionamos neste livro indicar os objetos nas linhas e as variáveis nas colunas. Veja Tabela 1.1.

Tabela 1.2. Matrizes de Dados

(a) Brutos

$$X = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

(b) Relativos (Padronizados).

$$Z = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & \cdots & Z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{np} \end{bmatrix}$$

Usualmente pretende-se agrupar objetos semelhantes segundo suas características (variáveis). Mas nada impede que o interesse seja o de agrupar variáveis segundo os valores obtidos pelos objetos.

É muito importante a definição do objeto e a correspondente atribuição do valor da característica. Por exemplo, o objeto pode ser pessoa e a variável de interesse salário. Ou o objeto pode ser família e a variável de interesse o salário do chefe. Observe que a característica é a mesma, mas associada a objetos distintos e com significado bem distinto para o processo de agrupar.

Tratamento dos Dados

No estudo-piloto os objetos serão agrupados segundo duas características com unidades distintas (centímetro e quilograma), desse modo, serão necessários alguns cuidados no instante de agrupar as informações. No momento, e sem maiores explicações, usar-se-á a padronização estatística para as duas variáveis, ou seja, subtrair a média de cada observação e dividir pelo respectivo desvio-padrão. As propriedades dessa transformação são bastante conhecidas e deixa-se aos leitores a responsabilidade inicial de procurarem entender as razões. Estes novos encontram-se na Tabela 1.2.

A importância, e os modos, da relativização dos dados serão discutidos nos Capítulos 2 e 4. Entretanto, as técnicas de Análise de Agrupamentos podem ser aplicadas a qualquer conjunto de dados, brutos ou relativizados. A escolha também depende dos objetivos. Também, fica evidente que os resultados serão distintos, e dependendo das variáveis envolvidas, o uso dos dados originais pode tornar muito difícil a interpretação do conceito de homogeneidade. Assim, o propósito dessa etapa é o de derivar a matriz de dados relativizados Z Tabela 1.1.(b), sobre a qual serão aplicadas as técnicas de Análise de Agrupamento.

Tabela 1.3. Construção dos Valores Padronizados do Peso e Altura do Estudo-Piloto

INDIVÍDUO	ALTURA	PESO	ZALT	ZPES
A	180	79	1,10	1,31
B	175	75	0,33	0,75
C	170	70	-0,44	0,05
D	167	63	-0,90	-0,93
E	180	71	1,10	0,19
F	165	60	-1,21	-1,35
MÉDIA	172,8	69,7	0,0	0,0
D. PADRÃO	6,5	7,1	1,0	1,0

Critérios de Parecença (Semelhança ou Proximidade)

Como verificar se o objeto *A* é mais parecido com *B* do que com *C*? Quando o número de atributos envolvidos é pequeno, a inspeção visual pode responder. Por exemplo, a representação gráfica da matriz *Z*, na Figura 1.1.(b), mostra que *A* está mais perto de *B* do que *C*. Observe que aqui está sendo usado o conceito usual de distância euclideana para definir a ideia de parecença. Outras definições poderiam ser usadas que não levariam às mesmas conclusões. Esta questão será o conteúdo do próximo Capítulo.

Voltando à Figura 1.1.(b), nota-se que para alguns pares de pontos ainda não é fácil fazer afirmações. Por exemplo, B está mais próximo de *A* ou *C*? Assim, torna-se importante construir um coeficiente de parecença, que quantifique essa proximidade. Aqui, seguindo o conceito natural de distância, usar-se-á a distância euclideana definida por:

$$d = \left[(z_1(\mathrm{A}) - z_1(\mathrm{B}))^2 + (z_2(\mathrm{A}) - z_2(\mathrm{B}))^2\right] \qquad (1.5.1)$$

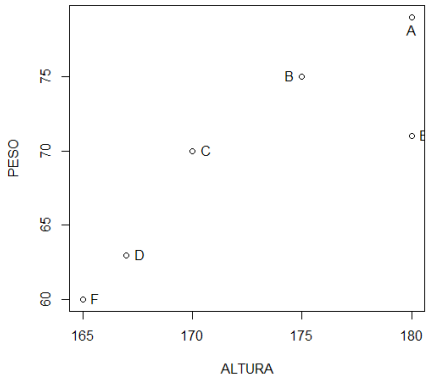
$$d = \left[(1.10 - 0.33)^2 + (1.31 - 0.75)^2\right] = 0.95$$

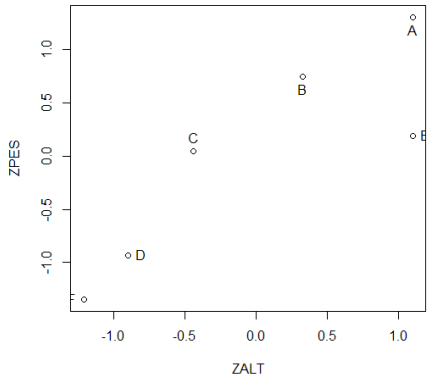
onde $z_i(.)$ indica o valor da variável $\mathbf{Z}_i$ para o ponto indicado. Aplicando esta fórmula para todos os pares da matriz **Z**, obtêm-se a matriz de parecença **D** derivada da matriz **Z**. Ela está construída no Tabela 1.2.(a). A inspeção desta matriz, além de confirmar os resultados observados na figura, explicita outras conclusões que não estavam tão claras. Por exemplo, *B* está mais próximo de *A* do que de *C*. Mais ainda, analisando apenas a matriz de similaridade chegar-se-iam aos mesmos resultados da inspeção gráfica.

Fig. 1.1. Representação Cartesiana do Peso e Altura do Estudo Piloto

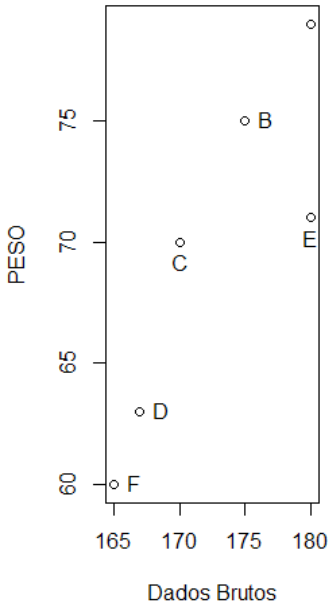
```
PESO=c( 79, 75,70,63,71,60)
ALTURA =c(180, 175,170,167,180,165)
nomes=c("A","B","C","D","E","F")
plot(ALTURA, PESO)
identify(ALTURA, PESO,#coordenadas gráficas dos pontos
nomes,#vetor de descrição dos pontos
n=6)#número de pontos a serem identificados
```

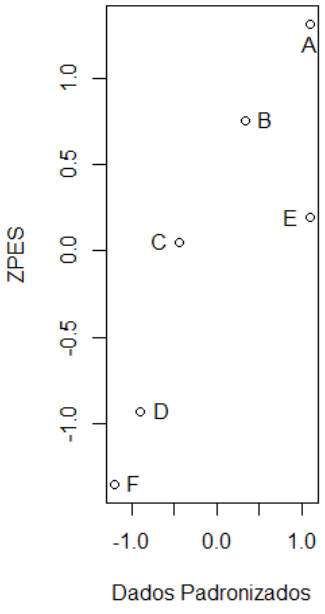
(a) Dados Brutos





a) Dados Brutos **b) Dados Padronizados**





```
par(mfrow=c(1,2))
PESO=c( 79, 75,70,63,71,60)
ALTURA =c(180, 175,170,167,180,165)
nomes=c("A","B","C","D","E","F")
plot(ALTURA, PESO, xlab="Dados Brutos")
identify(ALTURA, PESO,#coordenadas gráficas dos pontos
nomes,#vetor de descrição dos pontos

n=6)#número de pontos a serem identificados
ZALT=c( 1.10, 0.33 ,-0.44,-0.90, 1.10, -1.21)
ZPES=c( 1.31, 0.75, 0.05, -0.93, 0.19, -1.35)
plot(ZALT,ZPES, xlab="Dados Padronizados")
identify(ZALT, ZPES, nomes, n=6)
```

Tabela 1.4. Matriz de Similaridade entre os objetos do Estudo-Piloto, segundo a Distância Euclidiana dos Dados Padronizados

(a) Distância Usual

```
#######dados padronizados
library=require
library(cluster)
library(vegan)
library(ecodist)
library(MASS)
require(pvclust)

PESO=c( 1.10, 0.33,-0.44,-0.90,1.10,-1.21)
ALTURA =c(1.31, 0.75,0.05,-0.93,0.19,-1.35)
Dados=c(PESO, ALTURA)
tab.Dados=table(Dados)
x=matrix(Dados,nrow=6,ncol=2,byrow=TRUE)
y=c("A","B","C","D","E","F" )
colnames(x)<-y
x=t(x)
x=matrix(x,nrow=6,ncol=2,byrow=TRUE)
x=data.frame(PESO, ALTURA)
x
attach(x)
options(digits=2)
dist1<-distance(x,"euclidean")
dist1
```

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0.00 & & & & & \\ 0.95 & 0.00 & & & & \\ 1.99 & 1.05 & 0.00 & & & \\ 3.00 & 2.08 & 1.09 & 0.00 & & \\ 1.12 & 0.95 & 1.54 & 2.29 & 0.00 & \\ 3.52 & 2.60 & 1.60 & 0.52 & 2.77 & 0.00 \end{bmatrix} \end{array}$$

(b) Distância Reduzida

```
d=((1.10-0.33)^2+(1.31-0.75)^2)   #AB
sqrt(d/2)

d1=((1.10-(-0.44))^2+(1.31-0.05)^2)   #AC
sqrt(d1/2)

d2=((1.10-(-0.90))^2+(1.31-(-0.93))^2)   #AD
sqrt(d2/2)

d3=((1.10-1.10)^2+(1.31-0.19)^2)   #AE
sqrt(d3/2)

d4=((1.10-(-1.21))^2+(1.31-(-1.35))^2)   #AD
sqrt(d4/2)
```

$$D = \begin{array}{c} \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0.67 & & & & \\ 1.41 & 0.74 & & & \\ 2.12 & 1.47 & 0.77 & & \\ 0.79 & 0.67 & 1.09 & 1.62 & \\ 2.49 & 1.84 & 1.13 & 0.37 & 1.96 \end{bmatrix} \end{array}$$

Outra vantagem ocorre quando existem muitos atributos classificatórios onde torna-se inviável a inspeção gráfica, mas é possível criar coefi-

cientes de parecença entre os objetos. Um exemplo simples é a generalização da distância euclidiana para um espaço de dimensão *p*, a Saber.

$$d_{(A,B)} = \left[\sum_{i=1}^{p} \frac{(z_1(\mathrm{A}) - z_1(\mathrm{B}))^2}{p} \right]^{1/2} \qquad (1.5.2)$$

Neste livro, a menos que seja especificado, sempre será usada esta última expressão para a distância euclideana. No Tabela 1.2.(b), aparece a distância reduzida para o Estudo Piloto. Aproveitou-se também para eliminar uma linha e uma coluna da matriz, por terem significados óbvios.

Aplicação da Técnica de Agrupamento

Aplica-se a Análise de Agrupamento nas mais diversas áreas. Os resultados deste conjunto de técnicas podem contribuir para a definição de um esquema formal de agrupamento, podem também sugerir um conjunto de regras para agrupar novos objetos em novos grupos com fins de diagnóstico, apresentar sugestões de modelos estatísticos para descrever populações, encontrar objetos que representa grupos ou classes.

A escolha de um particular algoritmo de agrupamento exige o conhecimento de suas propriedades aliado aos objetivos da pesquisa. Neste exemplo ilustrativo supor-se-á, sem mais explicações, que a escolha recaiu no método da média das distâncias (M.M.D.). Este é um processo hierárquico e em cada passo diminui uma dimensão da matriz de parecença pela reunião de pares semelhantes até reunir todos os pontos em um único grupo.

Abaixo aparecem os diversos passos da aplicação do método ao exemplo ilustrativo.

Tabela 1.5. Matrizes de similaridade, para os Diversos Passos do Método das Médias das Distâncias

(a) Passo 0

	A	B	C	D	E
B	0,67	– 0,74	–	–	–
C	1,41	1,47	– 0,77	–	–
D	2,12	0,67	1,09	–	–
E	0,79	1,84	1,13	1,62	–
F	2,49				1,96
				0,37	

(b) Passo 1. Agrupar D com F

	A	B	C	E
B	0,67	–	–	–
C	1,41	0,74	–	–
E	0,79	0,67	1,09	
DF	2,30	1,66	0,95	1,79

(c) Passo 2. Agrupar A com B

	C	E	DF
E	1,09	– 1,79	–
DF	0,95	0,73	–
AB	1,08		1,98

(d) Passo 3. Agrupar E com AB

	C	DF
DF	0,95	–
ABE	1,08	1,92

(e) Passo 4. Agrupar C com DF

	CDF
ABE	1,64

Passo 1. Observando a matriz de similaridade repetida na Tabela 1.3.(a), nota-se que os indivíduos mais próximos de *D* e *F*, cuja distância entre eles é 0,37. Assim, os dois pontos são agrupados em um único, obtendo desse modo 5 grupos.

A, B, C, E, (DF)

É necessário reconstruir a nova matriz de similaridade. Como os pontos A, B, C e E não sofreram alterações as distâncias entre eles também continuam as mesmas. Veja no Tabela 1.1.(b) os resultados. É necessário definir a distância entre o conjunto (DF) e os demais pontos.

É aqui que a maioria dos métodos se diferencia e algumas das alternativas serão abordadas no Capítulo 3. O M.M.D define a distância entre dois grupos com a média entre os valores individuais dos objetos de um dos grupos com os do outro. Assim:

$$d(A, DF) = \frac{(\mathrm{d(A,D)} + \mathrm{d(A,F)})}{2} = \frac{(2.12 + 2.49)}{2} = 2.30$$

$$d(B, DF) = \frac{(\mathrm{d(B,D)} + \mathrm{d(B,F)})}{2} = \frac{(1.47 + 1.84)}{2} = 1.66$$

$$d(C, DF) = \frac{(0.77 + 1.11)}{2} = 0.95$$

$$d(\mathrm{E}, DF) = \frac{(1.62 + 1.96)}{2} = 1.79$$

Com a obtenção da matriz de parecença (Tabela 1.3.(b)), conclui-se o passo 1, que reuniu os pontos *D* e *F*, num nível igual a 0,37.

Passo 2. Analisando a nova matriz de similaridade nota-se que existem dois pares com a mesma proximidade *A* com *B* e *B* com E. Embora raro de acontecer na prática, o processo recomenda selecionar aleatoriamente um dos pares e criar o novo grupo. Porém, os pacotes computacionais, por facilidade de programação, escolhem o primeiro par que aparece para agrupar. Desse modo, neste passo agrupa-se *A* com *B*, obtendo-se os seguintes grupos: *C*, *E*, (*DF*) e (*AB*). Como no caso anterior, as distâncias entre *C*, *E* e (*DF*)

não se alteram, conforme aparece na Tabela (c) da Tabela 1.3. As distâncias de (*AB*) com os demais pontos serão:

$$d(\mathrm{C}, AB) = \frac{(\mathrm{d}(\mathrm{C},\mathrm{A}) + \mathrm{d}(\mathrm{C},\mathrm{B}))}{2} = \frac{(1.41 + 0.74)}{2} = 1.08$$

$$d(\mathrm{E}, \mathrm{AB}) = \frac{(\mathrm{d}(\mathrm{E},\mathrm{A}) + \mathrm{d}(\mathrm{E},\mathrm{B}))}{2} = \frac{(0.79 + 0.67)}{2} = 0.73$$

$$d(DF, AB) = \frac{[\mathrm{d}(\mathrm{D},\mathrm{A}) + \mathrm{d}(\mathrm{D},\mathrm{B}) + \mathrm{d}(\mathrm{F},\mathrm{A}) + \mathrm{d}(\mathrm{F},\mathrm{B})]}{4} =$$

$$= \frac{(2.12 + 1.47 + 2.49 + 1.84)}{4} = 1.98$$

Termina aqui o passo 2 com *A* sendo reunido a *B* ao nível 0,67.

Passo 3. Reunir *E* com (AB) ao nível 0,73 de similaridade, obtendo-se os grupos *C*, (*DF*) e (*ABE*). Recalculando as distâncias necessárias tem-se

$$d(\mathrm{C}, ABE) = \frac{[\mathrm{d}(\mathrm{C},\mathrm{A}) + \mathrm{d}(\mathrm{C},\mathrm{B}) + \mathrm{d}(\mathrm{C},\mathrm{E})]}{3} = \frac{(1.41 + 0.74 + 1.09)}{3} = 1.08$$

$$d(DF, \mathrm{ABE}) = \frac{[\mathrm{d}(D,\mathrm{A}) + \mathrm{d}(D,\mathrm{B}) + d(\mathrm{D},\mathrm{E}) + \mathrm{d}(\mathrm{F},\mathrm{A}) + \mathrm{d}(\mathrm{F},\mathrm{B}) + \mathrm{d}(\mathrm{F},\mathrm{E})]}{6}$$

$$= \frac{(2.12 + 1.47 + 1.62 + 2.49 + 1.84 + 1.96)}{6} = 1.92$$

.

Com a construção da matriz (d), Tabela 1.3, encerra-se este passo.

Passo 4. Reunir *C* com (*DF*), ao nível 0,95, obtendo-se a partição (*ABE*, *CDF*). A distância entre os dois grupos será:

$$d(ABE, CDF) = \frac{[\mathrm{d(A,C)} + \mathrm{d}(A, D) + \mathrm{d}(A, F) + \mathrm{d(B,C)} + \mathrm{d(B,D)} +}{9}$$
$$\frac{+ \mathrm{d(B,F)} + \mathrm{d(E,C)} + d(\mathrm{E,D}) + \mathrm{d(E,F)}]}{9} = 1.64$$

Conclui-se escrevendo a matriz (c) do Tabela 1.3.

Passo 5. O processo encerra reunindo num único grupo os conjuntos ABE e CDF, que são iguais a um nível 1,64 de parecença.

Como já foi dito, existem diferentes métodos para agrupar elementos que serão discutidos futuramente. O importante é conhecer suas propriedades, qualidades e deficiências, pois irá ajudar à escolha daquele que melhor responde aos objetivos do trabalho.

Apresentação dos Resultados

As etapas descritas na seção anterior, embora instrutivas acerca do processo de agrupar, não facilitam a interpretação dos resultados. Necessita-se de instrumentos mais apropriados e um deles é o resumo das etapas descritivas acima. A Tabela 1.3 mostra em cada etapa a formação dos grupos e os respectivos níveis em que eles são formados. É muito importante entender o significado desse nível e sugerimos ao leitor refletir um pouco mais acerca desse conceito. Dificilmente dois objetos serão exatamente iguais, mas sendo condescendentes no critério de "igual" pode-se aceitar que eles são "parecidos". Assim, os objetos *D* e *F* podem ser considerados semelhantes e esse grau de semelhança é avaliado com a nota 0,37. Observe que não existe um padrão com o qual podemos comparar este número para afirmar se é muito ou pouco. O conhecimento do processo e a familiaridade com as grandezas envolvidas é que irão ajudar. Duplicando esse nível, ou seja, relaxando um pouco mais o conceito de semelhança, concluir-se-ia que além de *D* e *F* também seriam considerados semelhantes entre si, os objetos *A*, *B* e *F*. E assim por diante seriam interpretados os dados da tabela mencionada.

Tabela 1.6. Resumo do M.M.D Aplicado aos Dados do Estudo-Piloto

Passo	Junção	Nível
1	D,F	0,37
2	A,B	0,67
3	AB,E	0,73
4	C,DF	0,95
5	ABE,CDF	1,64

A tabela 1.3 resumo, possui uma representação gráfica muito útil e muito usada em A.A, conhecida por dendrograma (gráfico em forma de árvore), ilustrado na Figura 1.2. A escala vertical à esquerda, indica o nível de similaridade. No eixo horizontal são marcados os objetos, numa ordem conveniente, as linhas verticais partindo dos objetos têm altura correspondente ao nível em que os objetos são considerados semelhantes.

A grande vantagem do dendrograma é mostrar graficamente o quanto é necessário "relaxar" o nível de parecença para considerar grupos próximos. Observando a Figura 1.2., notamos que o maior salto é observado na última etapa, sugerindo a existência de dois grupos homogêneos: (*A*, *B*, *E*) e (*C*, *D*, *F*).

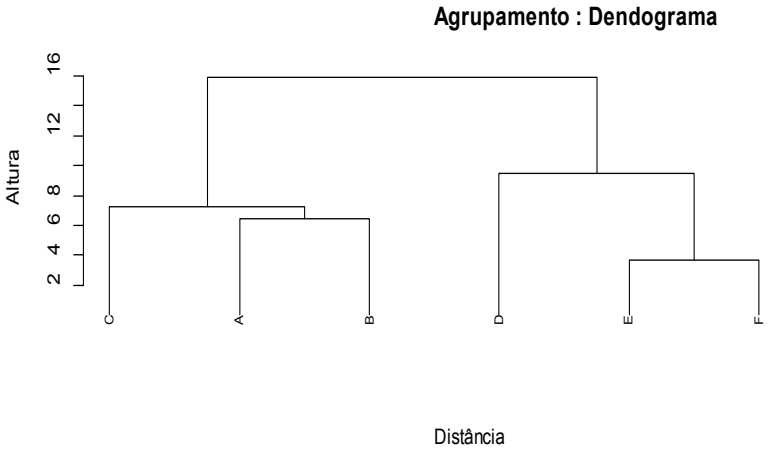
Tendo obtido esses resultados, é conveniente voltar aos dados para uma melhor compreensão do processo de agrupar. Baseado no dendrograma é conveniente reescrever os dados originais e a matriz de similaridade na ordem produzida pelo método de agrupamento. Estes procedimentos foram feitos na Tabela 1.4.

Tabela 1.7. Reordenação dos Dados do Estudo-Piloto de Acordo com o Dendrograma

(**Fig. 1.2.**)

Indivíduo	Z altura	Z peso	Altura	Peso
A	1.10	1.31	180	79
B	0.33	0.75	175	75
C	1.10	0.19	180	71
D	-0.44	0.05	170	70
E	-0.90	-0.93	167	63
F	-1.21	-1.35	165	60

Fig. 1.2. Dendrograma do M.M.D Aplicado aos Dados do Estudo-Piloto



Reordenação da Matriz de Similaridade de Acordo com o Dendrograma

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0 & & & & & \\ 0.67 & 0 & & & & \\ 0.79 & 0.67 & 0 & & & \\ 1.41 & 0.74 & 1.09 & 0 & & \\ 2.12 & 1.47 & 1.62 & 0.77 & 0 & \\ 2.49 & 1.84 & 1.96 & 1.13 & 0.37 & 0 \end{bmatrix} \end{array}$$

Os resultados indicam a existência de dois grupos de pessoas: as pequenas e as grandes. Observando a matriz de similaridade, e a Figura 1.1., nota-se que o objeto *B* está mais próximo de *C* do que de *E*, e usando dois grupos apenas, eles ficariam separados. Isso deve-se ao particular método de agrupamento usado e fatos como esse é que inspiram a criação de outros critérios de agrupar e que serão discutidos nos próximos Capítulos.

Avaliação e Interpretação dos Resultados

Dendrograma é uma representação matemática e ilustrativa de todo o procedimento de agrupamento através de uma estrutura de árvore.

Os nós do dendrograma representam agrupamentos, e nós são compostos pelos grupos e ou objetos (grupos formados apenas por ele mesmo) ligados a ele (nó). Se cortarmos o dendrograma em um nível de distância desejado, obteremos uma classificação dos números de grupos existentes nesse nível e dos indivíduos que os formam. O número de grupo dos indivíduos é obtido pelo corte do dendrograma em um nível desejado e então cada componente conectado forma um grupo.

O dendrograma pode ser considerado a representação simplificada da matriz de similaridade, e, portanto, se coloca a pergunta: é uma "boa" simplificação? Uma das maneiras de responder é verificar a capacidade do dendrograma em reproduzir a matriz de similaridade. O primeiro passo para isso é construir a matriz cofenética, que é a matriz de distância entre os objetos obtidos a partir do dendrograma. Por exemplo, a distância entre os pontos *A* e *C* é dada pelo nível em que os dois são agrupados e que é 1,64 pelo dendrograma. Já a distância entre *A* e *E* será 0,73. Procedendo de modo análogo para os demais pontos, constrói-se a matriz cofenética da Tabela 1.5.

Matriz Cofenética Baseada no Dendrograma da Figura 1.2

$$C = \begin{bmatrix} & & & & & & \\ 0.67 & & & & & & \\ 1.64 & 1.64 & & & & & \\ 1.64 & 1.64 & 0.95 & & & & \\ 0.73 & 0.73 & 1.64 & 1.64 & & & \\ 1.64 & 1.64 & 095 & 0.37 & 1.64 & & \end{bmatrix}$$

Tabela 1.8. Cálculo do Coeficiente de Correlação Cofenética

PAR	S	C	PAR	S	C	PAR	S	C
AB	0.67	0.67	BC	0.74	1.64	CE	1.09	1.64
AC	1.41	1.64	BD	1.47	1.64	CF	1.13	0.95
AD	2.12	1.64	BE	0.67	0.73	DE	1.62	1.64
AE	0.79	0.73	BG	1.84	1.64	DF	0.37	0.37
AF	2.49	1.64	CD	0.77	0.95	EF	1.96	1.64

s ... distância da matriz de similaridade.

c ... distância da matriz cofenética.

cc = corr(s, c) = 0,75 = 1,27 = 1,28 s_s = 0,63 s_c = 0,48

Deve-se agora verificar a proximidade das duas matrizes, e esta é fornecida pelo coeficiente de correlação entre os valores da matriz de similaridade e os correspondentes da matriz cofenética. Este índice é chamado Coeficiente de Correlação Cofenética. As operações necessárias aos cálculos estão na Tabela 1.6. No caso do Estudo-Piloto este indicador é 0,75. Quanto mais próximo da unidade melhor será a representação, e quanto mais próximo de zero será pior. O valor observado 0,75 é alto ou baixo? Responder a isto é tão difícil como responder, na maioria das situações, o que é um alto coeficiente de correlação entre duas variáveis. Depende da área de estudo e de padrões que vão se desenvolvendo com a prática. Pode-se adiantar que em Análise de Agrupamento, algo em torno de 0,8 já pode ser considerado bom ajuste.

Analisando todos os resultados do exemplo ilustrado, poder-se-ia concluir que a amostra piloto sugere dois tipos de indivíduos: pequenos e grandes. Para continuar o estudo retrospectivo bastaria escolher (ou sortear) apenas duas pessoas: Uma do conjunto (*A*, *B*, *E*) e outra de (*C*, *D*, *F*), e teríamos elementos "representativos" do grupo, segundo os critérios de altura e peso, na crença de que essas variáveis sejam substitutas da característica de interesse.

Sumário

As técnicas de Análise de Agrupamento exigem de seus usuários a tomada de uma série de decisões interdependentes que requerem o conhecimento das propriedades dos diversos algoritmos à disposição. Algumas dessas decisões envolvem conteúdos mais metodológicos, enquanto que outras mais, o caráter técnico. Deve-se iniciar explicitando claramente o objeto e os objetivos desejados com a aplicação da Análise de Agrupamento. Também devem ser explicitados os critérios (variáveis) que irão definir as semelhanças entre os objetos. Muitas vezes essas variáveis necessitam de transformações para tornarem-se mais adequadas aos objetivos enunciados. Obtida a matriz de dados transformados o próximo passo é a escolha de um coeficiente de *semelhança* entre os objetos. Em seguida escolher o método de obter os grupos homogêneos e a apresentação dos resultados obtidos. Finalmente, avaliar e interpretar, à luz dos objetivos, os resultados produzidos. Outras questões também aparecem como as de encontrar quantos grupos homogêneos existem nos dados.

Exercícios

1. Usando as informações sobre instrução e sexo dos dados na Tabela 1.1;
   a. procure construir uma matriz de parecença entre os objetos;
   b. construa um dendrograma descrevendo o processo de agrupamento.

Estudando a multicolinearidade entre 5 variáveis encontrou-se a seguinte matriz de correlação entre elas:

$$R = \begin{bmatrix} 1.00 & & & & \\ 0.38 & 1.00 & & & \\ 0.19 & 0.54 & 1.00 & & \\ 1.16 & 0.60 & 0.22 & 1.00 & \\ 0.06 & 0.46 & 0.13 & 0.94 & 1.00 \end{bmatrix}$$

2. Proponha e realize um procedimento análogo ao descrito neste Capítulo para agrupar as variáveis.

Um diretor de Marketing deseja agrupar 6 municípios de acordo com a porcentagem das vendas de bolachas doces (X) e salgadas (Y), indicadas no Tabela abaixo:

Município	A	B	C	D	E	F
X(%Doces)	9	14	13	10	12	14
Y(%Salgados)	19	22	22	18	31	20

a. Construa conglomerados através da inspeção gráfica;
b. Construa uma matriz de parecença entre os municípios;
c. Aplique o método M.M.D;
d. Construa o dendrograma;
e. Calcule a correlação cofenética;
f. Interprete os resultados.

CAPÍTULO 2

Medidas de Distância: Similaridade e Dissimilaridade (Parecença)

Um conceito fundamental na utilização das técnicas de Análise de Agrupamento é a escolha de um critério que meça a distância entre dois objetos ou que quantifique o quanto eles são parecidos. Esta medida será chamada de **coeficiente de parecença**. Cabe observar que tecnicamente pode-se dividir em duas categorias: medidas de similaridade e de dissimilaridade. Na primeira, quanto maior o valor observado menos parecido (mais dissimilares) serão os objetos. Coeficiente de correlação é um exemplo de medida de similaridade, enquanto que distância euclideana é um exemplo de dissimilaridade.

De um modo geral, as medidas de similaridade e de dissimilaridade e vice-versa são interrelacionadas e, facilmente, transformáveis entre si. Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na Literatura. Tais coeficientes podem ser, facilmente, convertidos para coeficientes de dissimilaridade: se a similaridade for denominada **s**, a medida de dissimilaridade será o seu complementar (**1** – **s**).

A maioria dos algoritmos de Análise de Agrupamento estão programados para operarem com o conceito de distância (dissimilaridade), exigindo do usuário o esforço da transformação (veja Exemplo 2.1.). Devido a essas duas observações neste livro, não será feita a distinção, a menos que a particular situação assim o exija. Deste modo, para facilitar a linguagem e chamar a atenção para a diferença, usar-se-á o termo matriz de parecença para indicar semelhança ou distância entre objetos.

Exemplo 2.1

Estudando o comportamento de 3 variáveis X_1, X_2 e X_3, usou-se o coeficiente de correlação como coeficiente de parecença (similaridade), com os seguintes resultados:

$$S = \begin{array}{c} \\ X1 \\ X2 \\ X3 \end{array} \begin{array}{c} \begin{array}{ccc} X1 & X2 & X3 \end{array} \\ \begin{bmatrix} 1.00 & & \\ 0.70 & 1.00 & \\ 0.60 & 0.75 & 1.00 \end{bmatrix} \end{array}$$

Note-se que as duas variáveis com comportamento mais parecidos são X_2 e X_3 e que possuem a maior correlação entre eles. Já X_1 e X_2 seriam as menos similares. Com a transformação d(.,.) = 1 – corr(.,.) obtém-se a matriz de dissimilaridades.

$$D = \begin{array}{c} \\ X1 \\ X2 \\ X3 \end{array} \begin{array}{c} \begin{array}{ccc} X1 & X2 & X3 \end{array} \\ \begin{bmatrix} 1.00 & & \\ 1.70 & 1.00 & \\ 1.60 & 0.25 & 1.00 \end{bmatrix} \end{array}$$

Indicando que quanto maior o valor observado, menos parecidos são os objetos.

O coeficiente de correlação negativo, às vezes, pode ter o mesmo significado que o positivo, ou seja, indica o mesmo grau de similaridade, nesses casos usa-se a transformação d(.,.) = 1 – |corr(.,.)|, produzindo

$$D = \begin{array}{c} \\ X1 \\ X2 \\ X3 \end{array} \begin{array}{c} \begin{array}{ccc} X1 & X2 & X3 \end{array} \\ \begin{bmatrix} 1.00 & & \\ 0.30 & 1.00 & \\ 0.40 & 0.25 & 1.00 \end{bmatrix} \end{array}$$

Muitas vezes os coeficientes de parecença não são definidos de um modo muito preciso e não definem uma métrica sobre o espaço dos objetos. Isso pode levar a alguns problemas sérios de interpretação. Aqui não serão aprofundadas estas questões e os interessados encontrarão mais informações em Späth (1980).

Como já foi discutido anteriormente, também a escolha das variáveis influi na definição das semelhanças. Neste Capítulo, serão definidas as variáveis e estudar-se-á diversos modos de definir parecença. Apenas, para facilitar a apresentação, serão tratadas inicialmente as variáveis quantitativas, depois as qualitativas nominais e as qualitativas ordinais. Finalmente sugerem-se maneiras para tratamento de variáveis mistas.

Seja M um conjunto, uma métrica em M é uma função d: MxM $\square$, tal que para quaisquer i, j, z $\square$ M, tenhamos:

1. d (i, j) = d (j, i) (simétrica). Esta regra afirma que a distância entre dois elementos não varia, não importando o ponto a partir do qual ela é medida. Por isto, a matriz de D é mostrada sendo triangular inferior. Tendo em vista que ela é simétrica, os valores acima da diagonal estão implicitamente definidos;
2. d (i, j) > 0, se i≠ j; para todos os objetos;
3. d (i, j) = 0, se e somente se, i = j e
4. d (i, j) ≤ d (i, z) + d (z, j). Esta é conhecida como a desigualdade triangular, e basicamente especifica que a menor distância entre dois pontos é uma reta.

Além disso, é esperado que d (i, j) aumente quando a dissimilaridade entre i e j aumentar.

Coeficientes de Parecença para Atributos Quantitativos

A distância euclidiana é a métrica de maior utilização como função de agrupamento e apresenta grande facilidade de cálculo.

Essa função define a distância entre duas parcelas como uma simples soma de p diferenças ou desvios ao quadrado. Geralmente, esta métrica é utilizada para variáveis padronizadas.

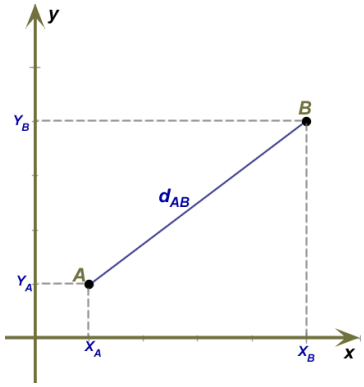


Fig. 2.1. Distância euclidiana entre os pontos A e B

Medidas Derivadas da Distância Euclideana

Considere o vetor **x** de coordenadas reais, (x_1, x_2, ..., x_p), como descritor dos objetos que serão investigados os assemelhamentos. A medida mais conhecida para indicar a proximidade entre os objetos A e B e a distância euclideana (D.E):

$$d(A,B) = \sqrt{\sum_{i=1}^{p} (X_i(A) - X_i(B))^2} \qquad 2.1.1$$

ou

$$d_{ij} = \sqrt{\sum_{f=1}^{p} (x_{if} - x_{jf})^{1/2}}$$

Ou em linguagem matricial

$$d(A,B) = \left((X(A) - X(B))'(X(A) - X(B)\right)^{1/2} \qquad 2.1.2$$

Uma primeira medida derivada desta, e muito usada em Análise de Agrupamento, é o coeficiente da distância euclideana média (D.E.M), onde a soma das diferenças ao quadrado é dividida pelo número de coordenadas envolvidas, isto é:

$$d(A,B) = \left[\frac{\sum_{i=1}^{p} (X_i(A) - X_i(B))^2}{p}\right]^{1/2} \qquad 2.1.3$$

Esta última expressão é apenas um reescalonamento da distância anterior, possuindo as mesmas propriedades, portanto produz os mesmos resultados se submetidos às mesmas técnicas de Análise de Agrupamento. Entretanto, este último coeficiente possui duas propriedades interessantes. A primeira é que ela pode ser usada na ausência de dados para algumas coordenadas (*missing values*). A segunda, de ordem prática, permite acumular evidências empíricas sobre níveis de parecença.

Para ilustrar os diversos coeficientes serão usados os mesmos dados do exemplo do Capítulo 1.

Exemplo 2.2

Na Tabela 2.1, estão apresentados os dados das variáveis quantitativas usadas no Capítulo 1.

Tabela 2.1

Indivíduo	Altura	Peso	Idade	Z1	Z2	Z3
A	180	79	30	1.10	1.31	1.08
B	175	75	25	0.33	0.75	0.00
C	170	70	28	-0.44	0.05	0.65
D	167	63	21	-0.90	-0.93	-0.86
E	180	71	18	1.10	0.19	-1.52
F	165	60	28	-1.21	-1.35	0.65
Média	172.8	69.7	25.0	0,0	0.0	0.0
Desvio padrão	6.5	7.1	4.6	1.0	1.0	1.0

(a) Distância Euclideana

Usando as variáveis altura e peso, a DE entre A e B é

$d_2(A,B) = [(180 - 175)^2 + (79 - 75)^2]^{1/2} = (41)^{1/2} = 6.40$ Usando como critérios as duas variáveis anteriores e mais a idade, a D.E passa a ser

$$d_3(A,B) = [(180 - 175)^2 + (79 - 75)^2 + (30 - 25)^2]^{1/2} = (66)^{1/2} = 8.12$$

(b) Coeficiente Médio da Distância Euclideana

Aplicando a expressão 2.2. Obtém-se

$$d_2(A,B) = \left(\frac{41.2}{2}\right)^{1/2} = 4.53$$

$$d_2(A,B) = \left(\frac{66}{3}\right)^{1/2} = 4.69$$

Observe que neste último coeficiente embora aumente o número de coordenadas, a magnitude dos coeficientes são comparáveis.

(c) Distância Euclideana Padronizada

Observando a distância do Exemplo 2.2.(a) nota-se que estão somadas grandezas não comparáveis (cm, kg e anos), mais ainda, a mudança de uma das unidades, pode alterar completamente o significado e o valor do coeficiente (veja mais explicação no Capítulo 4). Essa é uma das razões da padronização da variável usada no Capítulo 1.

Assim o uso da transformação.

$$Z_i = \frac{X_i(\cdot) - \bar{X}_i}{S_i} \qquad 2.3$$

Onde $\bar{X}_i$ e S_i, indicam respectivamente a média e o desvio padrão de i-ésima coordenada, é um dos modos para evitar esse inconveniente. Feita a transformação, a distância euclideana passa

a ser:

$$d(A,B) = \left[\sum_{i=1}^{p} (z_i(\mathrm{A}) - z_i(\mathrm{B}))^2\right]^{1/2} \qquad 2.4$$

substituindo por (2.3) obtém-se

$$d(A,B) = \left[\sum_{i=1}^{p} \frac{(z_i(\mathrm{A}) - z_i(\mathrm{B}))^2}{S_i}\right]^{1/2} \qquad 2.5$$

que é a soma dos desvios padronizados. É fácil verificar que a expressão acima pode ser escrita da seguinte forma em notação vetorial:

$$d(A,B) = \left[(X(A) - X(B))' D^{-1} (X(A) - X(B))\right]^{1/2} \qquad 2.6$$

onde **D** é uma matriz diagonal, tendo como i-ésimo componente a variância S_i^2 , isto é,

$$D = diag(\mathrm{s}_1^2, \mathrm{s}_2^2, \cdots, \mathrm{s}_p^2) \qquad (2.7)$$

De modo análogo, pode-se definir a distância euclideana média:

$$d(A,B) = \left[\frac{((X(A) - X(B))' D^{-1} (X(A) - X(B))}{p}\right]^{1/2} \qquad (2.8)$$

Exemplo 2.2 (Continuação)

(d) Coeficiente da Distância Euclideana Padronizada

Da Tabela 2.1. obtém-se:

$$d_2(A,B) = \left[(1.10 - 033)^2 + (1.31 - 0.75)\right]^{1/2} = 0.91$$

$$d_2(A,B) = \left[\left(\frac{180 - 175}{6.5}\right)^2 + \left(\frac{79 - 75}{7.1}\right)^2\right]^{1/2} = 0.91$$

Ou ainda

$$d_2(A,B) = \left[(180 - 175.79 - 75)^2 \begin{bmatrix} 42.25 & 0 \\ 0 & 50.41 \end{bmatrix}^{-1} \begin{bmatrix} 180 - 175 \\ 79 - 75 \end{bmatrix}\right]^{1/2}$$

$$= \left[(5.4) \begin{bmatrix} 1/42.25 & 0 \\ 0 & 1/50.41 \end{bmatrix} \begin{bmatrix} 5 \\ 4 \end{bmatrix}\right] = 0.91$$

Propõe-se ao leitor verificar que a distância com 3 coordenadas será $d_3(A,B)$= 1,44. Do mesmo modo as distâncias euclideanas médias serão $d_2(A,B)$ = 0,67 e $d_3(A,B)$ = 0,83.

(e) Distância Euclideana Ponderada

Na realidade este último coeficiente de parecença está associado a uma questão frequente de A.A., que é o da ponderação das variáveis, ou seja, o de dar mais peso para variáveis que o pesquisador julga mais importante para

definir semelhança, veja (EVERITT, 1974, p.50). Pode-se criar pesos arbitrários para a diagonal da matriz D, ou então criar uma matriz D, baseada em critérios estatísticos. Assim, dada uma matriz B de ponderação, define-se

$$D_{(i,j)} = \left[(x_i - x_j)' j (x_i - x_j)\right]^{1/2} \qquad (2.9)$$

$$D(A,B) = \left[(x(A) - x(B))' B (x(\mathrm{A}) - x(B))\right]^{1/2}$$

como sendo a distância ponderada por ***B ou j***. Os casos particulares mais importantes são:

v. B = I, a ponderação é a matriz identidade, tem-se então a distância euclideana usual;

vi. $B = \left[diag\ (s_1^2, s_2^2, \cdots, s_p^2\right]^{-1}$, e tem-se a distância das variáveis padronizadas;

vii. B = V^{-1}, onde V é matriz de covariâncias, tem-se então a "distância de Mahalanobis".

Esta última distância $B = \left[\mathrm{diag}\ (s_1^2, s_2^2, \cdots, s_p^2\right]^{-1}$ além de ponderar pela variabilidade de cada uma das componentes, leva em conta também o grau de correlação entre elas. Este fato torna muito difícil a interpretação de resultados baseados neste coeficiente de parecença. Sugere-se ao leitor a análise cuidadosa do exercício 1, para melhor entendimento desta distância.

Distância de Mahalanobis também chamada distância generalizada. Esta medida, ao contrário das apresentadas anteriormente, considera a matriz de covariância Σ para o cálculo das distâncias: Esta é a distância que leva em consideração a estrutura de correlação existente nos dados.

A distância generalizada D^2 de Mahalanobis também pode ser usada como técnica de comparação quando na separação entre diversos grupos, permitindo avaliar a extensão e a direção dos afastamentos entre os valores médios das variáveis usadas na discriminação. As diferenças entre cada par de grupos que estão sendo comparados são, assim, examinados simultaneamente por meio das diversas variáveis que podem ser correlacionadas, de modo que, a informação fornecida por uma delas pode não ser independente da fornecida pelas demais.

O valor numérico da maior separação possível entre dois grupos quaisquer é chamado distância generalizada entre os grupos e mede em escala

independente da originalmente utilizada para as várias variáveis, a clareza das disjunções entre elas.

Assim, o valor da distância generalizada D^2, ligando dois grupos, é um número puro, com propriedades da distância comum e mede a extensão com que diferem entre si em tamanho e forma.

A distância Generalizada de Mahalanobis entre os grupos i e j é usualmente estimada segundo (RAO, 1952) por:

$D^2 = (X_i - X_j)'.\sum^{-1}.(X_i - X_j)$, em que $D^2 = d(X_i - X_j)$; $\sum^{-1}$ é a inversa da matriz de covariância residual de X, e D^2 tem a característica de ser invariante para qualquer transformação linear não-singular. Ao contrário da distância euclidiana, D^2 pode ser utilizada quando existe correlação entre as variáveis, sendo os coeficientes de correlação nulos, o valor de D^2 equivale à distância euclidiana para variáveis padronizadas.

Pode ser mostrado que as superfícies onde as distâncias de Mahalanobis são constantes são elipsóides centradas na média do espaço amostral. No caso especial em que as características são não correlacionadas, estas superfícies são esteróides, tal qual no caso da distância euclidiana.

O uso da distância de Mahalanobis corrige algumas das limitações da distância euclidiana, pois leva em consideração automaticamente a escala dos eixos coordenados e leva em consideração a correlação entre as características. Entretanto, como não existe almoço grátis, há um preço (alto) a se pagar por estas vantagens. As matrizes de covariância podem ser difíceis de determinar e a memória e o tempo de computação crescem de forma quadrática com o número de características.

Pode-se usar também a métrica de Mahalanobis para medir a distância entre um elemento x e um cluster de elementos cuja média é dada por e que possua uma matriz de covariância dada por **Σ**. Neste caso, a distância é dada pela fórmula

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

Conceitualmente, é como se estivéssemos avaliando a pertinência de um elemento não só por sua distância ao centro do cluster, mas também pela distribuição do mesmo, determinando assim a distância do elemento *x* em termos de uma comparação com o desvio padrão do cluster. Quanto maior for o valor desta métrica, maior o número de desvios padrões que um

elemento está distante do centro do cluster e menor sua chance de ser um elemento do mesmo.

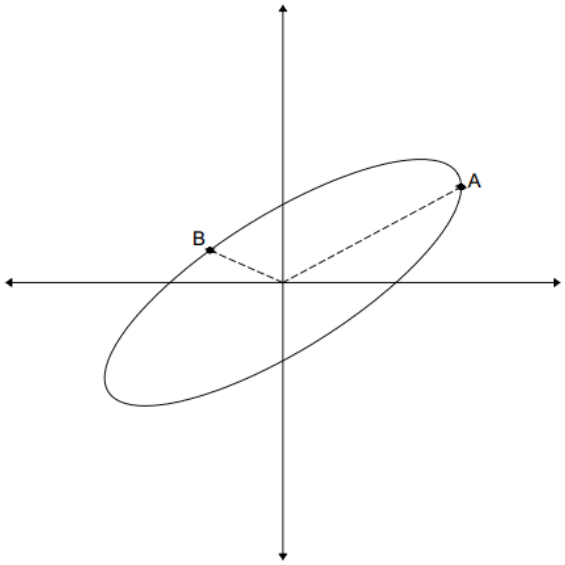


Fig. 3.4. Diagrama demostrando a distância de Mahalanobis; um conjunto da isodistância em duas dimensões.

Exemplo 2.2

(d) Para construir a distância de Mahalanobis encontra-se primeiro a matriz de covariâncias. Para as variáveis altura e peso, tem-se

$$V = Cov(X_1, X_2) = \begin{bmatrix} 42.25 & 40.11 \\ 40.11 & 50.41 \end{bmatrix}$$

Daqui observa-se que o coeficiente de correlação será

$$r = 0.8692$$

o que mostra forte correlação positiva entre as duas variáveis. Calculando a inversa obtém-se

```
V <- matrix(c(42.25,40.11,40.11,50.41), nrow = 2, ncol = 2, byrow =
TRUE,
            dimnames = list(c("row1", "row2"),
                            c("V.1", "V.2")))
V
solve(V)   # matriz inversa
```

$$V^{1} = \begin{bmatrix} 0.0968 & -0.0770 \\ -0.0770 & 0.0811 \end{bmatrix}$$

Assim a distância de Mahalanobis entre *A* e *B* será

$$d_{M} = \begin{bmatrix} \\ (5,4) \end{bmatrix} \begin{bmatrix} 0.0968 & -0.0770 \\ -0.0770 & 0.0811 \end{bmatrix} \begin{bmatrix} 5 \\ -4 \end{bmatrix}^{1/2} = 0$$

que é um pouco inferior do que a distância padronizada 0,91. Para verificar as diferenças compare: distância euclideana padronizada distância de Mahalanobis.

d(A,B) = 0,91 d(A,B) = 0,80

d(B,E) = 0,91 d(B,E) = 2,61.

Como interpretar esses resultados? Observe que a representação gráfica não ajuda a interpretar os resultados. A interpretação está associada à disposição espacial dos dados. A alta correlação positiva pode ser entendida como a direção e a força da correnteza de um rio e a distância de Mahalanobis como o "esforço" para ir de um ponto ao outro. Como A e B estão na direção da corrente e B e E estão no sentido contrário é mais fácil (0,80) ir de A a B do que de B a E (2,61), embora euclideanamente as distâncias sejam as mesmas (0,91). Quando o número de variáveis envolvidas é muito grande, é quase que impossível entender e interpretar esse coeficiente. Mas sabendo como ele age, deve-se usá-lo, se julgado o mais conveniente.

Alguns Outros Coeficientes

De um modo geral os coeficientes de parecença são criados com o intuito de moldar situações especiais de interesse do pesquisador, e por isso depara-se com uma série bem ampla de tais medidas. Muitas dessas medidas aparecem em publicações de áreas específicas onde é grande a utilização de técnicas de Análise de Agrupamento (taxonomia numérica, identificação de padrões, botânica, geologia, marketing, etc.). Um levantamento e análise das propriedades desses coeficientes ajuda a identificar alguns princípios gerais e encontrar alguns que melhor se ajustem aos interesses de uma particular pesquisa. Em Cormack (1971), encontra-se uma boa revisão dessas medidas. Abaixo são apresentados alguns coeficientes de uso frequente ou portadores de propriedades interessantes.

Valor Absoluto

Em vez do uso dos desvios quadráticos é muito comum o uso do valor absoluto, e tem-se

$$d(A,B) = \sum_{i=1}^{j} w\left|x_i(A) - x_i(B)\right| \qquad (2.10)$$

onde os w_i's representam as ponderações para as variáveis. Os valores mais usados são os da equiponderação $w_i = 1$ ou da média $w_i = 1/p$. Esta medida é conhecida como a "métrica do quarteirão" (métrica city-block). Usar-se-á como ilustração as variáveis padronizadas do Exemplo 2.2. Assim, com ponderação 1/p obtém-se

$$d(A,B) = \frac{\left\{\left|1.10 - 0.33\right| + \left|1.31 - 0.75\right|\right\}}{2} = 0.66$$

Distância Manhattan de Hamming ou Pombalina ou City Block

A distância de Manhattan também conhecida como distância pombalina, distância de quarteirões (city block) ou distância de taxi, desenvolvida por Herrmann Minkwski no século 19. Recebeu esse nome, pois define a

menor distância possível que um carro é capaz de percorrer numa malha urbana, rural reticulada ortogonal, tal como se encontram em zonas como Manhattan ou a Baixa Pombalina.

Ela é calculada através do somatório do valor absoluto das diferenças entre as observações e costuma suprimir grandes diferenças que venham a existir devido à presença de outliers.

$$d_{ij} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{iv} - x_{jv}|$$

A abordagem para calcular distância pode ser apropriada sob certas circunstâncias, mas provoca diversos problemas. Um é a suposição de que as variáveis não são correlacionadas uma com a outra; se elas forem correlacionadas, os agrupamentos não são válidos.

A distância entre dois elementos (i e j) é a soma dos valores absolutos das diferenças entre os valores das variáveis (v = 1, 2, ...,p) para aqueles dois casos.

$$d_{ij} = \sum_{v=1}^{p} |X_{iv} - X_{jv}|$$

Adaptado por Kaufman e Rousseeuw (1990) que comentam sobre a origem desse nome. Imagine uma floresta (de manejo florestal) ou uma cidade na qual é dividida em quarteirões de largura 1(Figura 5). Na figura 5 se quisermos nos mover entre os pontos A e B percorreremos, no mínimo, uma distância 3, uma vez que não podemos cruzar um quarteirão. Esse valor é obtido através da expressão acima.

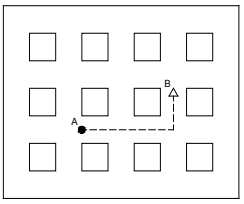


Fig. 5: Distância quarteirão entre os pontos A e B

A distância city block permite caminhar em quatro direções para ir de um ponto a outro, com isso, a distância de Manhatan nem sempre fornece a menor distância em linha reta entre dois pontos, assim como a distância euclidiana, mas é bastante utilizada por seu cálculo ser mais fácil do que o da distância euclidiana.

Exemplo: Um carro sai do ponto I (1, 1) para o ponto J(4, 3). Considerando que este carro não poderá passar por cima das àrvores ou das casas, prédios, como faremos para calcular a distância que ele percorrerá entre os dois pontos?

$$d_{ij}=\sum_{v=1}^{P}|x_{iv}-x_{jv}|$$

$$d_{ij}=|4-1|+|3-1|=3+2=5$$

Resposta: a distância percorrida pelo carro será de 5 quarteirões.

Às vezes deseja-se ressaltar a importância de certos atributos no cálculo da distância. Para isto, considera-se uma distância ponderada, que consiste em se associar pesos a cada uma das coordenadas do objeto.

$$d_{ij}=w_1|x_{i1}-x_{j1}|+w_2|x_{i2}-x_{j2}|+\ldots+w_p|x_{iv}-x_{jv}|$$

Os w_i, com i variando de 1 a p, da medida anterior, representam os pesos que podem ser dados a cada dupla de variáveis, sendo normalmente utilizada a equiponderação, com $w_i = 1$ (ou utiliza-se a média, com $w_i = \frac{1}{p}$).

Em muitos casos, a distância de Manhattan apresenta resultados similares ao da distância euclidiana. Entretanto, nessa medida, o efeito de uma grande diferença entre uma das dimensões de um elemento é minimizado, já que a mesma não é elevada ao quadrado.

(a) Distância de Minkowsky

A generalização da medida anterior que passa a ser:

$$d(A,B) = \left[\sum_{i=1}^{p} w_i \left| x_i(A) - x_i(B) \right|^k \right]^{1/k} \qquad (2.11)$$

para p = 1 e p = 2 passa a ser o caso anterior e a distância euclideana respectivamente.

para k = 3 e w_i = 1/2 tem-se, no exemplo 2.1

$$d(A,B) = \left[\frac{1}{2} \left| (1.10 - 0.33)^3 + (1.31 - 0.75)^3 \right| \right]^{1/3} = 0.60$$

(b) Coeficiente de Gower

Baseado na proporção da variação em relação à maior discrepância possível:

$$d(A,B) = -\log_{10} \left[1 - \frac{1}{p} \sum_{i=1}^{p} \frac{\left| x_i(A) - x_i(B) \right|}{x_{\max_i} - x_{\min_i}} \right] \qquad (2.12)$$

Exemplificando

$$d(A,B) = -\log_{10} \left[1 - \frac{1}{2} \left[\frac{\left| 1.10 - 0.33 \right|}{1.10 - (-1.21)} + \frac{\left| 1.31 - 0.75 \right|}{1.31 - (-1.35)} \right] \right] = 0.14$$

(c) Coeficiente de Similaridade de Cattel (HARTIGAN, 1974, p.67).

$$c(\mathrm{A},\mathrm{B}) = \frac{2\left[p - \frac{2}{3} \right] - d^2}{2\left[p - \frac{2}{3} \right] + d^2} \qquad (2.13)$$

onde d^2 é a distância euclideana com variáveis padronizadas. Assim

$$c(\mathrm{A},\mathrm{B}) = \frac{2\left(2\left(\frac{2}{3}\right)\right)0.9^2}{2\left(2\left(\frac{2}{3}\right)\right)+0.91^2} = 0.53$$

Uma outra derivada desta é devida a Cattel e Coulter (CORMACK, 1971).

$$c(\mathrm{A},\mathrm{B}) = \frac{\sqrt{2p}\,d^2}{\sqrt{2p}+d^2} = 0.41$$

(d) Coeficientes para Variáveis Positivas

Alguns coeficientes são baseados no fato dos critérios de similaridade assumirem valores estritamente positivos. Alguns foram ampliados para englobar valores negativos, mas fica mais fácil compreendê-los quando as variáveis envolvidas assumem apenas valores positivos. Eis algumas delas:

(e1) Coeficiente de Canberra

$$\mathrm{d}(\mathrm{A},\mathrm{B}) = \frac{1}{p}\sum_{i=1}^{p}\frac{|x_i(A)-x_i(B)|}{x_i(A)+x_i(B)} \qquad (2.15)$$

(e2) Coeficiente de Bray-Curtis

$$\mathrm{d}(\mathrm{A},\mathrm{B}) = \frac{\sum|x_i(A)-x_i(B)|}{\sum|x_i(A)+x_i(B)|} \qquad (2.16)$$

(e3) Coeficiente de Sokal e Sneath

$$\mathrm{d}(\mathrm{A},\mathrm{B}) = \left(\frac{1}{p}\sum\left(\frac{x_i(A)-x_i(B)}{x_i(A)+x_i(B)}\right)^2\right)^{1/2} \qquad (2.17)$$

Exemplo 2.3. Usando ainda os dados da Tabela 2.1, pode-se obter um outro tipo de relativização através da transformação

$$Z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Os resultados estão na Tabela 2.2

Tabela 2.2. Relativização das variáveis Peso e Altura

Indivíduo	Altura	Peso	Z1	Z2
A	180	79	1.00	1.00
B	175	75	0.67	0.79
C	170	70	0.33	0.53
D	167	63	0.13	0.16
E	180	71	1.00	0.58
F	165	60	0.00	0.00
x_{max} - x_{min}	15	19	1.00	1.00

(e1) Camberra

$$d(A,B) = \frac{1}{2}\left[\frac{|1.00 - 0.67|}{1.00 + 0.67} + \frac{|1.00 - 0.79|}{1.00 + 0.79}\right] = 0.16$$

(e2) Bray-Curtis

$$d(A,B) = \frac{1}{2}\left[\frac{|1.00 - 0.67| + |1.00 - 0.79|}{1.00 + 0.67 + 1.00 + 0.79}\right] = 0.16$$

(e3) Sokal-Sneath

$$d(A,B) = \left[\frac{1}{2}\left(\frac{0.33}{1.67}\right)^2 + \left(\frac{0.21}{1.79}\right)^2\right]^{1/2} = 0.16$$

O fato dos três valores acima serem iguais é simples coincidência e deve-se também ao nível de precisão adotado. Mas observando as expressões nota-se, que existe menos de casos muitos especiais, esses três coeficientes produzem resultados muito parecidos.

Coeficientes de Parecença para Atributos Qualitativos Nominais

É frequente o uso de critérios qualitativos na procura de elementos semelhantes, daí a necessidade de coeficientes que definam o grau de similaridade entre os objetos segundo variáveis desse tipo. Por facilidade de apresentação iniciar-se-á a apresentação pelo caso onde os critérios envolvidos são todos do tipo binário (sim ou não). Depois far-se-á a extensão para variáveis com múltiplos atributos.

São muitas as propostas deste tipo de coeficiente encontradas na literatura. Praticamente qualquer medida de associação para as chamadas tabelas de contingência, pode ser usada como medida de parecença.

Coeficientes de Parecença para Variáveis Dicotômicas

Será usado o exemplo apresentado no Tabela 2.1, para ilustrar as medidas apresentadas aqui. Ele resume para duas cidades (objetos) A e B, informações (variáveis) sobre a existência ou não de serviços assistenciais (Posto de Saúde, Creche, etc.). Deseja-se medir a similaridade entre as duas cidades.

Tabela 2.3

a. Resultados sobre a presença (1) ou não (0) de 10 serviços assistenciais em duas cidades (A e B)

Variável	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Cidade A	1	1	0	1	0	1	1	0	0	1
Cidade B	1	0	0	1	0	1	1	1	0	0

b. Tabela 2.1, de dupla entrada apurando o número observado de pares (1, 1), (1, 0), (0, 1) e (0, 0)

Tabela 2.4 – Número Observados

		Cidade A		Total
		1	0	
	1	a=4	b=1	a + b = 5
Cidade B	0	c=2	d=3	c + d = 5
Total		a + c = 6	b + d = 4	a + b + c + d

(a) Distância Euclideana Média

Nesta particular situação poder-se-ia calcular a distância euclideana entre os dois vetores da Tabela 2.4.(a), o que daria

$$d(A,B) = \left[\frac{1}{p}\sum_{i=1}^{p}\left(x_i(A) - x_i(B)\right)^2\right]^{1/2} = \left[\frac{b+d}{p}\right]^{1/2} = \left[\frac{b+d}{a+b+c+d}\right]^{1/2} \qquad (2.17)$$

conhecida na literatura como a distância binária de Sokal. Ela indica a proporção de atributos não coincidentes nos dois objetos. Quanto maior esse número, mais diferente serão os objetos, sendo, portanto, uma medida de dissimilaridade. Pode-se verificar que sua amplitude de variação é de 0 a 1. O valor nulo significa maior similaridade entre as cidades. No exemplo ilustrativo esse valor é

$$S(A,B) = \left[\frac{c+d}{a+b+c+d}\right] = \left[\frac{2+1}{4+1+2+3}\right]^{1/2} = 0.55$$

Uma série de outros coeficientes deriva deste conceito.

(b) Coeficiente de Concordância Simples

Uma primeira transformação é a de construir um coeficiente de similaridade. A proposta mais usada é a proporção de coincidências, isto é,

$$S(A,B) = \frac{a+d}{a+b+c+d} = 0.7 \qquad (2.18))$$

onde valores grandes significam maior similaridade entre os objetos. Este coeficiente também varia entre 0 e 1.

(b) Coeficiente de Concordâncias Positivas

Às vezes deseja-se medir a similaridade baseando-se apenas na presença da característica e não na ausência. Isso leva ao seguinte coeficiente:

$$S(A,B) = \left[\frac{a}{a+b+c+d}\right] = \left[\frac{4}{4+1+2+3}\right] = 0.40$$

ou um outro muito mais usado, o coeficiente de Jaccard

$$S(A,B) = \left[\frac{a}{a+b+c}\right] = \left[\frac{4}{4+1+2}\right] = 0.57 \qquad (2.20)$$

(Analise o significado deste coeficiente).

(d) Outras Medidas

Tentando ressaltar propriedades específicas criou-se uma série de coeficientes que são derivações dos anteriores. Alguns desses coeficientes estão na Tabela 2.5. Deixa-se ao leitor a responsabilidade de reconhecer e interpretar as qualidades que cada um pretende ressaltar. Maiores explicações podem ser encontradas em Romesburg (1984).

Tabela 2.5. Alguns Coeficientes de Semelhança para Variáveis Dicotômicas. Baseado em (ROMESBURG, 1984)

Coeficientes	Expressão da Similaridade(1)	Intervalos
Jaccard (1901)	$s_{ii'} = a/a+b+c$	0-1
Dice (1941) Sorensen (1948) /Nei e Li	$s_{ii'} = 2a/2a+b+c$	0-1
Concordância simples (Simple Matching) Sokal and Michener,(1958)	$s_{ii'} = a+d/a+b+c+d$	0-1
Russel e Rao (1940)	$s_{ii'} = a/a+b+c+d$	0-1
Yule	$s_{ii'} = ad-bc/ad+bc$	-1-1
Haman	$s_{ii'} = (a+d)-(b+c)/a+b+c+d$	-1-1
Pearson	$s_{ii'} = a+ad/a+b+c+ad$	0-1
Roger e Tanimoto (1960)	$s_{ii'} = a+d/a+2(b+c)+d$	0-1
Sokal e Sneath	$s_{ii'} = 2(a+d)/2(a+d)+b+c$	0-1
Ochiai (OCHIAI, 1957)	$s_{ii'} = a/\sqrt{(a+b)(a+c)}$	0-1
Pearson	$s_{ii'} = ad-bc/(a+b)(a+c)(b+d)(c+d)$	-1-1
Binária de Sokal*	$d_{ii'} = \sqrt{b+c/(a+b+c+d)}$	0-1

(1) a: número de coincidências do tipo 1-1 para cada par de genótipos; b: número de discordâncias do tipo 1-0 para cada par de genótipos; c: número de discordâncias do tipo 0-1 para cada par de genótipos; d: número de coincidências do tipo 0-0 para cada par de genótipos.
*Binária de Sokal (medida de dissimilaridade).

Coeficientes de Parecença para Variáveis Qualitativas

Quando a variável qualitativa nominal possui mais de dois níveis, o artifício usual é a transformação em variáveis binárias através da criação de variáveis fictícias (dummies), resolvendo a questão pelo emprego dos coeficientes definidos acima.

Suponha o vetor de critérios qualitativos nominais:

$$y' = (y_1, y_2, \ldots, y_i)$$

onde a *i*-ésima componente assume l_i níveis, codificados de modo que

$$y_i = j, \quad com \quad j = 1, 2, \ldots, l_i$$

Suponha também que $\sum l_i = p$. Cada componente irá dar origem a l_i variáveis binárias

$x_k(i)$ tal que

$$x_k(i) = \begin{cases} 1 & se \;\; y_i = k \\ 0 & em \;\; \text{c. c.} \end{cases}$$

Desse modo, o vetor **y** de dimensão **l** é transformado no vetor **x** de dimensão p, formado por componentes binárias. Esquematicamente tem-se:

$$y' = (y_1, y_2, \ldots, y_i) \Rightarrow x' = (\underbrace{0, \ldots, \overset{y_1}{1}, \ldots 0}_{l_1}; \ldots; \underbrace{0, \ldots, \overset{y_i}{1}, \ldots, 0}_{li})$$

Sem perda de generalidade o vetor **x** será indicado por *p* coordenadas binárias x_i, isto é,

$$x' = (x_1, x_2, \cdots, x_p)$$

e tem-se a situação anterior.

Exemplo 2.4

(a) Deseja-se medir a semelhança entre dois objetos, segundo 4 variáveis nominais, com 3, 4, 5 e 6 níveis cada uma. As características de dois objetos A e B são respectivamente:

$$y'(A) = (2,1,3,5)$$
$$y'(\mathrm{B}) = (3,3,3,3)$$

A transformação em variáveis fictícias leva aos vetores:

$$x'(A) = (0,1,0;\ 1,0,0,0;\ \ 0,0,1,0,0;\ \ 0,0,0,0,1,0)$$
$$y'(\mathrm{B}) = (0,0,1;\ \ 0,0,1,0;\ \ 0,0,1,0,0;\ \ 0,0,1,0,0,0)$$

Condensando em uma tabela de dupla entrada tem-se:

Tabela 2.6. Quantificação da semelhança entre as características A e B.

		Característica A		
		1	0	
	1	a=1	b=3	a + b=4
Característica B	0	c=3	d=11	c + d=14
		a + c =4	b + d =14	a + b + c d=18

Usando o coeficiente de Jaccard tem-se:

$$S(A,B) = \left[\frac{a}{a+b+c}\right] = \left[\frac{1}{1+3+3}\right] = 0.14$$

ou o de coincidência simples:

$$S(A,B) = \left[\frac{a+d}{a+b+c+d}\right] = \left[\frac{1+11}{1+3+3+11}\right] = 0.67$$

Observando este último exemplo nota-se claramente a importância da escolha do particular coeficiente de parecença. No exemplo, a coincidência

de zeros é um valor previsível a priori, o que deve orientar a escolha de uma particular família de coeficientes de parecença.

Outro fator de preocupação relaciona-se com o número de níveis de cada variável. A coincidência para uma variável de dois níveis, deve ter a mesma importância do que uma com 5 níveis? Harrison tem uma proposta para esta situação (SPÄTH, 1980), que é

$$s(A,B) = \frac{\sum_{i=1}^{l} I_n I_i I(y_i(A), y_i(B))}{\sum_{i=1}^{l} I_n I_i} = I(y_i(A), y_i(B))$$

Com $w_i = \frac{I_n I_i}{\sum I_n}$ onde a função I é indicadora de coincidência de níveis, isto é,

$$I(y_i(A), y_i(B)) = \begin{cases} 1 \; se \; y_i(A) = y_i(B) \\ 0 \; se \; y_i(A) \neq y_i(B) \end{cases} I$$

Nota-se que cada variável é ponderada pelo número de níveis que possui, atenuada pelo cálculo do seu logaritmo. Poderiam ser propostos outros sistemas de ponderação.

Exemplo 2.4

(b) Usando os dados do Exemplo 2.4.(a), pode-se calcular o coeficiente proposto por Harrison

Tabela 2.7

l_i	3	4	5	6	18
$l_n l_i$	1.099	1.386	1.609	1.792	5.886
100 w_i	18.7	23.5	27.3	30.5	100

$$d(A,B) = (0.187)(0) + (0.235)(0) + (0.273)(1) + (0.305)(0) = 0.273$$

Observe que uma coincidência em y_4 é "mais importante" que uma em y_1, cerca de 63% maior pela comparação dos pesos.

Coeficientes de Parecença para Variáveis Qualitativas Ordinais

Quando os critérios de parecença são estabelecidos por variáveis qualitativas do tipo ordinal, uma solução simples é considerá-las simplesmente como variáveis qualitativas e aplicar qualquer um dos coeficientes definidos na seção anterior. Esse procedimento deixa de considerar a importante propriedade da ordem. A seguir será apresentada uma extensão do conceito de variáveis fictícias para esse tipo de variável.

Utilização de Variáveis Fictícias

A mesma estratégia usada anteriormente em transformar cada possível realização em uma variável binária, de acordo com a ocorrência ou não daquele particular atributo, também pode ser usado aqui. Porém deve ser introduzida a questão da ordem. O exemplo abaixo ilustra melhor o procedimento.

Exemplo 2.5

Um dos critérios escolhidos para agrupar pessoas é a variável ordinal *Y*, nível de escolaridade, podendo assumir um dos seguintes valores:

1. Analfabeto;
2. Primário (completo ou não);
3. Secundário (completo ou não);
4. Universitário (completo ou não).

Pode-se criar quatro variáveis binárias, ou seja,

$y \to (x_1, x_2, x_3, x_4)$.

Uma pessoa com nível secundário ($x_3 = 1$) é considerada como portadora das características anteriores, por estar numa categoria de ordem acima dos outros. As alternativas acima teriam a seguinte representação vetorial

Analfabeto (1,0, 0, 0);

Primário (1, 1, 0, 0);

Secundário (1, 1, 1, 0);

Universitário (1, 1, 1, 1).

A variável ordinal *y* definida por

$$y = j, \quad j = 1, 2, \ldots, I$$

é transformada em L variáveis dicotômicas x_k, tal que:

$$y = k$$

então

$$x_i = 1, \text{ para } i = 1, 2, ..., k, \quad e \quad x_i = 0, \text{ para } i = k + 1, ..., l.$$

E novamente podem ser usados os coeficientes de semelhança definidos para variáveis binárias.

Exemplo 2.6

Suponha agora que as variáveis usadas no Exemplo 2.4.(a) sejam do tipo ordinal, com as mesmas observações para os objetos A e B, isto é,

y'(A) = (2, 1, 3, 5)

y'(B) = (3, 3, 3, 3)

Com a transformação proposta obtém-se:

X'(A) = (1,1,0; 1,0,0,0; 1,1,1,0,0; 1,1,1,1,1,0)

X'(B) = (1,1,1; 1,1,1,0; 1,1,1,0,0; 1,1,1,0,0,0)

Ou condensada na tabela 2.8 de dupla entrada

Tabela 2.8. Dupla entrada

		A		
B		1	0	Total
	1	9	3	12
	0	2	4	6
Total		11	7	18

Adotando o coeficiente de Jaccard usado no Exemplo 2.4, tem-se

$$S(\mathrm{A},\mathrm{B}) = \frac{a}{a+b+c} = \frac{9}{9+3+2} = \frac{9}{14} = 0.64$$

bem diferente de 0,14 quando as variáveis eram do tipo nominal.

As observações feitas para variáveis nominais continuam válidas para as variáveis ordinais.

Coeficientes de Parecença para Variáveis de Diferentes Tipos

É comum a presença de critérios dos diferentes tipos na procura de parecença entre objetos. Sem perder a generalidade, as variáveis podem ser reagrupadas de modo que apareçam primeiro as nominais, depois as ordinais e finalmente as quantitativas.

Esquematicamente tem-se:

$$y' = (y_1, y_2, \ldots, y_p)' = (y_n, y_0, y_p).$$

A seguir serão apresentadas algumas soluções para responder ao problema. Não serão discutidos em detalhes as razões de tais medidas que podem ser encontradas em Romesburg (1984), Späth (1980) e Everitt (1980).

Coeficiente Combinado de Semelhança

Constrói-se coeficientes de parecença c_0, c_n e c_q para cada um dos três vetores y_n, y_0 e y_q, e em seguida constrói-se um único ponderado, resumidamente,

$$S(A,B) = w_1 \times c_0(A,B) + w_2 \times c_n(A,B) + w_3 \times c_q(A,B)$$

onde os $w_i's$ são pesos escolhidos convenientemente. A construção deste coeficiente exige alguns cuidados especiais como:

i. Os coeficientes de semelhança de mesmo sentido (similaridade ou dissimilaridade);
ii. Mesmos (ou próximos) intervalos de variação dos coeficientes;
iii. Conjuntos de pesos adequados e interpretáveis.

Quanto a este aspecto tem sido muito usado ponderar pelo número de variáveis envolvidas.

A seguir faz-se uma aplicação segundo receita proposta por Romesburg (1984).

Exemplo 2.7

Suponha a situação do exemplo ilustrativo do Capítulo 1, com os dados apresentados na Tabela 1.1. Suponha agora que o objetivo é agrupar os seis indivíduos levando em conta todas as variáveis, seus tipos e recodificação.

TIPO	NOME	TRANSFORMAÇÃO	
Quantitativas	Altura	ZALT	Padronizada
	Peso	ZPES	Padronizada
	Idade	ZPES	Padronizada
Nominal	Cor	ZCR1=	1 Branca
			0 Outra
		ZCR2=	1 Preta
			0 Outra
	Sexo	ZSEX=	1 Homem
			0 Mulher
Ordinal	Instrução	ZIT1=	1 Prim., Sec., Univ. 0 Outra
		ZIT2=	1 Sec.,Univ. 0 Outra
		ZIT3=	1 Univ.
			0 Outra

As transformações acima encontram-se nas Tabelas 2.2.(a), (b) e (c), que foram separadas de acordo com seu tipo. Sendo esse o primeiro passo do processo, ou seja, dividir o conjunto de critérios em subgrupos convenientes ou de interesse.

O segundo passo é criar para cada subconjunto uma matriz de semelhança entre os objetos. Neste exemplo usou-se como coeficiente de semelhança, a distância euclideana média para as variáveis quantitativas e o coeficiente de Jaccard para os ordinais e nominais. As matrizes de Parecença, também aparecem na Tabela 2.2.(d), (e) e (f).

O terceiro passo é construir, a partir das matrizes componentes, uma matriz única composta. O primeiro cuidado que se deve tomar é o de ter medidas de mesma qualidade: todas de similaridade ou todas de dissimilaridade (distância) e duas de similaridade. Estas duas foram transformadas em dissimilaridade, multiplicando-as por (-1). Elas aparecem nas colunas TNOMI e TORDI na Tabela 2.2.(g). Nesta tabela estão apresentadas as etapas para obtenção da matriz composta de parecença. O segundo cuidado é o de encontrar uma mesma escala para as diversas medidas envolvidas.

Tabela 2.8. Ilustração do Exemplo 2.7

(a) Dados Transformados (Quantitativos)

Indivíduo	Z altura	Z peso	Z idade
A	1,10	1,31	1,17
B	0,33	0,75	0,72
C	-0,44	0,05	-1,10
D	-0,90	-0,93	0,04
E	1,10	0,19	-1,55
F	-1,21	-1,35	0,72

(b) Dados Transformados (Nominais)

Indivíduo	Cor			Sexo
	Zcor 1	Zcor 2	Zcor 3	
A	0	1	0	1
B	1	0	0	1
C	1	0	0	0
D	0	0	1	0
E	0	0	1	0
F	1	0	0	1

(c) Dados Transformados (Ordinais)

Indivíduo	Instrução		
	zit1	zit2	zit3
A	0	1	0
B	1	0	0
C	1	0	0
D	0	0	1
E	0	0	1
F	1	0	0

Dendrograma da Amostra Piloto com a Matriz de Semelhança Completa

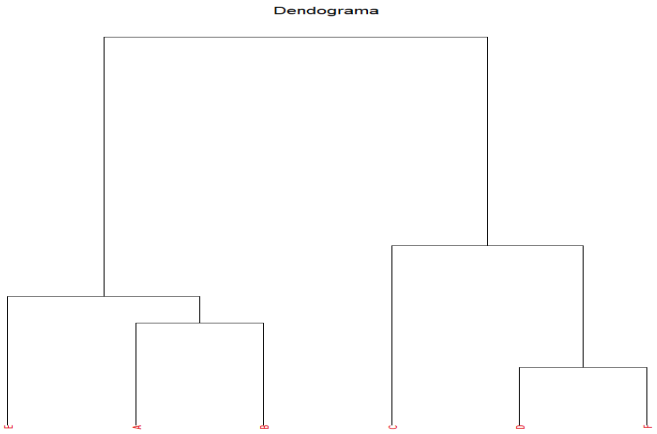


Fig. 2.1. Dendrograma da Amostra Piloto com a Matriz de Semelhança Completa

d) Matriz de Semelhança Quantitativa

$$\begin{array}{c} \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0.61 & & & & \\ 1.74 & 1.21 & & & \\ 1.85 & 1.26 & 0.91 & & \\ 1.69 & 1.42 & 0.93 & 1.61 & \\ 2.05 & 1.50 & 1.39 & 0.49 & 2.07 \end{bmatrix} \end{array}$$

e) Matriz de Semelhança Nominal

	A	*B*	*C*	*D*	*E*
B	0.33				
C	0.00	0.50			
D	0.00	0.00	0.00		
E	0.33	0,33	0.00	0.50	
F	0.00	0.00	1.00	0.00	0.00

f) Matriz de Semelhança Ordinal

	A	*B*	*C*	*D*	*E*
B	1.00				
C	0.67	0.67			
D	1.00	1.00	0.67		
E	0.67	0,67	1.00	0.67	
F	0.33	0.33	0.50	0.33	0.50

g) Pares e Coeficientes de Semelhança

par	quant	nomin	ordin	tnomi	tordi
Ab	0,608	0,330	1,000	-0,330	-1,000
AC	1,743	0,000	0,670	0,000	-0,670
Ad	1,853	0,000	1,000	0,000	1,000
Ae	1,698	0,330	0,670	-0,330	-0,670
Af	2,050	0,000	0,330	0,000	-0,330
BC	1,210	0,500	0,670	-0,500	-0,670
Bd	1,265	0,000	1,000	0,000	-1,000
Be	1,421	0,330	0,670	-0,330	-0,670
Bf	1,503	0,500	0,330	-0,500	-0,330
Cd	0,908	0,000	0,670	0,000	-0,670
Ce	0,930	0,000	1,000	0,000	-1,000
Cf	1,398	1,000	0,500	-1,000	-0,500
De	1,611	0,500	0,670	-0,500	-0,670
Df	0,494	0,000	0,330	0,000	-0,330
Ef	2,070	0,000	0,500	0,000	-0,500
Média	1,384			-0,233	-0,667
D. Padrão	0,469			0,290	0,237

(h) Pares e Coeficientes de Semelhança (Cont.)

par	zquan	znomi	zordi	c.semel	Dista
Ab	-1,655	-0,335	-1,406	-1,173	0,000
Ac	0,765	0,801	-0,011	0,648	1,821
Ad	1,000	0,801	-1,406	0,533	1,706
Ae	0,669	-0,335	-0,011	0,221	1,394
Af	1,420	0,801	1,425	1,214	2,388
BC	-0,371	-0,921	-0,011	-0,494	0,679
Bd	-0,254	0,801	-1,406	-0,094	1,079
Be	0,079	-0,335	-0,011	-0,074	1,099
Bf	0,253	-0,921	1,425	0,057	1,231
Cd	-1,015	0,801	-0,011	-0,242	0,931
Ce	-0,968	0,801	-1,406	-0,451	0,722
Cf	0,030	-2,643	0,707	-0,748	0,425
De	0,484	-0,921	-0,11	-0,067	1,106
Df	-1,898	0,801	1,425	-0,444	0,729
ef	1,462	0,801	0,707	1,116	2,289

i) Matriz de Semelhança Composta

$$\begin{array}{c} \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} A \\ 0.00 \\ 1.82 \\ 1.71 \\ 1.39 \\ 2.39 \end{array} \begin{array}{c} B \\ \\ 0.67 \\ 1.08 \\ 1,09 \\ 1.23 \end{array} \begin{array}{c} C \\ \\ \\ 0.93 \\ 0.72 \\ 0.42 \end{array} \begin{array}{c} D \\ \\ \\ \\ 1.11 \\ 0.73 \end{array} \begin{array}{c} E \\ \\ \\ \\ \\ 2.29 \end{array}$$

Tabela 2.3 Dados Pessoais de seis indivíduos da amostra piloto, ordenados segundo resultados da A.A.

Indivíduos	Altura	Peso	Idade	Instrução	Cor	Sexo
E	180	71	18	Secn	Parda	Masc
A	180	79	30	Univ	Preta	Masc
B	175	75	28	Univ	Branca	Masc
C	170	70	20	Secn	Branca	Fem
F	165	60	28	Prim	Branca	Fem
D	167	63	25	Univ	Parda	Fem

Aqui adotou-se a padronização estatística (subtração da média e divisão pelo desvio padrão). Agora tem-se 3 coeficientes de parecença para cada par de objetos, indicados nas colunas ZQUAN, ZNOMI e ZORDI. O último passo agora é definir uma medida comum e aqui foi escolhida:

$$s= (3ZQUAN + 2ZNOMI + ZORDI) / 6$$

onde cada coeficiente foi ponderado pelo número de variáveis (critérios) envolvidos. Estes resultados estão na coluna C.SEMEL. Alguns programas computacionais não aceitam valores negativos para dissimilaridade (distâncias), assim pode-se transformar os dados (-1,173 neste caso), obtendo-se os valores da coluna DISTA. Para terminar, os dados desta última coluna foram escritos em forma matricial, veja Tabela 2.1.(g).

Finalmente, apenas para recordar o objetivo deste livro, usou-se nesta matriz o algoritmo MD para produzir a árvore de semelhança da Figura 2.1. Na Tabela 2.2.(i) os dados foram ordenados de acordo com o dendrograma e é possível verificar as semelhanças entre os indivíduos.

Transformação em Variáveis Binárias

Um critério bem simples, mas que perde informações (sensibilidade), é transformar todas as variáveis em variáveis binárias. Por exemplo, a variável altura, pode ser reclassificada em baixo e alto, para os menores ou maiores do que a mediana. Algumas vezes, a variável quantitativa é transformada em variável ordinal para ganhar um pouco mais de sensibilidade. Após a transformação em variáveis binárias, usa-se um dos coeficientes para variáveis binárias definidas na seção 2.4.

Transformação dos Critérios em Variáveis, Assumindo Valores no Intervalo [0, 1]

O objetivo é transformar todas as variáveis de modo que o intervalo de variação fique entre 0 e 1. Em seguida usa-se a distância euclideana, ponderada ou não, como medida de semelhança.

As variáveis binárias já estão no intervalo [0,1]. Para as variáveis nominais e ordinais usam-se os artifícios descritos nas seções 2.3 e 2.4. Para variáveis quantitativas pode-se usar a transformação:

$$Z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Outros Coeficientes

Uma série de outras propostas ainda existe para construir coeficientes de semelhança usando variáveis de diferentes escalas de mensuração e outras poderiam ser criadas para atender situações particulares. Entretanto, duas delas merecem menção especial: uma devido à sua solução ingênua e outra bem mais elaborada.

Proposta de Romesburg

Romesburg (1984) sugere esquecer o tipo da variável e aplicar a distância euclideana simples, desde que todas as variáveis estejam codificadas através de números. Sua justificativa é empírica e afirma que esse procedimento tem a capacidade de produzir grupos semelhantes, e não muitos

diferentes, daqueles usando medidas mais sofisticadas de semelhança. A maior dificuldade reside na interpretação dos valores observados para os coeficientes de parecença. Basta observar que a recodificação das variáveis nominais leva a valores distintos dos coeficientes similares. Talvez seja uma boa medida para começar a explorar os dados.

Proposta de Gower

Gower (1971) propõe um coeficiente de parecença que é uma forma mais elaborada do coeficiente combinado definido em 2.5.1. Em primeiro lugar, ele exige que para cada variável x_j seja definido um coeficiente de parecença s_i, variando entre 0 e 1. Em seguida é criada uma variável indicadora I_i, assumindo o valor 1 quando os dois objetos podem ser comparados segundo o critério i, e zero em caso contrário. Assim a similaridade entre os objetos A e B, segundo as p variáveis, de qualquer tipo, passa a ser:

$$S(A,B) = \frac{\sum I_i(A,B)s_i(A,B)}{\sum I_i(A,B)}$$

só será indefinido quando todos $I_i(A,B) = 0$, ou seja, os dois objetos não podem ser comparados segundo nenhum critério. Esta medida permite diferenciar valores perdidos de valores inexistentes. Algumas medidas propostas são casos particulares desta.

Aqueles interessados em construir medidas de semelhança recomenda-se a leitura do artigo mencionado, o qual contém um bom material para compreensão do problema e algumas de suas soluções.

Sumário

Após a escolha das variáveis que serão usadas como critérios de semelhança, uma das questões vitais das técnicas de Análise de Agrupamento é a definição do coeficiente de similaridade. Ele deve ser escolhido com muito cuidado, pois deve ressaltar qualidades específicas explicitadas nos objetivos. Daí a grande variedade de tais medidas. Neste Capítulo foram apresentados os mais comuns e mais simples. Pretendeu-se com isto, fornecer ao leitor algumas medidas que satisfaçam, ou aproximem-se, de seus objetivos. A partir daqui ele poderá criar ou procurar novos coeficientes.

Uma palavra de advertência: antes de inventar um novo coeficiente, procure um já existente. Dificilmente o seu problema é totalmente original.

Exercícios

(1) Imagine uma situação onde as estatísticas abaixo pudessem ser usadas como coeficiente de parecença:

a. coeficiente de correlação.

b. Tau de Kendall.

c. t para testar diferenças entre médias.

d. Qui-quadrado em tabelas de dupla entrada.

(2) Pontos são medidos segundo o vetor (X,Y) com média (0,0) e matriz de covariância igual a de correlação:

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

Também são dados os pontos A=(0,0) e B=(1,1)

a. Calcule a distância de Mahalanobis entre A e B.

b. Quais os valores em (a) quando r = 0, 1 e -1? Interprete os resultados.

(3) Doze microcomputadores foram analisados segundo nove variáveis e os resultados estão na Tabela abaixo.

Tabela 2.4 - Doze Microcomputadores Analisados Segundo Nove Variáveis

MICRO	VARIÁVEL								
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
M1	A	0	64	140	1	0	80	0	2058
M2	A	0	64	140	0	0	80	0	1919
M3	A	0	32	0	0	0	38	0	470
M4	A	0	64	0	0	4	40	0	293
M5	A	1	128	320	1	10	80	1	3010
M6	A	1	64	180	1	10	80	0	2558
M7	B	0	64	195	1	0	80	1	1600
M8	C	0	64	204	1	0	80	1	1250
M9	A	0	16	0	0	0	28	0	205
M10	A	0	16	0	0	0	32	0	385
M11	A	0	64	184	1	4	80	0	1700
M12	A	0	16	0	0	4	22	0	193

a. Classifique as variáveis segundo o tipo que você julgar mais adequado.

b. Para cada grupo de variáveis, encontrar uma matriz de parecença.

c. Construir a matriz única de parecença.

d. Justifique suas escolhas.

CAPÍTULO 3

Formando os Agrupamentos

Introdução

A existência de uma definição formal de agrupamentos poderia facilitar bastante a criação de algoritmos para encontrá-los. Entretanto, essa definição envolve uma série de conceitos pessoais que nem sempre são aceitos universalmente. Porém, todas baseiam-se em duas ideias básicas: coesão interna dos objetos e isolamento externo entre os grupos (CORMACK, 1971). Como foi discutido no Capítulo anterior, cada investigador pode ter uma maneira diferente para medir as duas ideias acima, daí a existência do grande número de algoritmos para formar agrupamentos. Ainda no Capítulo 2 discutiu-se bastante o conceito de parecença entre objetos. Neste Capítulo a ideia chave é a de parecença entre grupos. As distintas definições darão origem aos diferentes algoritmos. As próprias técnicas de agrupar podem ser "classificadas" em grupos, e mesmo aqui, diferentes autores produzem diferentes classificações. Consultem, por exemplo, Cormack (1971) ou Everitt (1974). Por simplicidade, adotar-se-á aqui, a sugestão do primeiro que propõe a seguinte classificação para os algoritmos de "Análise de Agrupamentos":

i. Técnicas Hierárquicas: Na qual os objetos são classificados em grupos em diferentes etapas, de modo hierárquico, produzindo uma árvore de classificação;

ii. Técnicas de Partição: Nas quais os agrupamentos obtidos produzem uma partição do conjunto de objetos;

iii. Técnicas de Cobertura: Nas quais os agrupamentos obtidos recobrem o conjunto de objetos, mas podem-se sobrepor um ao outro.

Como qualquer classificação, existirão tipos que serão difíceis de classificar ou que poderão caber em mais de um grupo. Com esse risco em mente, esta é a classificação que será usada no livro, pois ela deve atender à maioria dos propósitos em Análise de Agrupamentos.

A seguir passar-se-á a descrever os algoritmos mais usados. Optou-se pela apresentação através de uma aplicação numérica, usando para isso o exemplo do Capítulo 1.

Exemplo 3.1

Para ilustrar os procedimentos de diversos algoritmos usar-se-á o mesmo exemplo do Capítulo 1, com apenas duas variáveis PESO e ALTURA, já padronizados (ZALT e ZPES) e a distância euclideana reduzida como medida de parecença. Assim, temos

(a) Coordenadas

Tabela 2.5

Objeto	ZALTURA	ZPESO
A	1.10	1.31
B	0.33	0.75
C	-0.44	0.05
D	-0.90	-0.93
E	1.10	0.19
F	-1.21	-1.35

b) Matriz de Distâncias

$$\begin{array}{c} \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0.67 & & & & \\ 1.41 & 0.74 & & & \\ 2.12 & 1.47 & 0.77 & & \\ 0.79 & 0.67 & 1.09 & 1.62 & \\ 2.49 & 1.84 & 1.13 & 0.37 & 1.96 \end{bmatrix} \end{array}$$

Técnicas Hierárquicas de Agrupamento

As técnicas hierárquicas podem ainda ser subdivididas em dois tipos: *aglomerativas*, onde através de fusões sucessivas dos n objetos, vão sendo obtidos $n-1$, $n-2$, etc., grupos até reunir todos os objetos num único grupo; *divisivos*, partem de um único grupo e por divisão sucessivas vão sendo obtidos 2,3 etc., grupos. O que caracteriza estes processos é que a reunião de dois agrupamentos numa certa etapa produz um dos agrupamentos da etapa superior, caracterizando o processo hierárquico. Os processos aglomerativos são mais populares do que os divisivos e neste Capítulo serão abordados entre os algoritmos hierárquicos deste tipo.

Método da Centróide (M.C)

Como foi mencionado o que caracteriza os algoritmos de produzir agrupamentos é o critério usado para definir a distância entre grupos. Este processo é o mais direto deles, pois substitui cada fusão de objetos num único ponto representado pelas coordenadas de seu centro. A distância entre grupos é definida pela distância entre os centros. Em cada etapa procura-se fundir grupos que tenham a menor distância entre si.

Exemplo 3.2

1° *Passo.* O processo inicia com dado objeto alocado a um grupo. Portanto, a distância entre os grupos é a distância entre os objetos, indicada no Exemplo 3.1, pela matriz de distâncias.

b) 2° *Passo*. A matriz das distâncias índica que os dois grupos mais parecidos são *D* e *F* e serão fundidos, dando origem ao grupo *DF*, cujas coordenadas do seu centro serão:

$$ZALTURA(DF) = (-0{,}90 - 1{,}21)/2 = -1{,}06$$

e

$$ZPESO(DF) = (-0{,}93 - 1{,}35)/2 = -1{,}14.$$

Novas Coordenadas

GRUPOS	ZALT	ZPES
A	1,10	1,31
B	0,33	0,75
C	-0,44	0,05
E	1,10	0,19
DF	-1,06	-1,14

De posse das novas coordenadas, pode-se construir a nova matriz de distância. Observe que só irão mudar distâncias envolvendo o grupo *DF*. Assim,

$$d(A,DF) = \left[\frac{[(1.10 - (-1.06))^2 + (1.31 - (-1.14))^2]}{2}\right]^{1/2} = 2.31$$

(Observe que está sendo usada a distância dividida pelo número de variáveis envolvidas). De modo análogo, obtém-se:

$$d(B,DF) = 1.65 \qquad d(C,DF) = 0.95 \qquad d(\mathrm{D},DF) = 1.79$$

E finalmente a nova matriz de distância entre os 5 grupos é:

$$\begin{array}{c} \\ B \\ C \\ E \\ DF \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.67 & & & \\ 1.41 & 0.74 & & \\ 0.79 & 0.67 & 1.09 & \\ 2.31 & 1.65 & 0.95 & 1.79 \end{bmatrix} \end{array}$$

(c) 3° *Passo*. Daqui para frente o processo é iterativo. Assim, a fusão será entre A e B, produzindo

$$D(AB)) = (1{,}10 - 0{,}33)/2 = 0{,}72$$

$$ZPESO(AB) = (1{,}31 - 0{,}75)/2 = 1{,}03.$$

Com as novas coordenadas calculam-se as novas distâncias:

$$d(\mathrm{DF}, AB) = \left[\frac{[(-1.06 - 0.72)^2 + (-1.14 - 1.03)^2]}{2}\right]^{1/2} = 1.98$$

de modo análogo

$$d(C, AB) = 1{,}07 \quad d(E, AB) = 0{,}65$$

Novas Coordenadas

GRUPOS	ZALT	ZPES
C	-0,44	0,05
E	1,10	0,19
DF	-1,06	-1,14
AB	0,72	1,03

Matriz de Distâncias

$$\begin{array}{c} \\ E \\ DF \\ AB \end{array} \begin{array}{c} \begin{array}{ccc} C & E & DF \end{array} \\ \begin{bmatrix} 1.09 & & \\ 0.95 & 1.79 & \\ 1.07 & 0.65 & 1.98 \end{bmatrix} \end{array}$$

(d) 4° *Passo*. Da matriz acima, a fusão agora será entre AB e E, dando origem ao grupo ABE. Voltando às coordenadas iniciais, obtém-se:

$$ZALTURA(ABE) = (1{,}10 + 0{,}33 + 1{,}10)/3 = 0{,}84$$

$$ZPESO(ABE) = 1{,}31 + 0{,}75 + 0{,}19)/3 = 0{,}75.$$

e procedendo de modo análogo obtém-se:

Novas Coordenadas

Grupos	Zaltura	Zpeso
C	-0.44	0.05
Df	-1.06	-1.14
ABE	0.84	0.75

Matriz de Distâncias

$$\begin{array}{c} \\ DF \\ ABE \end{array} \begin{array}{c} \begin{array}{cc} C & DF \end{array} \\ \begin{bmatrix} 0.95 & \\ 1.03 & 1.90 \end{bmatrix} \end{array}$$

(e) 5° *Passo*. Fusão de *C* com *DF*, originando *CDF*, cujo centro será:

$$ZALT = (-0{,}44 - 0{,}90 - 1{,}21)/3 = -0{,}85$$

$$ZPES = (0{,}05 - 0{,}93 - 1{,}35)/3 = -0{,}74.$$

novamente deriva-se a matriz de distâncias.

Novas Coordenadas

GRUPOS	ZALT	ZPES
ABE	0,84	0.75
CDF	-0,84	-0,75

Matriz de Distâncias

```
options(digits=2)
y=(((0.84-(-0.85))^2+(0.75-(-0.75))^2)/2)
sqrt(y)
```

ABE

CDF 1,59

(f) 6° *Passo*. Finalmente reúne-se ABE com CDF, obtendo um único agrupamento contendo todos os objetos. Resumindo, tem-se a seguinte Tabela do processo hierárquico:

```
options(digits=2)
x=(((1.10-0.33)^2+(1.31-0.75)^2)/2)
sqrt(x)
```

$$d(AB) = \sqrt{\frac{(1.10 - 0.33) + (1.31 - 0.75)}{2}} = 0.67$$

NÓ	JUNÇÃO	NÍVEL
1	D e F	0,37
2	A e B	0,67
3	AB e E	0,65
4	C e DF	0,95
5	ABE e CDF	1,59

(Observação: a diminuição do nível observado no terceiro nó, prende-se ao fato de *B* estar equidistante de *A* e *E*, e a junção deveria ser simultânea, mas nenhum algoritmo incorpora esta possibilidade. Na prática, e na presença de muitas variáveis, por sorte, isto raramente ocorre. Para desenhar a árvore vamos supor o mesmo nível).

Baseado na Tabela acima constrói-se a árvore da Figura 3.1.

Dendrograma do Método da Centróide Aplicado ao Exemplo Básico

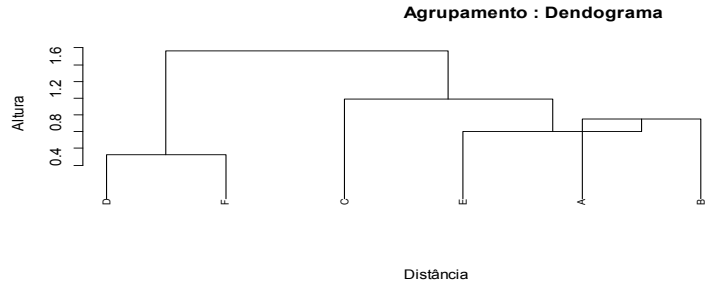


Figura 3.1. Dendrograma do Método da Centróide Aplicado

O maior empecilho ao uso desta técnica, é a necessidade a cada passo de voltar aos dados originais para recalcular as coordenadas e refazer as linhas da matriz de distâncias.

Quando muitas variáveis e objetos estão presentes, o processo torna-se impraticável quanto ao tempo de computação. Outras vezes, a matriz de parecença é construída de tal maneira que não é possível voltar às coordenadas dos objetos. Os processos hierárquicos a seguir utilizam-se apenas da matriz de distâncias entre os objetos.

Método das Médias das Distâncias (M.M.D)

Este método já foi descrito e usado no Capítulo I, portanto, não será repetido aqui. Convém ressaltar a facilidade de cálculo e rapidez em relação ao anterior. Os resultados dos dois processos são muito parecidos, como ilustração são apresentadas na Tabela abaixo, as distâncias calculadas pelos dois processos, onde pode-se constatar a veracidade da afirmação.

DIST	MC	MMD	DIST	MC	MMD
A, DF	2,31	2,30	E, AB	0,65	0,73
B, DF	1,65	1,66	DF, AB	1,98	1,98
C, DF	0,95	0,95	C, ABE	1,03	1,08
E, DF	1,79	1,79	DF, ABE	1,90	1,92
C, AB	1,07	1,08	ABE, CDF	1,59	1,64

Método da Ligação Simples ou do Vizinho mais Próximo (M.L.S)

Este método define como parecença entre dois grupos daquela dada pelos dois membros mais parecidos. Ou seja, entre todos os coeficientes de parecença entre elementos de um grupo e de outro, escolhe-se o de maior parecença como o coeficiente entre dois grupos. Assim, dados os conjuntos de objetos X e Y, a distância entre eles será definida como:

$$d(X,Y) = \min\left\{d(i,j) : i \in X \; e \; j \in Y\right\}$$

ou no caso de similaridade

$$s(X,Y) = \max\left\{s(i,j) : i \in X \; e \; j \in Y\right\}$$

Exemplo 3.3

a. **Passo 1.** Como nos processos anteriores, tem-se inicialmente 6 grupos individuais e a matriz de distância calculada no Exemplo 3.1.

b. **Passo 2.** Os dois grupos mais próximos são D e F que são reunidos no grupo DF. Necessita-se agora das distâncias deste grupo aos demais. A partir da matriz inicial tem-se:

$$d(A,DF) = \min\left\{d(A,D), \mathrm{d(A,F)}\right\} = \min\left\{2.12; 2.49\right\} = 2.12$$
$$d(B,DF) = \min\left\{d(B,D), \mathrm{d}(B,\mathrm{F})\right\} = \min\left\{1.47; 1.84\right\} = 1.47$$
$$d(C,DF) = \min\left\{d(C,D), \mathrm{d}(C,\mathrm{F})\right\} = \min\left\{0.77; 1.13\right\} = 0.77$$
$$d(E,DF) = \min\left\{d(E,D), \mathrm{d}(E,\mathrm{F})\right\} = \min\left\{1.62; 1.96\right\} = 1.62$$

que irá fornecer a seguinte matriz de distâncias:

$$\begin{array}{c} \\ B \\ C \\ E \\ DF \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & E \end{array} \\ \begin{bmatrix} 0.67 & & & \\ 1.41 & 0.74 & & \\ 0.79 & 0.67 & 1.09 & \\ 2.12 & 1.47 & 0.77 & 1.62 \end{bmatrix} \end{array}$$

c. **Passo 3.** Agrupar A e B ao nível de 0,67 e recalcular:

$$d(C,\mathrm{AB}) = \min\left\{d(\mathrm{C,A}), \mathrm{d(C,B)}\right\} = \min\left\{1.41; 0.74\right\} = 0.74$$
$$d(E,\mathrm{AB}) = \min\left\{d(E,A), \mathrm{d}(E,B)\right\} = \min\left\{0.79; 0.67\right\} = 0.67$$
$$d(DF,\mathrm{AB}) = \min\left\{d(D,A), \mathrm{d}(D,B), \mathrm{d(F,A)}, \mathrm{d}(F,B)\right\} = \min\left\{2.12; 1.47, 2.49, 1.84\right\} = 1.47$$

Observe que estes resultados podem ser obtidos da matriz do passo anterior. As duas primeiras são as mesmas e a terceira pode ser escrita do seguinte modo:

$$d(DF,\mathrm{AB}) = \min\left\{d(DF,A), \mathrm{d}(DF,B)\right\} = \min\left\{2.12; 1.47\right\} = 1.47$$

não sendo necessário recorrer à matriz inicial. A matriz resultante será

$$\begin{array}{c} \\ E \\ DF \\ AB \end{array} \begin{array}{c} C \quad E \quad DF \\ \begin{bmatrix} 1.09 & & \\ 0.77 & 1.62 & \\ 0.74 & 0.67 & 1.47 \end{bmatrix} \end{array}$$

d. Passo 4. Nesta etapa reúne-se *AB* com *E* ao nível de 0,67. Calculam-se as distâncias:

$$d(C, \text{ABE}) = \min\left\{d(\text{C}, \text{A}), \text{d}(\text{C}, \text{B}), \text{d}(\text{C}, \text{E})\right\} = \min\left\{1.41; 0.74; 1.09\right\} = 0.74$$

$$d(DF, \text{ABE}) = \min\left\{d(D, A), \text{d}(D, B), d(D, \text{E}), \text{d}(F, A), d(\text{F}, \text{B}), \text{d}(\text{F}, \text{E})\right\}$$

$$= \min\left\{2.12; 1.47; 1.62; 2.49; 1.84; 1.96\right\} = 1.47$$

ou baseando-se na matriz anterior.

$$d(C, \text{ABE}) = \min\left\{d(C, AB), \text{d}(C, \text{E})\right\} = \min\left\{0.74; 1.09\right\} = 0.74$$

$$d(DF, \text{ABE}) = \min\left\{d(DF, AB), \text{d}(DF, \text{E})\right\} = \min\left\{1.47; 1.62\right\} = 1.47$$

Daqui vem:

$$\begin{array}{c} \\ DF \\ ABE \end{array} \begin{array}{c} C \quad DF \\ \begin{bmatrix} 0.77 & \\ 0.74 & 1,47 \end{bmatrix} \end{array}$$

(e) Passo 5. Reunir C com ABE ao nível de 0,74

$$d(DF, \text{ABE}) = \min\left\{d(D, A), \text{d}(D, B), d(D, \text{E}), \text{d}(F, A), d(\text{F}, \text{B}), \text{d}(\text{F}, \text{E})\right\}$$

$$= \min\left\{2.12; 1.47; 0.77; 1.62; 2.49; 1.84; 1.96\right\} = 0.77$$

ou mais rápido:

$$d(DF, \text{ABCE}) = \min\left\{d(DF, C), \text{d}(DF, \text{ABE})\right\} = \min\left\{0.77; 1.47\right\} = 0.77$$

E tem-se a matriz

$$\begin{matrix} & DF \\ ABCE & (0.77) \end{matrix}$$

(f) Passo 6. O último passo cria um único agrupamento contendo os 6 objetos que serão similares a um nível de 0,77. Resumindo, tem-se

Resumo:

NÓ	JUNÇÃO	NÍVEL
1	D e F	0.37
2	A e B	0.67
3	AB e E	0.67
4	ABE e C	0.74
5	ABCE e DF	0.77

O dendrograma correspondente encontra-se na Figura 3.2.

Dendrograma do Método da Ligação Simples Aplicado ao Exemplo Básico

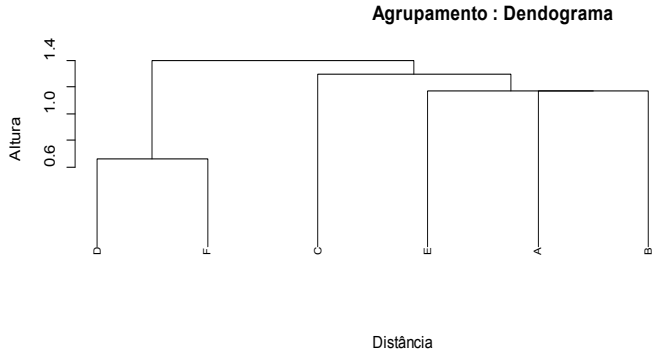


Figura 3.2. Dendrograma do Método da Ligação Simples

Uma das críticas mais sérias a este algoritmo é o de criar agrupamentos com objetos bem distintos, devido principalmente ao efeito do encadea-

mento. Observe, por exemplo, na Figura 1.1, a posição do objeto C e A que estaria num mesmo grupo por este método. O processo seguinte age na direção contrária deste procedimento

Método da Ligação Completa ou do Vizinho mais Longe (M.L.C)

Aqui a parecença entre dois grupos é definida pelos objetos de cada grupo que menos se parecem. Ou seja, formam-se todos os pares com um membro de cada grupo, a parecença entre os grupos é definida pelo par que menos se parece. No caso da parecença ser definida pela distância, a distância entre os grupos X e Y será:

$$d(X,Y) = \max\left\{d(i,j) : \text{i} \in X \text{ e } j \in Y\right\}$$

Convém ressaltar que a fusão ainda é feita com os grupos mais parecidos, menor distância.

Exemplo 3.4

(a) Passo 1. Como todo processo hierárquico inicia-se com 6 grupos individuais e as distâncias da matriz do Exemplo 3.1.

(b) Passo 2. Novamente D e F, os mais parecidos, dão origem ao grupo DF. As distâncias ao novo grupo são calculadas do seguinte modo:

$$d(A,DF) = \max\left\{d(A,D), d(A,F)\right\} = \max\left\{2.12; 2.49\right\} = 2.49$$
$$d(\text{B},DF) = \max\left\{d(\text{B},D), d(\text{B},F)\right\} = \max\left\{1.47; 1.84\right\} = 1.84$$
$$d(\text{C},DF) = \max\left\{d(\text{C},D), d(\text{C},F)\right\} = \max\left\{0.77; 1.13\right\} = 1.13$$
$$d(\text{E},DF) = \max\left\{d(\text{E},D), d(\text{E},F)\right\} = \max\left\{1.62; 1.96\right\} = 1.96$$

Daqui escreve-se a nova matriz de distâncias

$$\begin{array}{c} \\ B \\ C \\ E \\ DF \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & E \end{array} \\ \begin{bmatrix} 0.67 & & & \\ 1.41 & 0.74 & & \\ 0.79 & 0.67 & 1.09 & \\ 2.49 & 1.84 & 0.13 & 1.96 \end{bmatrix} \end{array}$$

(c) Passo 3. Os dois mais parecidos são A e B ao nível 0,67. Novas distâncias:

$$d(\mathrm{C},\mathrm{AB}) = \max\left\{d(\mathrm{C},\mathrm{A}), d(\mathrm{C},\mathrm{B})\right\} = \max\left\{1.41; 0.74\right\} = 1.41$$
$$d(\mathrm{E},\mathrm{AB}) = \max\left\{d(\mathrm{E},\mathrm{A}), d(\mathrm{E},\mathrm{B})\right\} = \max\left\{0.79; 0.67\right\} = 0.79$$
$$d(DF, AB) = \max\left\{d(D, A); d(F, A); d(\mathrm{D},\mathrm{B}); d(F, B)\right\} = \max\left\{2.12; 2.49; 1.47; 1,84\right\} = 2.49$$

Os mesmos resultados podem ser obtidos da matriz obtida no passo anterior. Por exemplo:

$$d(DF, AB) = \max\left\{d(DF, A); d(\mathrm{D}\,F,\mathrm{B})\right\} = \max\left\{2.49; 1,84\right\} = 2.49$$

Consequentemente tem-se

$$\begin{array}{c} \\ E \\ DF \\ AB \end{array} \begin{array}{c} \begin{array}{ccc} C & E & DF \end{array} \\ \begin{bmatrix} 1.09 & & \\ 1.13 & 1.96 & \\ 1.41 & 0.79 & 2.49 \end{bmatrix} \end{array}$$

(d) Passo 4. Nesta etapa reúne-se *AB* com *E* ao nível de 0,79. Calculam-se as distâncias:

$$d(\mathrm{C},\mathrm{ABE}) = \max\left\{d(\mathrm{C},\mathrm{AB}), d(\mathrm{C},\mathrm{E})\right\} = \max\left\{1.41; 1.09\right\} = 1.41$$
$$d(DF, ABE) = \max\left\{d(DF, AB); d(\mathrm{DF},\mathrm{E})\right\} = \max\left\{2.49; 1,92\right\} = 2.49$$

Daqui vem:

$$\begin{array}{c} \\ DF \\ ABE \end{array} \begin{array}{c} \begin{array}{cc} C & DF \end{array} \\ \begin{bmatrix} 1.13 & \\ 1.41 & 2.49 \end{bmatrix} \end{array}$$

(e) Passo 5. Agora o processo é repetitivo, *C* reúne-se à *DF*, cujas distâncias serão:

$$d(ABE;CDF) = \max\left\{d(ABE,C), d(ABE,DF)\right\} = \max\left\{1,41; 2.49\right\} = 2.49$$

(f) Passo 6. Todos reunidos num único grupo, ao nível 2,49.

Resumo:

Nó	Junção	Nível
1	D e F	0.37
2	A e B	0.67
3	AB e E	0.79
4	DF e C	1.13
5	ABE e CDF	2.49

O dendrograma correspondente encontra-se na Figura 3.3

Dendrograma do Método da Ligação Completa Aplicado ao Exemplo Básico

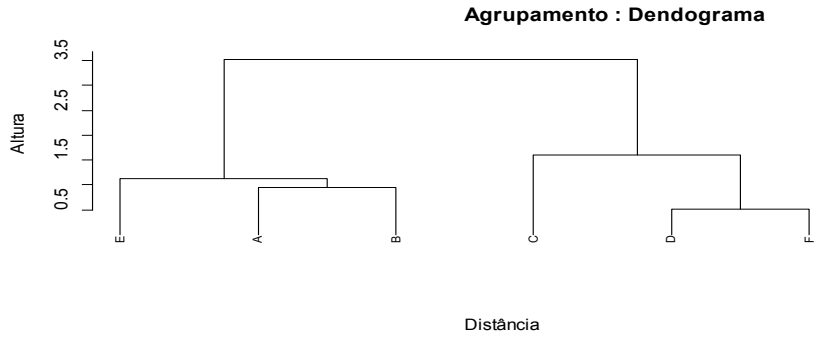


Figura 3.3. Dendrograma do Método da Ligação Completa

Ward

Este método, tem como característica a obtenção da soma dos quadrados, a qual chamaremos de SQ, para todos os possíveis grupos. A reunião definitiva dos objetos irá contemplar os menores valores de Sq. Este método pode ser usado diretamente na matriz de dados iniciais **n** x **p.**

O valor de E para dois grupos, G_1 e G_2, pode ser obtido através de

$$E_{(G_1G_2)} = \sum_{V}^{p} \sum_{i=1}^{n} \left(x_{iv} - \bar{x}_v\right)^2, \quad i \in G_1$$

Em que $\bar{x}_v$ é a média do grupo para cada variável V.

Para ilustrar o procedimento, apresentarei uma matriz de dados com cinco objetos e apenas duas variáveis:

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} V1 \\ \begin{bmatrix} 4 \\ 16 \\ 10 \\ 14 \\ 8 \end{bmatrix} \end{array} \begin{array}{c} V2 \\ \begin{bmatrix} 16 \\ 14 \\ 14 \\ 10 \\ 16 \end{bmatrix} \end{array}$$

Os resultados de todas as possíveis somas de quadrados são apresentados na Tabela 1. O primeiro passo é calcular o valor de SQ para cada um dos possíveis pares de objetos:

$$\text{em que } \bar{x}_{V1} = \frac{4+16}{2} = 10 \text{ e } \bar{x}_{V2} = \frac{16+14}{2} = 15$$

$$SQ_{(AB)} = (4-10)^2 + (16-10)^2 + (16-15)^2 + (14-15)^2 = 74$$

Esse procedimento é feito para todas as possíveis combinações de dois objetos. O menor valor de SQ indica a formação de um grupo.

Tabela 3.1 – Passos Possíveis de Agrupamentos e Valores de SQ

Passo		Possíveis	grupos		E
	(AB)	C	D	E	74
	(AC)	B	D	E	20
	(AD)	B	C	E	68
	(AE)	B	C	B	08
1	(BC)	A	D	E	18
	(BD)	A	C	E	10
	(BE)	A	C	B	34
	(CD)	A	B	E	16
	(CE)	A	B	D	04*
	(DE)	A	B	C	36
	(CE)	(AB)	D		78
	(CE)	(AD)	B		72
2	(CE)	(BD)	A		14*
	(CEA)	B	D		21.3
	(CEB)	A	D		37.5
	(CED)	A	B		37.5
	(CEA)	(BD)			31*
3	(CEBD)	A			59
	(CE)	(BDA)			105
4	(ABCDE)				115

Os valores ótimos de SQ (menores) estão indicados por*

Como pode-se observar, o primeiro grupo é formado pelos objetos C e E, pois, o valor de SQ é o menor (SQ = 4). Dessa forma podemos iniciar o passo 2, que consiste na combinação do grupo (CE) com todas as demais possibilidades de agrupamento. Para ilustrar esse passo, calcula-se:

$$SQ_{(CE)(AB)} = \frac{(10-9)^2 + (8-9)^2 + (14-15)^2 + (16-15)^2}{grupo\,(CE)} +$$

$$+ \frac{(4-10)^2 + (16-10)^2 + (16-15)^2 + (14-15)^2}{grupo\,(AB)} = 7.8$$

$$SQ_{(CEA)} = (4 - 7.3)^2 + (10 - 7.3)^2 + (8 - 7.3)^2 + (16 - 15.3)^2 + (14 - 15.3)^2 + (16 - 15.3)^2 = 21.3$$

As médias para o grupo (CEA) são obtidas somando-se os três valores correspondentes aos objetos para cada uma das duas variáveis. Nota-se, na Tabela 1, que o menor valor de SQ é para $SQ_{(CE)(AB)} = 14$; sendo assim, temos um segundo grupo formado por (BD). O terceiro passo consiste, neste exemplo, em verificar como irão se reunir os grupos (CE), (BD) e A. Por meio da Tabela 1 verificamos que o objeto A se une ao grupo (CE).

Chega-se, assim, ao final do processo, quando todos os objetos se unem.

Dendrograma do Método da Ward Aplicado ao Exemplo Básico

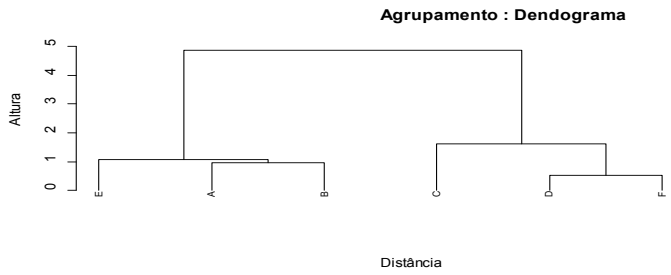


Figura 3.4. Dendrograma do Método da Ward

Este método não garante a melhor partição, pois o valor mínimo de SQ depende de resultados obtidos em passos anteriores.

Método de Divisão

Os métodos hierárquicos de divisão procedem de maneira inversa aos aglomerativos: separa-se o conjunto inicial em dois menores, até que, finalmente, todos os grupos contenham apenas um objeto. Muitos livros a respeito de Análise de Agrupamento dão pouca atenção a este método que também não tem sido considerado por grande parte dos *softwares* especializados nisso. No primeiro passo do algoritmo aglomerativo todas as reuniões de dois objetos são dadas por

C_2^n Possíveis combinações

Esse número cresce consideravelmente quando se usa o algoritmo do método de divisão, ou seja

$$\left(2^{n-1}\right)-1 \text{ possibilidades}$$

Todavia, é possível construir métodos divisivos que não consideram todas as divisões, muitos dos quais podem ser totalmente inapropriados, relatam Kaufman e Rousseeuw (1990).

Para ilustrar o procedimento deste método, retorna-se aos dados da matriz de dissimilaridade:

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & & & & \\ 12.2 & 0 & & & \\ 6.3 & 6.0 & 0 & & \\ 11.7 & 4.5 & 5.6 & 0 & \\ 4.0 & 8.2 & 2.8 & 8.5 & 0 \end{bmatrix} \end{array}$$

O processo assume, inicialmente, que todos os objetos estejam em um único grupo (ABCDE). No segundo passo, esse grupo é dividido em dois, de maneira a alocar em um deles o objeto mais dissimilar. Uma maneira para se obter esse objeto é usar a média aritmética das distâncias de cada objeto para os demais. Para o exemplo, observa-se na Tabela 3.2, logo temos

Tabela 3.2 - Média Aritmética para Demais Objetos

Objeto	Média
A	(12.2+6.3+11.7+4.0)/ 4 = 8.6
B	(12.2+6.0+4.5+2.8) / 4 = 7.7
C	(6.3+6.0+5.6+8.5) / 4 = 5.2
D	(11.7+4.5+5.6+8.5) / 4 = 7.6
E	(4.0+8.2+2.8+8.5) / 4 = 5.9

A distância de A para os demais é, em média, igual a 8.6, sendo este o objeto mais dissimilar. Definem-se, então, os grupos A e (BCDE).

Caso haja um empate entre dois ou mais objetos, sorteia-se um deles aleatoriamente. A seguir, calcula-se a média aritmética para o grupo remanescente e compara-se com a média do novo grupo, observa-se na Tabela 3.3

Tabela 3.3 - Média Aritmética para Grupo Remanescente, Novo Grupo e Diferença

Objeto	Média para o grupo remanescentes (a)	Média aritmética para os objetos do novo grupo (b)	Diferença (a - b)
B	(6.0+4.5+8,2) / 3 = 6.2	12.2	6.2 - 12.2 = -6.0
C	(6.0+5.6+2.8) / 3 = 4.8	6.3	4.8 – 6.3 = -1.5
D	(4.5+5.6+8.5) / 3 = 6.2	11.7	6.2 – 11.7 = -5.5
E	(8.2+2.8+8.5) / 3 = 6.5	4.0	4.0 – 6.5 = 2.5

Observa-se na Tabela 3.3, duas novas colunas: a coluna (b), que encerra a média dos novos grupos (neste caso pelo objeto A) e a coluna (a - b), resultante da diferença entre a média dos objetos remanescentes e a média dos objetos dos novos grupos. Seguindo o raciocínio de retirar o objeto cuja média seja maior, retira-se o objeto E, que se agrupa ao objeto A, dessa forma, tem-se (AE). O processo segue conforme a Tabela 2.9.

Tabela 3.4 - Média Aritmética para Grupo Remanescente, Novo Grupo e Diferença

Objeto	Média para o grupo remanescentes (a)	Média aritmética para os objetos do novo grupo (b)	Diferença (a - b)
B	(6.0+4.5) / 2 = 5.3	(12.2+8.2)/2 = 10.2	5.3 - 10.2 = -4.9
C	(6.0+5.6) / 2 = 5.8	(6.3+2.8)/2 = 4.6	5.8 – 4.6 = 1.2
D	(4.5+5.6) / 2 = 5.1	(11.7+8.5)/2 = 10.1	5.1 – 10.1 = -5.0

A coluna da diferença é utilizada com objetivo de interromper o processo quando todos os valores desta forem negativos. Este resultado indica que a dissimilaridade dos objetos para os novos grupos é menor; assim, os grupos formados pela extração não irão receber novos objetos, o que é mostrado na Tabela 3.5.

Tabela 3.5 - Média aritmética para grupo remanescente, novo grupo e diferença

Objeto	Média para os grupos remanescentes (a)	Média aritmética para os objetos do novo grupo (b)	Diferença (a - b)
B	4.5	(12.2+6.0+8.2)/3 = 7.0	4.5 – 7.0= - 2.5
D	4.5	(11.7+6.0+8.5)/3 = 8.7	4.5 – 8.7 = -4.2

O processo é interrompido, pois as diferenças são negativas. Logo temos os grupos (AEC) e (BD). Como o objetivo do método é que todos os grupos contenham apenas um objeto, devemos separar o grupo (AEC). Para isso se usa a matriz de dissimilaridade para os objetos citados:

$$\begin{array}{c} \\ A \\ E \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & E & C \end{array} \\ \begin{bmatrix} 0.0 & & \\ 4.0 & 0.0 & \\ 6.3 & 2.8 & 0.0 \end{bmatrix} \end{array}$$

Tabela 3.6 - Média Aritmética para Demais Objetos

Objeto	Média
A	(4.0+6.3) / 2 = 5.2
E	(4.0+2.8) / 2 = 3.4
C	(6.3+2.8) / 2 = 4.5

O objeto A é extraído devido seu valor médio ser o maior em relação aos demais. Dessa foram, obtemos os grupos (EC), A e (BD). Divide-se primeiramente o grupo (BD) por ter maior dissimilaridade em relação ao grupo (EC). A Figura mostra as sucessivas divisões.

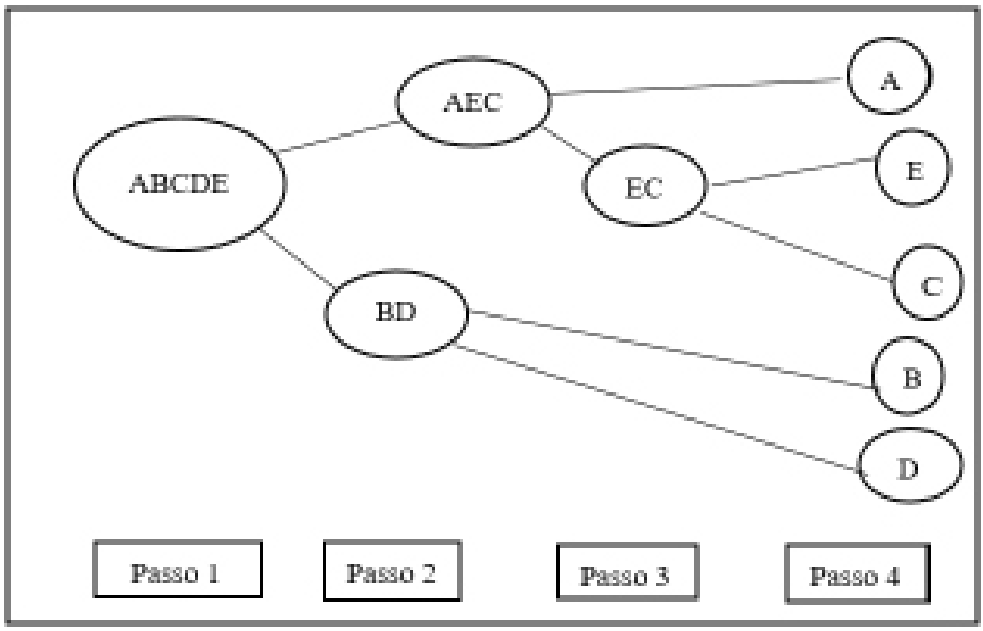


Figura 3.5 - Representação dos passos

Estas técnicas agem na direção contrária dos métodos aglomerativos. Isto é, o processo inicia com todos os elementos formando um único grupo. Este grupo é particionado em dois bens distintos. Move-se para cada um dos grupos resultantes e reproduz-se o procedimento até cada elemento ficar isolado num único agrupamento. O processo apresentado no exemplo a seguir pode ser considerado como o correspondente ao método das médias das distâncias M.M.D – Seção 3.2.2.

$$D = \begin{array}{c} \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0.67 & & & & \\ 1.41 & 0.74 & & & \\ 2.12 & 1.47 & 0.77 & & \\ 0.79 & 0.67 & 1.09 & 1.62 & \\ 2.49 & 1.84 & 1.13 & 0.37 & 1.96 \end{bmatrix} \end{array}$$

Exemplo 3.6. Voltando ao exemplo básico, parte-se da partição inicial

P(1) = {A, B, C, D, E, F}.

Procura-se o elemento mais afastado dos demais, definido como aquele com maior distância média aos demais. Por exemplo, a distância média de *A* aos demais pontos é:

$$d_m(A) = \frac{\{d(A,B) + d(A,C) + d(A,D) + d(A,\mathrm{E}) + d(A,F)\}}{5} = 1.50$$

$$d_m(A) = \frac{0.67 + 1.41 + 2.12 + 0.79 + 2.49}{5} = 1.50$$

O resultado para todos os pontos encontra-se na Tabela 3.2.(a). De lá, observa-se que F é o ponto mais "diferente" dos demais e ele será a semente do próximo grupo.

Assim, provisoriamente tem-se

$P_1(2)$ = {A, B, C, D, E}

$P_2(2)$ = {F}.

Calcula-se agora a distância média de cada ponto aos dois novos grupos. Por exemplo:

$$d_m(A, p_1(2)) = \frac{\{d(A,B) + d(A,C) + d(A,D) + d(A,\mathrm{E})\}}{4} = 1.25$$

$$d_m(A, p_1(2)) = \frac{0.67 + 1.41 + 2.12 + 0.79}{4} = 1.25$$

e

$$d_m(A, p_2(2)) = d(A,F) = 2.49$$

e a diferença

$$\square(A) = d_m(A, p_1(2)) \square d_m(A, p_2(2)) = \square 1.49$$

Valores negativos significam mais próximo de p_1 do que de p_2. Se o ponto é de p_1, tudo bem, mas sendo de p_2, ele deveria ser transferido de grupo. Os resultados dessa operação estão no Tabela 3.2.(b). Conclui-se que o ponto *D* deve ser transferido para junto de *F* e a nova partição provisória será:

$p_1(2)$ = {A, B, C, E}

$$p_2(2) = \{C, D, F\}.$$

As novas distâncias médias são calculadas e os resultados estão na Tabela 3.2.(d), revelando que essa é a partição inicial definitiva.

Move-se agora para o grupo {A, B, E}, repete o processo e irá fornecer (deixa-se ao leitor a obtenção do resultado) as partições:

$$\{A, B\} \quad \text{e} \quad \{E\}.$$

idem para {C, D, F}, depois para {A, B}, etc.

O resumo das operações está na Tabela 3.7.(e), onde a distância média entre d(X,Y) com X percorrendo a partição p_1 e Y a partição p_2, indicando o grau de distância entre as duas partições. Na parte (f) da Tabela aparece uma possível representação gráfica do procedimento. Observando o gráfico e o nível de distância em que os grupos são divididos o processo hierárquico pode ser ordenado do seguinte modo:

Tabela 3.7

Ordem	Partição	Agrupamentos
1	ABE, CDF	2
2	ABE, C, DF	3
3	AB, E, C, DF	4
4	A, B, E, C, DF	5
5	A, B, E, C, D, F	6

esta ordenação é importante para a escolha do número de grupos.

Métodos Não Hierárquicos

Nesses métodos, o número de grupos é especificado antes do processo de agrupamento que se inicia com a divisão dos objetos em *k* grupos. Uma

das formas de alocar os objetos nesses grupos é a maneira aleatória. O processo é efetuado diretamente na matriz de dados.

Descrição Geral

Estes métodos procuram diretamente uma partição dos ***n*** objetos, de modo que satisfaçam duas premissas básicas: coesão interna e isolamento dos grupos. Portanto, estas técnicas exigem a prefixação de critérios que produzam medidas sobre a qualidade da partição produzida. O uso dos métodos de partição pressupõe também o conhecimento do número ***k*** de partições desejadas. No Capítulo 4 serão discutidos alguns métodos para escolha desse número quando ele não é fixado a priori. Assim, o problema passa a ser a procura de uma partição dos ***n*** objetos em ***k*** grupos, de modo que tornem ótimo um critério de adequacidade da partição.

Isto poderia ser feito, construindo todas as partições possíveis, determinar o valor da medida para cada uma e selecionar a melhor partição. Porém, o número de possibilidades é muito grande, assintoticamente da origem de k^{n-1} veja (DURAN; ODELL, 1970). Ou seja, para resolver um problema pequeno com 16 objetos e 3 grupos, seria necessário investigar cerca de 14 milhões de partições. Desse modo, é inviável a solução através da investigação completa das partições. Portanto, os processos de partição tendem a investigar algumas partições, procurando encontrar a partição ótima ou uma alternativa que seja quase ótima.

Os algoritmos de partição diferem um do outro pela escolha diferente de um ou mais dos seguintes procedimentos:

a. Método de iniciar os agrupamentos;

b. Método de designar objetos aos agrupamentos iniciais;

c. Método de redesignar um ou mais objetos já agrupados para outros agrupamentos.

Antes de apresentar pormenores deste modo, ilustrar-se-á esta técnica através de um algoritmo muito conhecido.

K-Means

Apresentarei o método k-Means, um dos não-hierárquicos mais utilizados como técnica de agrupamento de acordo com Johnson (1982).

O processo consiste em:

1. Separar os *n* objetos em *k* grupos, de forma aleatória;
2. Calcular os centroides (médias) de cada grupo;
3. Percorrer o conjunto de objetos, associando-os ao agrupamento cujo centroide está mais próximo (utilizam-se as distâncias com ou sem a padronização das variáveis); recalcular o centroide do agrupamento que recebeu o novo objeto e do agrupamento que perdeu o objeto;
4. Repetir o Item 3 até que nenhuma reassociação tenha lugar.

Para ilustrar o procedimento, apresentarei uma matriz de dados com cinco objetos e apenas duas variáveis:

$$\begin{array}{cc} V1 & V2 \\ A\begin{bmatrix} 4 & 16 \end{bmatrix} \\ B\begin{bmatrix} 16 & 14 \end{bmatrix} \\ D = C\begin{bmatrix} 10 & 14 \end{bmatrix} \\ D\begin{bmatrix} 14 & 10 \end{bmatrix} \\ E\begin{bmatrix} 8 & 16 \end{bmatrix} \end{array}$$

Divide-se esse conjunto de dados em dois grupos, $k = 2$.

Para ilustrar o algoritmo, agruparei (AEB) e (CD). É possível notar que, na situação proposta, o melhor agrupamento seria (AE) e (BCD); o objetivo B está mais próximo do centroide de (BCD). O método *k-means* irá proporcionar o melhor agrupamento, ou seja (AE) e (BCD).

Seguindo o que foi descrito anteriormente, o primeiro passo consiste na escolha do número de grupos, assim $k = 2$.

O segundo passo é o cálculo dos centroides dos grupos (AED) e (CD):

Tabela 3.9 – Centroide para cada grupo

Grupo	Centroide V1	Centroide V2
(AEB)	(4+8+16)/3 = 9.3	(16+16+14)/3 = 15.3
(CD)	(10+14)/2 = 12	(14+10)/2 = 12

Como foi mencionado no passo 3, irei computar a distância de cada objeto em relação aos centroides. Para isso, utilizarei a distância euclidiana:

$$d(A,(AEB)) = \sqrt{(4-9.3)^2 + (16-15.3)^2} = 5.3$$
$$d((A,(CD)) = \sqrt{(4-12)^2 + (16-12)^2} = 8.9$$

Observa-se que a maior distância é a do objeto A para o grupo (AEB); dessa forma, o objeto permanece em (AEB).

Esse procedimento é realizado para todos os objetos, como é mostrado a seguir:

$$d(E,(AEB)) = \sqrt{(8-9.3)^2 + (16-15.3)^2} = 2.2$$
$$d((\mathrm{E},(CD)) = \sqrt{(8-12)^2 + (16-12)^2} = 5.6$$

Como a menor distância é a de B para o grupo (CD), este objeto é realocado neste grupo. Dessa forma, inicia-se novamente o processo, calculando-se os novos centroides para os novos grupos (AE) e (BCD).

Tabela 3.8 – Centroide Para cada Grupo

Grupo	Centroide V1	Centroide V2
(AE)	(4+8)/2 = 6	(16+16)/2 = 16
(BCD)	(16+10+14)/3= 13.3	(14+14+10)/3 = 12.6

Novamente iremos computar a distância de cada objeto para os demais centroides:

$$d(A,(AE)) = \sqrt{(4-6)^2 + (16-16)^2} = 2$$
$$d(A,(\mathrm{BCD})) = \sqrt{(4-13.3)^2 + (16-12.6)^2} = 9.9$$

O objeto A permanece no grupo (AE)

$$d(\mathrm{E},(AE)) = \sqrt{(8-6)^2 + (16-16)^2} = 2$$
$$d(\mathrm{E},(\mathrm{BCD})) = \sqrt{(8-13.3)^2 + (16-12.6)^2} = 5.6$$

O objeto E permanece no grupo (AE)

Inicia-se o mesmo procedimento para o outro grupo:

$$d(\mathrm{B},(AE)) = \sqrt{(16-6)^2 + (14-16)^2} = 10.2$$
$$d(\mathrm{B},(\mathrm{BCD})) = \sqrt{(16-13.3)^2 + (14-12.6)^2} = 3.0$$

O objeto B permanece no grupo (BCD).

$$d(\mathrm{C},(AE)) = \sqrt{(10-6)^2 + (14-16)^2} = 4.5$$
$$d(\mathrm{C},(\mathrm{BCD})) = \sqrt{(10-13.3)^2 + (14-12.6)^2} = 3.6$$

O objeto C permanece no grupo (BCD).

$$d(\mathrm{D},(AE)) = \sqrt{(14-6)^2 + (10-16)^2} = 10.0$$
$$d(\mathrm{D},(\mathrm{BCD})) = \sqrt{(14-13.3)^2 + (10-12.6)^2} = 7.3$$

O objeto D permanece no grupo (BCD).

Observa-se, pelos valores das distâncias que os grupos permanecem imutáveis.

Método das *K*-Médias

Este método, com pequenas variações, talvez seja um dos mais usados em Análise de Agrupamentos quando se têm muitos objetos.

Em primeiro lugar aparece a escolha do critério de homogeneidade dentro do grupo e heterogeneidade entre os grupos. O critério mais usado é o da soma de quadrados residual, inspirado em Análise de Variância. Suponha obtida uma partição dos $\boldsymbol{n}$ objetos em $\boldsymbol{k}$ grupos. Indicar-se-á do seguinte modo

$$p(1) = \{o_i(1) : 1 \leq i \leq n_1\}$$
$$p(2) = \{o_i(2) : 1 \leq i \leq n_2\}$$
$$p(\mathrm{j}) = \{o_i(\mathrm{j}) : 1 \leq i \leq n_j\}$$
$$p(\mathrm{k}) = \{o_i(\mathrm{k}) : 1 \leq i \leq n_k\}$$

O centro do grupo *p*(*j*), ou seja, o ponto formado pela média das coordenadas de seus membros, será representada por $\overline{o}(j)$. Desse modo, a soma de quadrados residuais dentro de *j*-ésimo grupo será

$$SQ\,\mathrm{Re}\,s(j) = \sum_{L(\times)} d^2(o_i(j); \overline{o}_1(j)) \; [1 \leq i \leq n_j]$$

onde d^2 representa o quadrado da distância euclideana do objeto *i*, do grupo *j*, ao seu centro. Para a partição toda, a soma de quadrados residual será

$$SQ\,\mathrm{Re}\,s = \sum \mathrm{SQRes}(j) \quad [1 \leq i \leq n_j]$$

Quanto menor for este valor, mais homogêneos são os elementos dentro de cada grupo e "melhor" será a partição.

Como anteriormente os demais passos serão ilustrados através do exemplo básico.

Exemplo 3.5

Os dados são os mesmos do Exemplo 3.1, e suponha que deseja-se encontrar uma partição com 2 grupos, ou seja, k = 2. A investigação de todas as partições possíveis iria exigir a construção e cálculo da SQRes para 31 combinações. Eis como o processo das *k*-médias reduz o número de partições investigadas.

i. Sementes dos Agrupamentos

Como a partição será formada por dois conjuntos, necessita-se de dois centros provisórios (duas sementes) para começar o processo. Serão escolhidos os dois primeiros objetos, na ordem de leitura, isto é, A será o centro do primeiro grupo, enquanto que B será do segundo. Observe estas operações nas duas primeiras etapas do Tabela 3.10 (a).

ii. Designação dos Objetos

Os dois primeiros objetos já foram designados, agora é a vez dos demais, e será feito de modo interativo, segundo a ordem de leitura. O próximo indivíduo na fila é *C*. Ele será colocado no grupo mais próximo, definido pela sua proximidade ao centro. Mas para evitar o cálculo da distância será usado um procedimento mais simples: usar-se-á apenas a primeira variável (ZALT) para medir a proximidade. Assim *C* de coordenada 0,44, está mais próximo do grupo 2 (0,33) do que de 1 (1,10). Assim *C* é alocado ao grupo 2 que tem as coordenadas do seu centro recalculadas.

Procede-se sequencialmente para os demais objetos e conforme aparece no final da tabela 3.10 (a), a segunda fase termina com os agrupamentos:

$$P(1) = \{\mathrm{A}, E\} \quad e \quad p(2) = \{B, C, D, E\}$$

Tabela 3.10. Etapas do Exemplo 3.5

(a)

ETAPA	LEITURA			GRUPO 1			GRUPO2		
				CENTRO			CENTRO		
	PONT	ZALT	ZPES	PONT	ZALT	ZPES	PONT	ZALT	ZPES
1	A	1,10	1,31	A	1,10	1,31	–	–	–
2	B	0,33	0,75	A	1,10	1,31	B	0,33	0,75
3	C	-0,44	0,05	A	1,10	1,31	BC	-0,06	0,40
4	D	-0,90	-0,93	A	1,10	1,31	BCD	-0,34	-0,04
5	E	1,10	0,19	AE	1,10	0,75	BCD	-0,34	-0,04
6	F	-1,21	-1,35	AE	1,10	0,75	BCDF	-0,56	-0,37

(b)

Grupo I					Grupo II				
Etapa	Objeto	Centro			Objeto	Centro			
		Z Altura	Z Peso	$d^2(., \overline{O}_1)$		Z Altura	Z Peso	$d^2(., \overline{O}_1)$	$L(\lambda)$
1	A	1.10	1.31	0.3136				5.5614	3.8219
2	E	1.10	0.19	0.3136				3.0526	(1.8149)
3				0.5929	B	0.33	0.75	2.0376	-2.3215
4				2,8616	C	-0.44	0.05	0.1896	(1.6549)
5				6,8224	D	-0.90	-0.93	0.4326	(3.9715)
6				9.2941	F	-1.21	-1.35	1.3894	(4.3435)
Centro		1.10	0.75	SQRes(1)		-0.56	-0.37	SQRes(2)	SQRes=
SQRes		0.00	0.6272	0.6272		1.3445	2.7048	4.0493	4.6765

c)

Grupo I					Grupo II				
Etapa	Objeto	Centro			Objeto	Centro			
		Z Altura	Z Peso	$d^2(., \overline{O}_1)$		Z Altura	Z Peso	$d^2(., \overline{O}_1)$	$L(\lambda)$
1	A	1.10	1.31	0,3795				5,5614	5,54446
2	B	0.33	0.75	0,2635				3,0526	2,3214
3	E	1.10	0.19	0,3795				2,0376	2,9358
4				2,1368	C	-0.44	0.05	0,1896	0,4065
5				5,8615	D	-0.90	-0.93	0,4326	4,3400
6				8,6215	F	-1.21	-1.35	1,3894	5,7230
Centro		0.84	0.75	SQRes(1)		-0.85	-0.37	SQRes(2)	SQRes=
SQRes		0.3953	0.6272	1,0125		0,3002	1,0323	1,3325	2,3550

Calcula-se agora o grau de homogeneidade interna SQRes que foi a medida adotada para avaliar a "bondade" da partição. As informações necessárias para o cálculo encontram-se no Tabela 3.1. (b). Assim

$$SQ\,\mathrm{Re}\,s(1) = d^2(A,\overline{o}(1)) + d^2(E,\overline{o}(1)) = 0.6272$$
$$SQ\,\mathrm{Re}\,s(2) = d^2(\mathrm{B},\overline{o}(2)) + d^2(\mathrm{C},\overline{o}(2) + \mathrm{d}^2(\mathrm{F},\overline{\mathrm{o}}(2)) = 4.0493$$
$$SQ\,\mathrm{Re}\,s = SQ\,\mathrm{Re}\,s(1) + SQ\,\mathrm{Re}\,s(2) = 4.6765$$

iii. Realocação dos Objetos

Como essa é uma partição arbitrária procura-se agora passar para outra melhor, isto é, uma que diminua a SQRes. Novamente o processo é feito iterativamente, ponto a ponto. Move-se o primeiro objeto para os demais grupos e verifica-se se há ganho na SQRes. Havendo, muda-se o objeto para aquele grupo que produz o maior ganho, recalculam-se as estatísticas e passa-se ao ponto seguinte. Não havendo ganho, deixa-se o objeto no grupo original e passa-se ao seguinte. Quando não houver mais mudanças ou após um certo número de iterações, o processo para.

A diminuição na soma de quadrados residual ao mover o objeto o que está no grupo 1, para um grupo qualquer J, é dado por (veja Exercício 2).

$$L(o,j;1) = \frac{n(j)d^2(o,\overline{o}(j))}{n(j)+1} - \frac{n(1)d^2(o,\overline{o}(j))}{n(1)-1}$$

onde n(.) indica o número de elementos do conjunto referido. Observe que as informações desta fórmula se referem à partição original antes da mudança do objeto.

Voltando ao exemplo básico, tem-se para o ponto A (Tabela 3.1.(b))

$$d^2(A,\overline{o}(1)) = 0.3136$$
$$d^2(A,\overline{o}(2)) = 5.5514$$

que já é uma indicação de que o objeto A está bem localizado, mudando para o grupo 2 a diferença na SQRes será

$$L(A,2:1) = \frac{4(5.5514)}{5} - \frac{2(0.3136)}{1} = 3.8219$$

ou seja, esta mudança irá aumentar a SQREs. Dada a maneira como foi construída, quanto mais negativo for *L*, maior será o ganho.

Portanto, *A* continua no grupo 1, e investiga-se agora o ponto *B*, que está no grupo 2.

$$d^2(B,\overline{o}(1)) = 0.5929$$
$$d^2(B,\overline{o}(2)) = 2.0376$$

que revela estar o ponto B mais próximo do grupo 1 do que de 2. Isto é reforçado por:

$$L(B,1:2) = \frac{2(0.5929)}{3} - \frac{4(2.0376)}{3} = -2.3215$$

Desse modo, *B* é transferido do grupo 2 para 1 e os centros dos novos grupos são recalculados. Estes dados estão na Tabela 3.1.(c). Recalculando as estatísticas obtém-se:

$$\text{SQRes} = 2.3550$$

que também obtida pela anterior (4,6765) menos a diminuição calculada (2,3215). O processo continua agora com o ponto C e da Tabela 3.1.(c) obtém-se:

$$d^2(\text{C},\overline{o}(1)) = 2.1368$$
$$d^2(\text{C},\overline{o}(2)) = 0.1896$$
$$e$$
$$L(\text{C},1:2) = \frac{3(2.1368)}{2} - \frac{3(0.1896)}{2} = 0.4065$$

E o ponto *C* fica onde está. Repetindo o procedimento com os pontos *D*, *E* e *F* termina-se a primeira iteração com:

p(1) = {A, B, E} e p(2) = {C, D, F}.

Recomeçando a segunda iteração com A, depois B, etc, não será feito nenhum outro movimento. Assim, o processo termina produzindo a partição acima. Verifica-se neste exemplo particular que esta é a melhor partição entre os 31 possíveis, tendo sido obtida com apenas duas tentativas. Nem sempre a solução encontrada por este processo coincide com o ótimo global, algumas vezes a solução é um ótimo local, por isso é que se sugerem algumas estratégias de análise que serão discutidas mais à frente e em outros Capítulos.

Modificações nos Procedimentos

Com o intuito de aperfeiçoar, tornar mais rápido e mais eficientes os algoritmos de partição, autores têm proposto diferentes alternativas, principalmente para a escolha das sementes e do procedimento de realocação de objetos. A seguir serão vistos algoritmos das modificações mais usadas.

(a) Escolha das Sementes e Alocação Inicial

A importância deste passo pode ser constatada pelo exemplo acima. Suponha que em vez de usar as duas primeiras observações como sementes, o critério adotado seja o de ordenar os objetos segundo a variável ZALT e selecionar como semente os dois pontos mais distantes. Seriam escolhidos A e F, e mantidos os outros procedimentos, a partição ideal seria obtida na primeira iteração. Muitos algoritmos escolhem os centros iniciais e a alocação inicial dos objetos de uma só vez.

i. **Escolha aleatória.** Muitas vezes as sementes, ou mesmo a primeira partição, é escolhida de forma aleatória. O Exemplo 3.5 ilustra esse procedimento. Tem a vantagem da simplicidade e a desvantagem da ineficiência. É muito usado para gerar diferentes sementes e verificar o encontro de um ótimo global e não local.

ii. **Escolha baseada em uma variável aleatória.** Essa variável é dividida em k intervalos iguais e cada agrupamento tem origem nesse intervalo. Por exemplo, cada semente é o ponto mais próximo do centro do intervalo. A variável escolhida como guia geralmente é aquela com maior variância ou ainda uma combinação linear de todas: a soma ou a primeira componente principal. Ou seja, aqueles que po-

tencialmente já produzem partições com baixa soma de quadrados residual.

iii. **Prefixado.** O usuário tem conhecimento sobre a distribuição dos pontos e pode escolher centros convenientes.

iv. **Resultado de etapas anteriores.** É comum a aplicação repetida das k-médias para valores sequenciais de k, assim os $k-1$ conglomerados servem de início para o passo seguinte, com o objeto menos ajustado servindo como a última semente.

Quando o processo escolhe pontos iniciais como centro, e sequencialmente vai designando os demais, pode-se distinguir dois casos:

(i) recalcule as estatísticas no final da interação.

(ii) recalcule as estatísticas a cada nova alocação.

(b) Critérios de Realocação

Tendo conseguido uma partição inicial deve-se buscar pontos para serem deslocados para outros grupos. Este movimento visa otimizar algum critério de homogeneidade ou heterogeneidade prefixado. Uma primeira classificação dos procedimentos pode ser:

(i) move-se um objeto de cada vez (o mais usado).

(ii) move-se grupos de objetos.

(iii) Ainda segundo a ordem de movimentação pode-se diferenciar:

(iv) tente cada caso na ordem em que ele aparece, fazendo a mudança quando recomendado, antes de verificar o seguinte;

(v) calcule o ganho para o movimento de cada objeto e no final mova aquele que leva ao maior ganho.

(c) Critério de Homogeneidade

O procedimento de partição leva a um resultado onde pode ser aplicada a equação matricial básica de Análise de Variância:

$$T = W + B$$

onde **T** é a matriz de dispersão total, **W** é a matriz de dispersão dentro de grupo e **B** a dispersão entre grupos. A matriz **W** pode ser escrita como:

$$W = \sum W_i \quad (1 \leq i \leq k)$$

onde W_i é a matriz de dispersão dentro do agrupamento i.

Para um dado conjunto de pontos T é fixo e B e W irão depender da particular partição obtida. Quanto "menor" for W, consequentemente B será "maior", maior homogeneidade interna e separação entre os grupos.

Para dimensão 1 é fácil entender o critério de maior, mas para dimensões superiores esse critério já não é tão intuitivo, recorre-se então aos conceitos usados em Análise de Variância Multivariada (MANOVA).

(i) Traço de W

Este critério foi usado no Exemplo 3.5, através da distância euclideana de cada objeto ao seu centro. Quase todos os algoritmos baseados em distância euclidiana são inspirados por esta medida de homogeneidade interna. Seu apelo é intuitivo: melhor partição é aquela que produz a menor soma de variâncias dentro dos grupos.

Observe que:

Traço (T) = traço (W) + traço (B).

Portanto a partição ótima leva também à grande separação entre os grupos.

(ii) Determinante de W

Outra medida de variabilidade global de dados multivariados é o determinante de *W* conforme (JOHNSON; WICHERN,1982), que além da variância entre grupos, consideram também a covariância entre as variáveis envolvidas. Assim, procura-se a partição que leva ao menor valor |**W**|, o determinante de **W**. Os comentários feitos para o uso da distância de Maha-

lanobis cabem também aqui na hora de interpretar os conglomerados obtidos.

(iii) Traço e Determinante de BW^{-1}

Quando tem-se apenas uma variável envolvida um bom critério para julgar a partição é o valor da estatística *F*, definida por:

$$F = \frac{\frac{B}{(K-1)}}{\frac{W}{(n-k)}}$$

ou resumidamente B/W. A generalização multivariada dessa estatística é o produto matricial

$$BW^{-1}$$

o que pode dar origem a dois outros critérios:

$$traço(BW^{-1}) = \sum \lambda_i$$

e

$$\left|BW^{-1}\right| = \frac{|B|}{|W|} = \lambda_1\lambda_2 \ldots \lambda_k$$

onde λ_i é o i-ésimo autovetor de BW^{-1}

(iv) Outros critérios

Alguns outros critérios foram propostos baseados no conhecimento da distribuição de probabilidade e usando função de verossimilhança ou ainda usando funções especiais de "estabilidade média" e centros de atrações ou ainda teoria da informação. Sugere-se aos usuários de pacotes compu-

tacionais a identificação do critério adotado pelo mesmo, verificando se o mesmo atende às suas necessidades.

Outros Métodos

É muito grande a variedade de algoritmos para produzir agrupamentos, cada um procurando ressaltar determinados critérios, restrições, objetivos, etc. No Capítulo 5, são indicados alguns programas aplicativos e descritos os algoritmos presentes. Entretanto, com o intuito de informar o leitor dos princípios que os diferenciam, abaixo são apresentadas algumas outras técnicas, procurando cobrir as principais "famílias" de algoritmos.

Tabela 3.11. Etapas do Exemplo 3.6

(a) Distância Média de Cada Ponto aos Demais

PONTO	A	B	C	D	E	F
DISTÂNCIA MÉDIA	1,50	1,08	1,03	1,27	1,23	1,56

(b) Distância Média de Cada Ponto às Partições

OBJETO	DIST. MÉTIDA À {F}	DIST. MÉDIA À {A, B, C, D, E}	DIFERENÇA
A	2,49	1,25	-1,24
B	1,84	0,89	-0,95
C	1,13	1,00	-0,13
D	0,37	1,50	1,13
E	1,96	1,04	-0,92
F	0,00	1,56	-1,56

(c)

OBJETO	DIST. MÉDIA À {D, F}	DIST. MÉDIA À {A, B, C, E}	DIFERENÇA
A B	2,30	0,96	-1,34
C	1,66	0,69	-0,97
E	0,95	1,08	0,13
	1,79	0,85	-0,94
D F	0,37	1,50	1,13
	0,37	1,86	1,49

(d)

OBJETO	DIST. MÉDIA À {C, D, F}	DIST. MÉDIA À {A, B, E}	DIFERENÇA
A	2.01	0.73	-1.28
B	1.35	0.67	-0.68
C	1.56	0.73	-0.83
D	0.95	1.08	0.13
E	0.57	1.74	1.17
F	0.75	2.10	1.35

(e) Resumo das Partições

ETAPA	PARTIÇÃO INICIAL	FINAL p_1	p_2	DIST. MÉDIA
1	A B C D E F	A B E	C D F	1,64
2	A B E	A B	E	0,75
3	C D F	D F	C	0,95
4	A B	A	B	0,67
5	D F	D	F	0,37

(f) Representação Gráfica

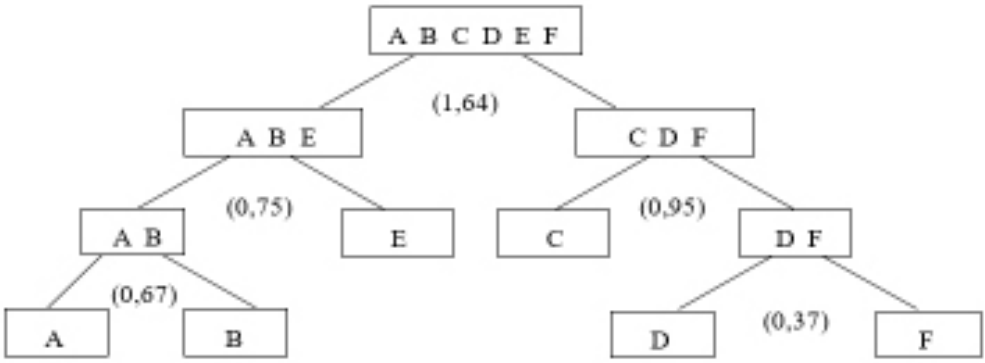


Figura 3.6

Técnicas AID

Uma outra família de algoritmos muito semelhante à anterior é conhecida pela sigla AID (Automatic Interaction Detector). São processos hierárquicos divisivos onde o critério a ser otimizado está relacionado como uma variável "resposta". É situação semelhante àquelas usadas em modelos lineares com a existência de uma variável resposta (Y) e várias preditoras (X_1, X_2, ..., X_p), estas usualmente categóricas. O objetivo é procurar quais combinações de níveis dos preditores que melhores discriminam a variável resposta. O critério mais usado para julgar a discriminação é a soma de quadrados entre os grupos formados da variável resposta. O exemplo abaixo explica melhor o procedimento.

Exemplo 3.6. Deseja-se explorar quais os fatores que explicam a diferença salarial (Y) medida em unidades monetárias (UM) convenientes. Foram escolhidas as 5 variáveis categóricas abaixo como preditoras, a saber:

CÓDIGO	NOME	DESCRIÇÃO DOS NÍVEIS
Idade	Idade	Trabalhadores de 18 a 65 anos divididos em 6 faixas etárias
Esta	Estado civil	1 - solteiro 2 - outro
Ocupação	Ocupação	1 - Cargos de diretoria 2 - Chefias 3 - Especializados 4-5-6 Operários
Educação	Educação	1 - menos de um ano de educ. formal 2 - entre 1 e 4 anos 3 - entre 4 e 8 anos 4 - entre 8 e 11 anos 5 - nível universitário
Tamanho	Tamanho da empresa	1 - pequena 2 - média 3 - grande 4 - muito grande

O resultado típico do algoritmo AID está na Figura 3.4 e o processo será melhor entendido observando a mencionada figura.

O processo começa com todos os 1427 indivíduos num único conglomerado, cujo salário médio é 8,52 UM. O segundo passo deve procurar a divisão dos dados em dois grupos que produzam duas médias bem distintas. Esta distinção é medida pela soma de quadrados entre os grupos:

$$SQEnt = \sum N_i(\bar{y}_i - \bar{y})^2$$

Para produzir partição de fácil interpretação, o algoritmo usa cada variável preditora separadamente. Ilustrando através da variável Tamanho da Empresa, por exemplo, tem-se a seguinte distribuição da variável no grupo 1:

CATEGORIA	MÉDIA	FREQUÊNCIA
1	7,64	294
2	6,48	462
3	10,64	233
4	10,12	438
TOTAL	8,52	1.427

Deve-se formar todas as partições dicotômicas possíveis usando as 4 categorias, ou seja, as sete seguintes partições:

(1, 234), (2, 134), (3,124), (4,123), (12,34), (13,24), (14,23)

e para cada uma delas calcular a medida de distância entre elas. Por exemplo, para a segunda partição (2,134), tem-se

$N_1 = 462$ $\quad \bar{y}_1 = 6{,}48$ $\quad N_2 = 965$ $\quad \bar{y}_2 = 9{,}49$ $\quad N = 1.427$ $\quad \bar{y} = 8{,}52$

consequentemente:

$$SQEnt(2.134) = 462(6.48 - 8.52)^2 + 965(9.49 - 8.52)^2 = 2830.63$$

Repete-se o processo para as demais partições e seleciona-se a partição que melhor separa os grupos, ou seja, a maior SQEnt. Entretanto não é necessário investigar todas as partições. Pode-se provar que ordenando as categorias segundo as suas médias, a melhor partição dicotômica será uma daquelas obtidas respeitando essa ordem. Assim, deve-se investigar apenas:

(2,143) (21,43) (214,3)

A primeira já foi calculada, para as duas seguintes têm-se:

(21.43)	$N_1 = 756$	$N_2 = 671$	
	$Y_1 = 6.93$	$Y_1 = 10.30$	SQEnt =4037.24
(214.3)	$N_1 = 1.194$	$N_2 = 233$	
	$Y_1 = 8.10$	$Y_1 = 10.64$	SQEnt =1257.82

Comparando os 3 resultados conclui-se que a melhor partição é (21,43), assim para esta variável a melhor partição produz SQEnt(TAMA) = 4.037,24.

Repete-se agora o processo para as demais variáveis, e seleciona-se a partição que produz a maior SQEnt, neste caso será SQEnt(OCUP) = 5.931,00, devido à partição (123,456), dando origem a dois grupos indicados na Figura 3.4, pelos números 2 e 3. Move-se agora para o grupo potencialmente mais promissor e repete-se o processo até a transgressão de alguma lei da parada. Procura interpretar a construção da árvore na Figura 3.4, para entender a utilidade deste algoritmo.

Hoje existem muitas versões para algoritmos do tipo AID, veja, por exemplo, Du Toit et alii (1986) ou Ho (1989). Mas recomenda-se a consulta da proposta original de Morgam e Sonquist (1964).

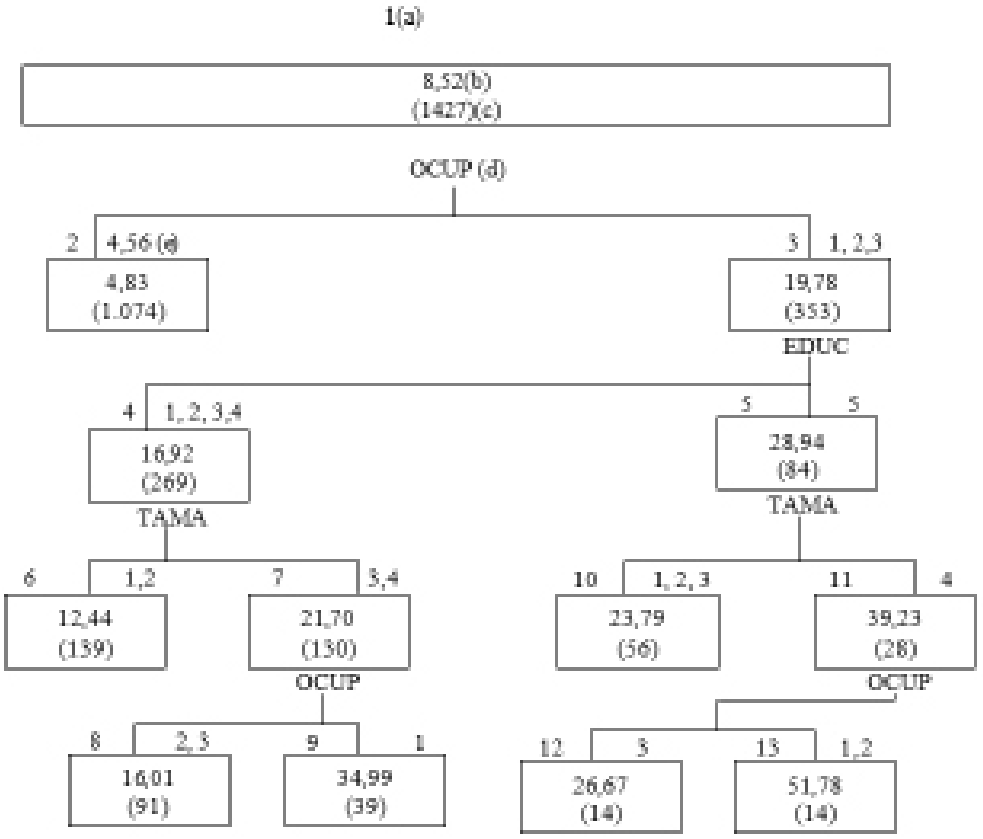


Figura 3.7. Árvore da Técnica AID

(a) número do grupo (b) média (c) número de elementos
(d) variável de partição (e) níveis da variável.

Outras Técnicas

Muitas outras propostas foram feitas para procurar agrupamentos num conjunto de dados. Algumas dessas técnicas procuram modelar distribuições de probabilidade através da procura de misturas de densidade de probabilidade ou ainda procurando os pontos centrais através da moda. Ainda dentro de algoritmos mais "teóricos", existe toda uma família de propostas baseadas nas teorias de "fuzzy sets".

Outro conjunto de algoritmos que vem ganhando bastante espaço é aquele envolvendo algum tipo de restrição. Por exemplo, procedimentos onde só podem ser agrupados objetos que respeitem uma contiguidade geográfica.

A descrição dessas várias opções seria muito extensa e cansativa. O leitor interessado encontrará na bibliografia mencionada material suficiente para iniciar sua procura.

Sumário

A opção por um particular algoritmo é outra das etapas cruciais de aplicação de A.A. Deve-se escolher aquele que melhor responda aos seus objetos e seja eficiente frente aos lados. Algoritmos que produzem árvores são difíceis de analisar na presença de muitos objetos, enquanto que processos de partição não são reveladores das parecenças entre objetos. O conhecimento dos princípios dos algoritmos mais comuns é a melhor ajuda para guiar a seleção do procedimento mais adequado. Outras vezes, a escolha fica restrita à disponibilidade dos pacotes para aproximar da solução imaginada como ideal. Por exemplo, construir distância de Mahalanobis num módulo matricial e M.M.D no módulo de A.A. Por isto, nesta seção procurou-se mostrar as linhas básicas dos principais algoritmos.

Exercícios

1. Usando a tabela de parecença do Exemplo 2.7.(h), explore os diversos algoritmos hierárquicos discutidos neste capítulo. Interprete os resultados obtidos.
2. Prove que ao mover uma observação **o** do grupo 1 para o grupo J, a soma de quadrados diminui em:

$$L(o,\mathrm{J}:1)=\frac{n(\mathrm{J})d^2(o,\overline{o}(\mathrm{J}))}{n(\mathrm{J})+1}\square\frac{n(1)d^2(o,\overline{o}(1))}{n(1)1}$$

3. Usando os dados do Exercício 2.10.3

Explore os diversos métodos hierárquicos com a matriz única de parecença.

a. Usando o M.M.D, faça os dendrogramas para cada tipo de variável.
b. Através do método k-means, procure a partição em dois grupos para as variáveis quantitativas.

CAPÍTULO 4

Tópicos Especiais

No Capítulo anterior, várias técnicas de Análise de Agrupamento foram descritas. Neste Capítulo serão tratados alguns tópicos relacionados com a utilização dessas técnicas. Devido às peculiaridades de cada método, cada um deles está exposto a problemas específicos. Existem, porém, questões de âmbito mais geral e a estas será dedicada maior atenção.

Seleção de Variáveis

O resultado de uma Análise de Agrupamento deve ser um conjunto de grupos que podem ser consistentemente descritos através de suas características, atributos e outras propriedades. Conjuntamente, esses descritores são as variáveis do problema. Assim, um dos fatores que mais influencia o resultado de uma Análise de Agrupamento é, indiscutivelmente, a escolha de variáveis.

Embora esta questão seja inerente ao campo de aplicação do problema em estudo e deva assim refletir o julgamento do pesquisador sobre a relevância das variáveis para o tipo de classificação procurada, algumas considerações são cabíveis.

Variáveis que assumem praticamente o mesmo valor para todos os objetos são pouco discriminatórias e sua inclusão pouco contribuiria para a de-

terminação da estrutura do agrupamento. Para ilustrar considere os dados do Capítulo 1. Se o grau de instrução fosse o mesmo para os seis indivíduos essa informação de nada auxiliaria na tarefa de busca de grupos naturais. Por outro lado, a inclusão de variáveis com grande poder de discriminação, porém irrelevantes ao problema, pode mascarar os grupos e levar a resultados equivocados.

Além disso, é desejável que os objetos sejam comparáveis segundo o significado de cada uma delas. Por exemplo, o tamanho da frota de ônibus das cidades brasileiras informa mais sobre as populações dessas cidades do que suas condições de transporte público. Uma variável mais adequada seria o número de ônibus por habitante.

Frequentemente, o número de variáveis medidas é grande, dificultando a análise. Deve-se, então, respeitando o princípio da parcimônia, procurar diminuir o seu número de forma que sua seleção contemple tanto a sua relevância como seu poder de discriminação face ao problema em estudo. Em último caso, pode-se ainda utilizar técnicas estatísticas para redução da dimensionalidade da matriz de dados, tais como a Análise de Componentes Principais e a Análise Fatorial.

Escala de Variáveis

Um aspecto importante a ser considerado é a homogeneidade entre variáveis. Ao se agrupar observações é necessário combinar todas as variáveis em um único índice de similaridade, de forma que a contribuição de cada variável depende tanto de sua escala de mensuração como daquelas das demais variáveis. Há casos em que a variação de uma unidade em uma variável expressa em toneladas é menos significante que a variação de uma unidade medida em kg em outra variável. Visando reduzir o efeito de escalas diferentes surgiram várias propostas de relativização das variáveis. Serão apresentadas algumas das mais comuns. Um tratamento mais completo pode ser encontrado em Späth (1980).

Considere as observações originais $x_1, ..., x_n$.

- A transformação mais comum é aquela definida por

$$z_i = \frac{x_i - \bar{x}}{s}, i = 1, \ldots, n$$

onde e s denotam a média e o desvio-padrão das observações. Esta transformação, já utilizada no Capítulo 1, faz com que os dados transformados tenham média zero e variância unitária. A Tabela 1.2 contém os valores brutos e transformados das variáveis Peso e Altura do Exemplo 1. A desvantagem desta padronização é reduzir todas as variáveis ao mesmo grau de agrupabilidade.

- Outra forma de se transformar variáveis é tornar-se os desvios em relação ao menor valor e normalizá-los pela amplitude, ou seja,

$$z_i = \frac{x_i - x_{(1)}}{x_{(n)} - x_{(1)}}, i = 1, \ldots, n$$

onde $x_{(1)}$ e $x_{(n)}$ denotam o mínimo e o máximo da amostra, respectivamente. A Tabela 2.2, mostra os dados de Peso e Altura do Exemplo 1 devidamente transformados.

- Tomando-se a média como fator normalizador pode-se definir

$$z_i = \frac{x_i}{\bar{x}}, i = 1, \ldots, n$$

A Tabela 4.1 contém as observações de Peso e Altura do Capítulo 1 devidamente transformados.

A despeito da variedade de propostas, recomenda-se que a escala das variáveis seja definida através de transformações sugeridas pelo bom senso e pela área de conhecimento da aplicação. Hartigan (1975) fornece um bom exemplo ao tentar agrupar um conjunto de alimentos (carne, peixes e aves) segundo alguns de seus nutrientes (Energia, Lipídeo, Cálcio e Ferro).

Os dados foram obtidos tornando-se as unidades em três onças de peso e estão nas Tabelas 5.1 e 5.2. Como se vê, o desvio padrão de Energia é cerca de setenta vezes o desvio padrão de Ferro, indicando que a variação de uma unidade de energia tem um significado menor que a mesma variação de Ferro. Para garantir que esses "significados" sejam mantidos faz-se necessário transformar as variáveis de forma a tornar suas variâncias mais homogêneas. Como já foi visto, a padronização usual é indesejável por reduzir

todas as variáveis a um mesmo grau de discriminação. Para manter o significado original das variáveis, Hartigan propôs que as observações fossem transformadas em percentagem das necessidades diárias de cada um dos nutrientes, segundo o Yearbook of Agriculture (1959). Conforme pode-se observar nas Tabelas 5.3 e 5.4, os dados transformados são mais homogêneos sem que a transformação tenha alterado a importância individual de cada variável.

Ponderação das Variáveis

Um outro aspecto a ser considerado é a ponderação das variáveis, já que elas não têm a mesma importância para o problema. No Capítulo 2, apresentou-se algumas métricas que permitem ordenar as variáveis segundo seu grau de relevância, por exemplo, a métrica do valor absoluto, a distância de Minkowsky, etc. Convém observar que os pesos são atribuídos de forma subjetiva e a sua escolha também depende do contexto e do grau de conhecimento que o pesquisador tem do problema. Como a A.A, é uma técnica exploratória que busca a partir dos dados, a formulação de hipóteses, o mais comum é se atribuir o mesmo peso para todas as variáveis.

Número de Grupos

Os algoritmos vistos no Capítulo 3 produzem grupos que constituem uma proposição sobre a organização básica e desconhecida dos dados. Entretanto, todos estes procedimentos esbarram em uma dificuldade comum que é a determinação do número ideal de grupos. Como ilustração, considere o Exemplo 1, onde a aplicação de M.M.D levou a dois grupos, posteriormente identificados como dos indivíduos grandes e o dos indivíduos pequenos.

O processo aglomerativo poderia também ter sido interrompido no estágio imediatamente anterior, produzindo os grupos {A, B, E}, {D, F} e {E}, o qual seria o de pessoas medianas, digamos. Para auxiliar na decisão de qual das duas propostas melhor mimetiza a população de interesse, podem ser adotadas algumas estratégias bastante simples. A seguir serão apresentadas algumas delas.

A determinação do número de grupos para uma base de dados é uma das tarefas mais difíceis no processamento de agrupamento.

O número de grupos pode ser definido, a priori, por meio de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador, por simplicidade, ou ainda pode ser definido, posteriormente, com base nos resultados da análise ou pela experiência do pesquisador.

Para determinar o número apropriado de grupos, existem diversas abordagens possíveis: em primeiro lugar, o pesquisador pode especificar antecipadamente o número de agrupamentos. Talvez, por motivos teóricos e lógicos, esse número seja conhecido. O pesquisador pode, também, ter razões práticas para estabelecer o número de agrupamentos com base no uso que pretende fazer dele. Em segundo lugar, o pesquisador pode especificar o nível de agrupamento de acordo com um critério.

Se o critério de agrupamento for de fácil interpretação, tal como a média de (dis)similaridade interna do agrupamento, é possível estabelecer certo nível que ditaria o número de agrupamentos. As distâncias entre os agrupamentos, em etapas sucessivas, podem servir de guia e o pesquisador pode escolher interromper o processo quando as distâncias excederem um valor estabelecido.

Uma terceira abordagem é representar, graficamente, a razão entre a variância total interna dos grupos e a variância entre os grupos em relação ao número de agrupamentos. O ponto em que surgir uma curva acentuada, um ponto de inflexão, poderá ser a indicação do número adequado de agrupamentos. Aumentar esse número, além desse ponto, seria inútil; e diminuí-lo, seria correr o risco de misturar parcelas diferentes.

Qualquer que seja a abordagem empregada, geralmente é aconselhável observar o padrão total de agrupamentos. Isso pode proporcionar uma medida da qualidade do processo de agrupamento e do número de agrupamentos que emergem nos diversos níveis do critério de agrupamento. De maneira geral, mais de um nível de agrupamento é relevante.

Técnicas de Partição

Um método simples e eficiente na escolha do número de grupos é a análise gráfica.

Geralmente os agrupamentos são obtidos de forma a minimizar (maximizar) uma função, por exemplo, o traço de **W**. Assim, a ideia é verificar o ganho dessa função ao se passar de k para k + 1 grupos. Pode-se, portanto, obter as configurações para 1, ..., g e os correspondentes valores da função

objetivo. O gráfico do número de grupos contra a função objetivo permitirá uma visualização da perda (ganho) da função conforme se aumenta o número de grupos. Decide-se por k ($k \leq g$) grupos quando a variação da função ao se passar de k para k + 1 grupos for pequena em relação às demais.

Técnicas Hierárquicas

As técnicas hierárquicas descritas no Capítulo 3 não apresentam um indicador intuitivo para o número de grupos, como é o caso dos métodos de partição. Uma sugestão é o exame do dendrograma em busca de grandes alterações dos níveis de similaridade para as sucessivas fusões. Por exemplo, a Figura 1.2 mostra que há um salto no nível de similaridade no passo 5, indicando que esta última fusão reuniu dois grupos pouco similares. Neste caso, portanto, seria conveniente interromper o processo aglomerativo no passo 4, confirmando-se assim a partição em apenas dois grupos.

A próxima regra de decisão a ser apresentada é baseada em uma única variável. Para facilitar a apresentação, denote a configuração de n objetos em k grupos por P(k), e a matriz de dispersão dentro dos grupos da k-ésima configuração por W(k). A dispersão da j-ésima variável é dada pelo j-ésimo elemento da diagonal da matriz de dispersão e será denotado por $w_{jj}(k)$.

Se os dados forem normalmente distribuídos e com a mesma matriz de covariância, $w_{jj}(k)$ tem distribuição proporcional a com $n - i$ graus de liberdade. Analogamente, para P(k + 1), $w_{jj}(k + 1)$ tem distribuição proporcional a com $n - k - 1$ graus de liberdade. Ainda, se P(k + 1) for o resultado da partição de um dos grupos de P(k) em dois, a estatística:

$$R_j = \left(\frac{W_j j(k)}{W_j j(k+1)} - 1 \right) \div n - k - 1$$

representa a economia relativa na variância dentro dos grupos ao se passar de k para $k + 1$ grupos, quando se considera tão somente a j-ésima variável. Dessa forma, valores altos de R indicam a existência de $k + 1$ grupos na variável j. Para um teste global deve-se repetir o procedimento acima em cada uma das variáveis e aceita-se P(k + 1) se todas as razões forem suficientemente grandes.

Se além das suposições mencionadas anteriormente for possível supor que as variáveis são independentes, pode-se obter um teste englobando-as

simultaneamente. Neste caso, o traço da matriz de dispersão dentro dos grupos tem distribuição proporcional a com *p* graus de liberdade. Assim,

$$R = \left(\frac{trW(k)}{trW(k+1)} - 1 \right) \div n - k - 1$$

tem distribuição F com p e n – k – 1 graus de liberdade.

O valor da estatística R aplicada ao resultado de uma A.A será, em geral, maior do que aquele observado em dados cuja estrutura é conhecida, uma vez que a A.A produz grupos maximizando-se a disparidade entre eles. Apesar da estatística R não fornecer um teste estatístico no caso de A.A, ainda pode ser utilizado como um indicador do número de grupos. Para isso, procede-se como se fosse fazer o teste e calcula-se o valor de R. Se o valor observado da estatística for menor que o valor crítico dado por uma tabela de distribuição F, tem-se a indicação clara da existência de k grupos. Para que R indique uma estrutura de k + 1 grupos é necessário que o seu valor observado seja não somente maior que o valor crítico da distribuição F, mas sim algumas vezes maior. Hartigan (1975) sugere, para amostras grandes, R > 10.

Exemplo

Considere novamente o exemplo do Capítulo 1, e sejam as configurações P(2) formada por {A, B, E} e {C, D, F}; e P(3) definida por {A, B, E}, {C} e {D, F}. Neste caso, tem-se:

$$w(2) = \begin{bmatrix} 0.6972 & 0.5567 \\ 0.5567 & 1.6460 \end{bmatrix} \div \begin{bmatrix} 0.0481 & 0.0651 \\ 0.0651 & 0.0882 \end{bmatrix}$$

e finalmente, R_1 =40,48 e R_2 = 52,99.

O valor crítico a 1% da distribuição *F* com 1 e 3 graus de liberdade é 34,12. A sugestão de Hartigan fica prejudicada devido ao pequeno número de objetos na amostra.

Outros Métodos

Alguns outros métodos foram propostos. Por exemplo, escolher o valor de k que minimiza $g^2 \det(\mathbf{W})$, onde g é o número de grupos ou ainda o valor de k que maximiza:

$$C = \frac{(n - k)tr(B)}{(k - 1)tr(w)}$$

onde B é a matriz de dispersão entre grupos.

Quando a técnica de mistura de multinormais é usada, o número de grupos pode ser determinado através do teste da razão de verossimilhança. Para se testar a hipótese eles existem *k* grupos versus a alternativa de que são *l* grupos, sugere-se a estatística

$$\frac{-2}{n}\left[n - 1 - p - \frac{1}{2}\right]\log$$

que converge para uma distribuição de com 2p(k – l) graus de liberdade. Wolfe (1971) mostrou que a estatística acima converge mais rapidamente para a distribuição limite do que $-2\log$.

Medidas de Semelhança entre Variáveis

A discussão de A.A nos Capítulos anteriores se centrou em agrupar as observações conforme elas são descritas pelas variáveis. Entretanto, é igualmente razoável agrupar variáveis segundo seu comportamento mútuo manifestado nas unidades amostrais. A fim de se agrupar variáveis é necessário definir coeficientes numéricos de similaridade que caracterizem as relações entre variáveis.

Nesta seção algumas medidas de similaridade entre variáveis serão apresentadas. Para facilitar a exposição serão tratadas inicialmente as variáveis quantitativas e depois as qualitativas. Mais detalhes podem ser encontrados em Romesburg (1984).

Variáveis Quantitativas

Considere o comportamento das variáveis X e Y, representadas pelos vetores $x = (x_1, ..., x_n)'$ e $y = (y_1, ..., y_n)'$.O i-ésimo componente de cada vetor representa o valor observado da variável no i-ésimo objeto, i = 1,...,n.

- Ângulo entre dois vetores – Uma medida de similaridade entre X e Y é dada pelo coseno do ângulo α entre os vetores x e y:

$$s(x,y) = \cos\square = frac\square_{i=1}^{n} x_i y_i \sqrt{\square_{i=1}^{n} x_i \square_{i=1}^{n} y_i}$$

À medida que o vetor x se "aproxima" de y, maior vai se tornando o co-seno de α.

Exemplo: Os dados brutos de Peso e Altura do Capítulo 1 fornecem s(X,Y) = 0,998.

- Coeficiente de correlação de Pearson – Uma variação do co-seno do ângulo formado por x e y é o co-seno do ângulo formado por **u** e **v**, onde para i = 1,...,n,

$$u_i = x_i \square \overline{x}$$
$$v_i = y_i \square \overline{y}$$

Assim,

$$r(X,Y) = \frac{\square_{i=1}^{n} u_i v_i}{\sqrt{\square_{i=1}^{n} u_i^2 \square_{i=1}^{n} v_i^2}}$$

$$r(X,Y) = \frac{\square_{i=1}^{n} (x_i \square \overline{x})(y_i \square \overline{y})}{\sqrt{(x_i \square \overline{x})^2 (y_i \square \overline{y})^2}}$$

A medida acima é o *coeficiente de correlação* observado entre X e Y.

Exemplo: O coeficiente de correlação entre Peso e Altura, com base nos dados do exemplo anterior, é r(X,Y) = 0,869.

Variáveis Binárias

Convencione-se os valores 1 e 0 para representar as duas categorias das variáveis dicotômicas X e Y. A classificação dos n objetos segundo X e Y pode ser representada da seguinte forma:

Tabela 4.1 - Dupla entrada para variáveis binárias

Y \ X	1	0	Total
1	a	b	a + b
0	c	d	c + d
Total	a + c	b + d	a + b + c + d

Nesta situação e tomando-se os rótulos 1-0 como os valores assumidos por X e Y, tem-se que: $\sum_{i=1}^{n} x_i y_i = a, \sum_{i=1}^{n} y_i^2 = a + c \quad e \sum_{i=1}^{n} x_i^2 = a + b$. O co-seno do ângulo entre os vetores **x** e y é dado por:

$$\cos\square = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Verificando ainda que $\sum_{i=1}^{n} x_i = a + b$ e que $\sum_{i=1}^{n} y_i = a + c$, tem-se que o coeficiente de

correlação é dado por $r = \dfrac{ad - bc}{\sqrt{(a+b)(c+d(a+c)(b+d)}}$

Exemplo:

Considere as variáveis: Sexo e Instrução para os dados da Tabela 1.1. Para estes dados, obteve-se cosα = 0,58 e r = 0.

Variáveis Multinominais

Considere o par de variáveis (X,Y) que assumem respectivamente as modalidades 1, ..., p e 1, ..., q. Suponha ainda uma tabela de contingência onde n_{ij} representa a frequência absoluta do par (i,j), e, n_i e n_j as frequências marginais de X e Y, respectivamente.

Analogamente, defina as frequências relativas, $f_{ij,}\ f_i\ e\ f_j$

Qui-quadrado – A medida de associação mais comum entre X e Y é o de Pearson dado por:

$$\chi^2 = n\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

Esta medida tem uma característica indesejável que é o fato dela ser um múltiplo do número de observações. A fim de eliminar esse efeito surgiram algumas variações.

- Coeficiente de contingência quadrática média

$$\chi^2 = \frac{\chi^2}{n}$$

- Coeficiente de contingência de Pearson

$$P = \left[\frac{\chi^2}{1+\chi^2}\right]^{1/2} \quad 7$$

- Coeficiente de Tschuprow

$$T = \left(\frac{\chi^2}{n\sqrt{(p-1)(q-1)}} \right)^{1/2}$$

- Coeficiente de Cramér

$$C = \frac{\chi^2}{n \min(p-1, q-1)}$$

Exemplo:

Para os dados de Instrução e Cor do Capítulo 1, tem-se que

$$\chi^2 = 7.56, \quad \chi^2 = 2.26 \quad P = 075, \; T = 0.22 \; e \; C = 1.26$$

É fácil verificar que no caso de variáveis binárias $\chi^2 = T^2 = C^2 = r^2$. Há outras propostas de coeficientes de associação entre variáveis na literatura. Alguns exemplos são: a correlação canônica e o coeficiente de Goodman e Kruskal.

Medidas de Associação entre Variáveis Ordinais

Serão apresentadas três medidas de associação entre variáveis ordinais, todas baseadas no coeficiente de correlação.

Tabela 4.2 - Variáveis binárias

S / I	M	F	Total
Universitário	2	2	4
Secundário	1	1	2
Total	3	3	6

Coeficiente de Kendall – Denote por $X_i < XX_j$ se a observação X_j precede X_j na relação de ordem implícita em X. Para i,j = 1,...,n, seja:

$$U_{ij} = \begin{cases} 1 se\ x_i > x_j \\ -1 se\ x_i < x_j \end{cases}$$

Analogamente, define-se:

$$V_{ij} = \begin{cases} 1 se\ y_i > y_j \\ -1 se\ y_i < y_j \end{cases}$$

Note que a definição acima descarta a possibilidade de $x_i = x_j (y_i = y_j)$ para i ≠ j. O *coeficiente* τ de *Kendall* é definido por:

$$\tau = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} U_{ij} V_{ij}}{\sqrt{\left[\sum_{i=1}^{n} \left[\sum_{j=1}^{n} Uij^2 - \left[\sum_{i=1}^{n} \left[\sum_{j=1}^{n} \left[\sum_{j=1}^{n} V_{ij}^2\right.\right.\right.\right.}}$$

Não é difícil verificar que

$$\tau = \frac{F - C}{n(n-1)'}$$

onde F é o número de partes em que a variação de X e Y são no mesmo sentido $(U_{ij} V_{ij} = 1)$, e C é o número de partes em que a variação se dá no sentido inverso $(U_{ij} V_{ij} = -1)$

- Coeficiente generalizado de Kendall – Esta medida generaliza a anterior quando permite a igualdade entre observações das variáveis de interesse. Assim, defina:

$$U_{ij} = \begin{cases} 1\ se\ x_i > x_j \\ 0\ se\ x_i = x_j \\ -1\ se\ x_i < x_j \end{cases}$$

e

$$V_{ij} = \begin{cases} 1 \; se \; y_i > y_j \\ 0 \; se \; y_i = y_j \\ -1 \; se \; y_i < y_j \end{cases}$$

O coeficiente generalizado de Kendall é:

$$\tau_b = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}V_{ij}}{\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} U_{ij}^2\right)\left(\sum_{i=1}^{n} \sum j = 1^n V_{ij}^2\right)}}$$

Coeficiente de Spearman – Outra medida de similaridade entre variáveis ordinais é o Coeficiente de Spearman, denotado por r_s, e calculado pela fórmula anterior, onde

$$U_{ij} = \text{posto}(x_i) - posto(x_j)$$
$$V_{ij} = \text{posto}(y_i) - posto(y_j)$$

É fácil verificar que

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} d_i^2$$

$$\text{onde } d_i = \text{posto}(x_i) - \text{posto}(y_i), \; i = 1, \ldots, n$$

Escolha da Técnica

Outra questão muito importante em Análise de Agrupamento é a escolha da técnica a ser utilizada. Possivelmente a grande variedade de técnicas se deve às diferentes concepções de agrupamento. Um conceito intuitivo é o do

chamado agrupamento natural, que é aquele formado por regiões contendo uma alta densidade de pontos, separados por regiões quase vazias. As Figuras 4.1 e 4.4, ilustram alguns agrupamentos naturais cujas estruturas são óbvias e nem por isso seriam facilmente identificados pela maioria das técnicas. Isso porque os métodos de A.A, fazem suposições implícitas sobre o tipo de estrutura presente nos dados e cabe ao analista verificar se tais suposições são razoáveis para o seu particular conjunto de observações. Por exemplo, o agrupamento da Figura 4.1 seria identificado pela maioria das técnicas porque os grupos são esféricos e estão bastante separados. O método do vizinho mais próximo (M.L.S) aplicado aos dados da Figura 4.4 produziria um resultado equivocado porque o par de objetos mais próximos consiste de uma observação de cada grupo. Este exemplo, ilustra o fato que o M.L.S é incapaz de delinear grupos pouco separados. Por outro lado, esta técnica tem a propriedade de "encadeamento" que é a tendência em produzir grupos do tipo serpentina e possivelmente seria o único a reproduzir o agrupamento da Figura 4.3.

Independentemente da organização dos dados, a minimização do traço de **W** tende a originar grupos homogêneos e esféricos, enquanto a minimização do determinante de **W** tende a produzir grupos com a mesma forma, embora não necessariamente esféricos.

Já as técnicas de mistura de distribuições esbarram em outro problema que é a suposição da normalidade. Ainda é pouco estudado o impacto que a violação dessa hipótese causa na análise.

Como cada método de A.A impõe um certo grau de estrutura nos dados e para se assegurar que o resultado obtido não é um artefato da técnica utilizada, recomenda-se que o analista aplique diferentes critérios de agrupamento e aceite a estrutura resultante da maior parte deles.

Outra maneira de verificar a estabilidade do agrupamento consiste em particionar ao acaso, o conjunto de observações em dois subconjuntos e aplicar o mesmo critério em cada um deles. Se o agrupamento for estável, a alocação dos objetos nas sub-amostras será semelhante àquela na amostra integral.

Avaliação dos Agrupamentos

Nas situações em que os grupos são claramente distintos qualquer uma das estratégias estudadas no Capítulo 3, produzirá o mesmo agrupamento. Porém, mesmo nos casos mais comuns em que tal fato não ocorre, ainda é possível se avaliar o desempenho da Análise de Agrupamento através de

indicadores quantitativos. Far-se-á a apresentação, em separado, para as representações hierárquicas e para os métodos de partição.

Técnicas Hierárquicas

O coeficiente de correlação entre os correspondentes elementos da matriz de similaridade original e a matriz de similaridade produzida por uma classificação hierárquica é o *coeficiente de correlação cofenética*, como já foi visto no Capítulo 2. Esta medida foi inicialmente proposta como uma medida de concordância entre uma classificação obtida e a estrutura original. Posteriormente, esta mesma estratégia de avaliação passou a ser utilizada para se comparar diferentes classificações advindas de um mesmo conjunto de dados. Os resultados de aplicações distintas são comparados calculando-se o coeficiente de correlação entre os elementos correspondentes de suas matrizes de similaridade.

Diversas outras formas de avaliação são utilizadas. Cormack (1971) enumera as seguintes medidas de distorção entre agrupamentos com os respectivos autores:

$$\frac{\sum (s_{ij} - \bar{s})(r_{ij} - \bar{r})}{\sqrt{(s_{ij} - \bar{s})^2 (r_{ij} - \bar{r})^2}} \qquad \textit{Sokal e Rohlf}\ (1962)$$

$$\frac{\sum d_{ij} d^*_{ij}}{\sqrt{d^2_{ij} d^{*2}_{ij}}} \qquad \textit{Guttman}\,(1968)$$

$$\sum d^2_{ij} d^{*2}_{ij} \qquad \textit{Gower}\,(1966,1970)$$

$$\left(\frac{1}{2}\sum abs(S_{ij} - r_{ij})^{i/\alpha}\right)^{\alpha}, 0 < \alpha < 1 \ \textit{Jardine}\,(1967)$$

$$\sum W_{ij}(s_{ij} - r_{ij})^2 \qquad \textit{Hartigan}\,(1967)$$

$$\sum W_{ij}\ \mathrm{d}_{ij} - d^{*2}_{ij} \qquad \textit{Anderson}\,(1971)$$

$$\sum \mathrm{K}(s_{ij} - r_{ij})^2 \qquad \textit{Shepard}\ (1962)$$

$$\sum \frac{d_{ij}}{\sum d_{ij}} d_{ij} - d^{*2}_{ij} \qquad \textit{Sammon}\ (1969)$$

onde r_{ij} e s_{ij} denotam as similaridades e d e d_{ij}^{*} as dissimilaridades entre o par de elementos (i,j) em duas representações distintas. Além disso, foi utilizada a letra *w* para denotar a função peso.

Medida de Similaridade entre Partições

Neste caso, uma das soluções é construir uma tabela de contingência para representar a classificação cruzada dos objetos nas duas partições. Para ilustrar considere duas partições de 10 objetos: a partição 1 com dois grupos definidos por {B, C, F, J} e {A, D, E, G, H, J}, a partição 2 com três grupos: {B, F, H}, {A, C, E, J} e {D, G, I}. A tabela abaixo mostra o cruzamento das duas partições. Dessa forma, as medidas apresentadas anteriormente podem

Tabela 4.3 - Classificação cruzada

P_2 \ P_1	1	2	Total
1	2	1	3
2	1	3	4
3	1	2	3
Total	4	6	10

ser usadas para se avaliar a similaridade entre as duas partições.

Validação: Coeficiente R^2

Em cada passo do algoritmo de agrupamento, é possível calcular a soma de quadrados entre os grupos e dentro dos grupos da partição correspondente. Os critérios de formação de grupos que constituem um agrupamento que mais se destacam, são os critérios de formação de grupos usados na análise de uma matriz de dados contínuos, , que usam a decomposição da matriz de dispersão T, dada por:

$$T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})(x_{ij} - \bar{x})^T$$

e é o vetor de dimensão das médias de cada variável.

$$\bar{x} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}$$

Esta matriz da variabilidade total pode ser decomposta em:

- matriz da dispersão dentro do grupo, W, definida por:

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T$$

em que é o vetor de dimensão das médias das variáveis dentro do grupo .

- matriz da dispersão entre grupos, B, definida por :

$$B = \sum_{j=1}^{k} \sum_{i=1}^{n_j} n_j(\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T, \text{ com } \sum_{j=1}^{k} n_j = n$$

Então $T = B + W$

onde T, W e B são as matrizes associadas à variabilidade total dos dados, à variabilidade dentro dos grupos e à variabilidade entre os grupos, respectivamente.

Para dados univariados, a equação representa a decomposição da soma total dos quadrados da variável em soma dos quadrados dentro dos grupos e a soma dos quadrados entre grupos, que é fundamental na análise de variância.

Como T é fixo, porque não depende do agrupamento que se realize, a melhor partição é aquela em que W é mínimo ou B máximo, isto é quanto

maior a homogeneidade interna dos grupos maior é a separação entre os grupos.

Define-se o coeficiente R^2 da partição como:

$$R^2 = \frac{B}{T}$$

Estatística Pseudo F

Este critério que pode-se utilizar tanto nos métodos hierárquicos como nos não hierárquicos e que baseia-se em uma aproximação F, para compararmos duas soluções de agrupamento.

Conforme Calinski e Harabasz (1974) sugerem, o cálculo da estatística F em cada passo do agrupamento, isto é,

$$F = \frac{B/(k^* - 1)}{W/(n - k)} = \left(\frac{n - k^*}{k^* - 1}\right)\left(\frac{R^2}{1 - R^2}\right)$$

Em que k é o número de grupos relacionado com a partição do respectivo estágio de agrupamento.

Teste de Wilks

A razão (lambda de Wilks), que é a estatística do teste para a hipótese proposta e para a análise de variância simples, resulta do quociente entre os determinantes das matrizes de somas de quadrados e produtos cruzados dentro dos grupos e total.

$$\Lambda^* = \frac{|W|}{|W + B|}$$

Uma análise de variância permite que, comparando-se vários grupos a um só tempo, utilizem-se variáveis contínuas. O teste é paramétrico (a variável de interesse deve ter distribuição normal) e os grupos devem ser independentes.

Índice de Rand Ajustado

O índice de validação externo Rand ajustado é muito utilizado na comparação de algoritmos de agrupamento e oferece como vantagens a independência do número de grupos. O índice de Rand ajustado determina a semelhança entre duas parcelas P_1 e P_2 examinando a qual grupo pares de espécies pertencem nos dois grupos. Isso quer dizer que se duas espécies pertencerem ao mesmo grupo P_1 e P_2 o valor do índice aumenta; por outro lado, se as duas espécies pertencerem, ao mesmo grupo em P_1 mas pertencem a grupo diferentes em P_2 o valor do índice diminui. O índice de Rand ajustado é a versão normalizada do índice Rand, onde: e são os números de grupos das parcelas P_1 e P_2; n é a quantidade de dados do conjunto inicial; é o número de espécies do grupo e é o número de espécies do grupo é o número de espécies que pertencem aos grupos e , ou seja, o número de espécies comuns a P_1 e P_2.

$$Rand\ ajustado = \frac{A - B}{C - D}$$

$$Rand\ ajustado = \frac{\sum_{i=1}^{k_{P_1}} \sum_{j=1}^{k_{P_2}} \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_{i=1}^{k_1} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P_2 1}} \binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k_{P_1}} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P_2}} \binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1} \sum_{i=1}^{k_{P_1}} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P_2}} \binom{n_{.j}}{2}}$$

Valores próximos a 0 para índice de Rand ajustado indicam parcelas aleatórias, que pouco revelam sobre a relação entre as espécies, enquanto valores próximos a 1 são obtidos por parcelas mais relevantes.

Método Silhueta

Dentre as abordagens existentes para auxiliar na decisão do número de grupos, foi utilizado o método Silhueta, silhouette, proposto por (ROUSSE-

EUW, 1987), que subsidia na escolha de um número ótimo de grupos, avaliando os particionamentos encontrados e permite visualizar graficamente os agrupamentos.

A silhueta é um gráfico do cluster C composto por um valor de silhueta s(i), i =1; ..., n, que reflete a qualidade da alocação dos objetos nos grupos. Cada objeto (indivíduo) do cluster é representado por i. E para cada objeto i o valor s(i) é calculado (Equação 1):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Onde a(i) é a dissimilaridade média do objeto i em relação a todos os objetos do mesmo grupo C, e b(i) é a dissimilaridade média entre o objeto i em relação a todos os objetos do grupo vizinho mais próximo a ele, grupo X.

O valor de s(i) varia no intervalo entre -1 e 1, sendo adimensional. Quando um valor de $s(i) \approx 1$, significa que o objeto i foi bem classificado no grupo C, pois a(i) < b(i). Se o valor de $s(i) \approx -1$, significa que o objeto foi mal classificado, pois a(i) > b(i), ou seja, o objeto i, em média, está mais distante dos objetos do seu próprio grupo, isto é, o objeto do grupo C está mais próximo dos objetos do grupo X. Por sua vez, se $s(i) \approx 0$, o objeto **i** está entre os grupos C e X, isso ocorre quando a(i) = b(i), indicando que o objeto está num ponto intermediário a dois grupos. Logo, quanto mais próximo a 1, melhor será a qualidade do agrupamento (SOUZA, 2007).

Uma interpretação subjetiva para este método foi proposta por (KAUFMAN; ROUSSEEUW, 1990), que subsidia na avaliação do agrupamento encontrado (Tabela 3). O coeficiente de silhueta (CS(i)) é uma medida de qualidade para toda estrutura de agrupamento que foi descoberta pelo algoritmo de classificação.

Dados Artificiais

Os dados consistem de n parcelas simuladas a partir de sua distribuição normal bivariado com vetores médios () e da matriz de variância e covariância original.

$$\sum = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$$

onde

$$s_{11} = var(x_1) \; s_{22} = var(x_2) \, e \, s_{12} = \; s_{21} = cov(x_1, \; x_2)$$

Sumário

O analista interessado na aplicação de Análise de Agrupamento tem à sua disposição uma diversidade de técnicas, como foi visto no Capítulo anterior. Vários problemas associados com estas técnicas, bem como estratégias para contorná-los, foram objeto de estudo deste Capítulo. Foram também discutidos alguns métodos para validação de agrupamentos.

CAPÍTULO 5

Suporte Computacionais e Aplicação

Introdução

Neste Capítulo são apresentados vários programas de computador disponíveis para Análise de Agrupamento e uma aplicação para discussão de conceitos introduzidos nos Capítulos anteriores.

A técnica Análise de Agrupamento foi implementada utilizando o ambiente R. O R é uma linguagem de programação que permite manipular dados, fazer cálculos e construir gráficos estatístico. Caracteriza-se como um sistema complementar planejado e coerente e não apenas um conjunto ampliado de ferramentas muito especificas e inflexíveis, como são frequentemente em outros programas de análise de dados, como SAS, MINITAB etc. O R é um veículo para o desenvolvimento de novos métodos interativos de dados.

O software pode ser obtido pelo site do CRAN(R, 2018). No CRAN é possível baixar não só o pacote principal do R, mas também os pacotes opcionais chamados de contribuídos (Contributed Packages) e uma série de manuais. O R dispõe de um grupo de discursão na internet, R_Stat, onde muitas dúvidas do funcionamento do software podem ser sanadas, além de possibilitar a troca de experiências pelos participantes.

Com o uso deste software, soluciona-se o problema de indisponibilidade de pacotes estatísticos específicos para aplicação dos diversões métodos da Análise de Agrupamento, pois além de gratuito é amigável, aberto e com inúmeros recursos disponíveis.

Aplicação

Para finalizar, apresentamos nesta seção uma análise do exemplo sobre alimentos, introduzido na seção 4.2., com o objetivo de discutir conceitos apresentados neste livro.

Conforme referido anteriormente, os dados apresentados na Tabela 5.1, correspondem a quantidades de energia, proteína, gordura, cálcio e ferro encontradas em 3 onças de peso de 27 diferentes tipos de alimentos preparados com carnes e peixes. A Tabela 5.2, apresenta algumas estatísticas básicas relativas aos cinco atributos observados.

Tabela 5.1: Dados de Nutrientes em Carnes e Pexie. (A = Assado, C = Cozido, D = Defumado, E = Enlatado, F = Frito, FE = Fervido, G = Grelhado)

Descrição dos alimentos	Energia (cal)	Proteína (g)	Gordura (g)	Cálcio (mg)	Ferro (mg)
Bife	340	20	28	9	2.6
Hambúrguer	245	21	17	9	2.7
Carne/A	420	15	39	7	2.0
Churrasco	375	19	32	9	2.6
Carne/E	180	22	10	19	3.7
Galinha/G	115	20	3	8	1.4
Galinha/E	170	25	7	12	1.5
Coração	160	26	5	14	5.9
Perna Ovelha/A	265	20	20	9	2.6
Diant. Ovelha/A	300	18	25	9	2.3
Presunto/D	340	20	28	9	2.5
Porco/A	340	19	29	9	2.5
Porco/FE	355	19	30	9	2.4
Língua	205	18	14	7	2.5
Costela Vitela	185	23	9	9	2.7

Bluefish/C	135	22	4	25	0.6
Marisco/CRU	70	11	1	82	6.0
Marisco/E	45	7	1	74	5.4
Siri/E	90	14	2	38	0.8
Haddock/F	135	16	5	15	0.5
Cavala/G	200	19	13	5	1.0
Cavala/E	155	16	9	157	1.8
Perca/F	195	16	1	14	1.3
Salmão/E	120	17	5	159	0.7
Sardinha/E	180	22	9	367	2.5
Atum/E	170	25	7	7	1.2
Camarão/E	110	23	1	98	2.6

Tabela 5.2: Estatísticas Básicas para os Dados Originais

Atributo	Mínimo	Máximo	Média	Desvio Padrão
Energia	45.0	420	207.41	101.21
Proteína	7.0	26	19.00	4.25
Gordura	1.0	39	13.48	11.26
Cálcio	5.0	367	43.96	78.03
Ferro	0.5	6	2.38	1.46

Com a finalidade de agrupar os alimentos em conglomerados de modo a obter alimentos próximos entre si, em algum sentido, dentro dos conglomerados e alimentos afastados entre si em diferentes conglomerados, os atributos energia, proteína, cálcio e ferro foram relativizados em relação aos valores diários recomendáveis: 3.200 calorias para energia alimentar, 70 g para proteína, 800 mg para cálcio e 10 mg para ferro.

Esta transformação nos dados surge do fato das necessidades diárias recomendadas para os atributos serem diferentes, implicando em grande variação nos valores dos respectivos desvios padrões, conforme pode ser visto na Tabela 5.2 acima.

As Tabelas 5.3 e 5.4 apresentam, respectivamente, os dados e algumas estatísticas básicas para os atributos transformados. Da Tabela 5.4, podemos ver que os valores dos desvios padrões associados aos atributos transformados estão relativamente próximos entre si e ao valor do desvio padrão associado ao atributo gordura que não sofreu transformação.

Tabela 5.3: Porcentagem de Nutrientes em Carnes e Peixe em Relação à Quantidade Diária Recomendável, Exceto para Gordura. (A = Assado, C = Cozido, D = Defumado, E = Enlatado, F = Frito, FE = Fervido, G = Grelhado)

Descrição dos alimentos	Energia %	Proteína %	Gordura %	Cálcio %	Ferro %
Bife	11	29	28	1	26
Hambúrguer	8	30	17	1	27
Carne/A	13	21	39	1	20
Churrasco	12	27	32	1	26
Carne/E	6	31	10	2	37
Galinha/G	4	29	3	1	14
Galinha/E	5	36	7	2	15
Coração	5	37	5	2	59
Perna Ovelha/A	8	29	20	1	26
Diant. Ovelha/A	9	26	25	1	23
Presunto/D	11	29	28	1	25
Porco/A	11	27	29	1	25
Porco/FE	11	27	30	1	24
Língua	6	26	14	1	25
Costela Vitela	6	33	9	1	27
Bluefish/C	4	31	4	3	06
Marisco/CRU	2	16	1	10	60
Marisco/E	1	10	1	9	54
Siri/E	3	20	2	5	08
Haddock/F	4	23	5	2	05
Cavala/G	6	27	13	1	10
Cavala/E	5	23	9	20	18
Perca/F	6	23	1	2	13
Salmão/E	4	24	5	20	07
Sardinha/E	6	31	9	46	25
Atum/E	5	36	7	1	12
Camarão/E	3	33	1	12	26

Tabela 5.4: Estatísticas Básicas para os Dados Transformados

Atributo	Mínimo	Máximo	Média	Desv. Padrão
Energia	1	13	6,48	3,25
Proteína	10	37	27,18	6,08
Cálcio	1	46	5,52	9,78
Ferro	5	60	23,81	14,61

O método hierárquico das médias das distâncias e o não-hierárquico das K-médias foram aplicados aos dados da Tabela 5.3 com a distância euclideana reduzida como medida de parecença e os resultados estão mostrados abaixo.

Método das Médias das Distâncias (M.M.D)

A Tabela 5.5, mostra como os conglomerados foram formados com este método

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Bife	Carne/E	Marisco/CRU	Cavala/E	Sardinha/E
Hambúrguer	Galinha/G	Marisco/E	Salmão/E	
Carne/A	Galinha/E	Camarão/E		
Churrasco	Coração			
Perna Ovelha/A	Língua			
Diant. Ovelha/A	Costela Vitela			
Presunto/D	Bluefish/C			
Porco/A	Siri/E			
Porco/FE	Haddock/F			
	Cavala/G			
	Perca/F			
	Atum/E			

Tabela 5.5: Dados da Formação dos Conglomerados com o Método das Médias das Distâncias (M.M.D)

1	0,4472	Bife	Presunto/D
2	0,6325	Porco/A	Porco/FE
3	1,2466	Cluster 1	Cluster 2
4	1,4142	Galinha/E	Atum/E
5	1,4832	Hambúrguer	Perna Ovelha/A
6	1,7430	Cluster 3	Churrasco
7	2,4495	Perca/F	Cavala/G
8	2,6368	Cluster 6	Diant. Ovelha/A
9	2,7203	Haddock/F	Siri/E
10	2,8657	Cluster 5	Língua
11	3,7147	Galinha/G	Cluster 4
12	3,8471	Marisco/CRU	Marisco/E
13	4,1771	Cluster 11	Bluefish/C
14	4,4010	Cluster 10	Costela Vitela
15	5,1793	Cluster 7	Cluster 9
16	5,2726	Cavala/E	Salmão/E
17	5,7625	Cluster 13	Cluster 15
18	5,7634	Cluster 14	Carne/E
19	6,0306	Cluster 8	Carne/A
20	8,1340	Cluster 19	Cluster 18
21	8,9608	Camarão/E	Cluster 16
22	9,3752	Cluster 21	Cluster 17
23	11,6141	Coração	Cluster 12
24	12,0880	Cluster 20	Cluster 22
25	20,3549	Cluster 24	Cluster 23
26	20,9040	Cluster 25	Sardinha/E

A representação gráfica desta formação de conglomerados e a evolução do valor do nível de junção dos objetos podem ser vistas nos gráficos 5.1 e 5.2, respectivamente.

Gráfico 5.1: Representação Gráfica (Dendrograma) da Formação dos Conglomerados pelo Método das Distâncias Médias (M.M.D)

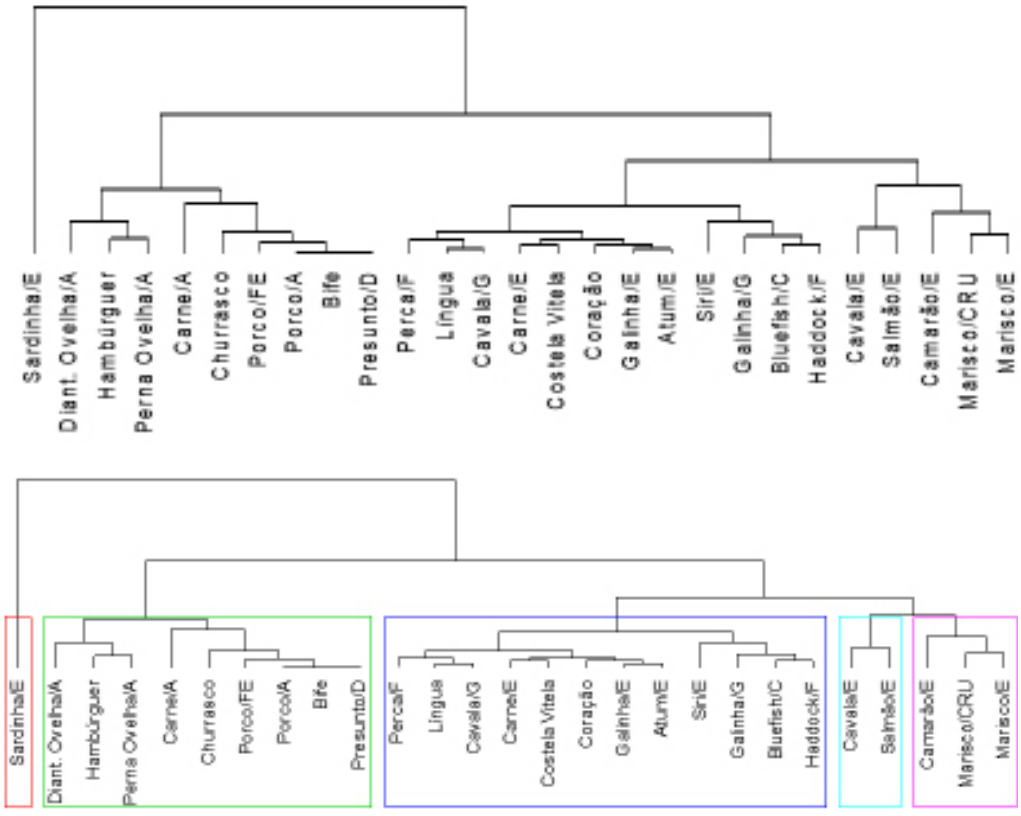
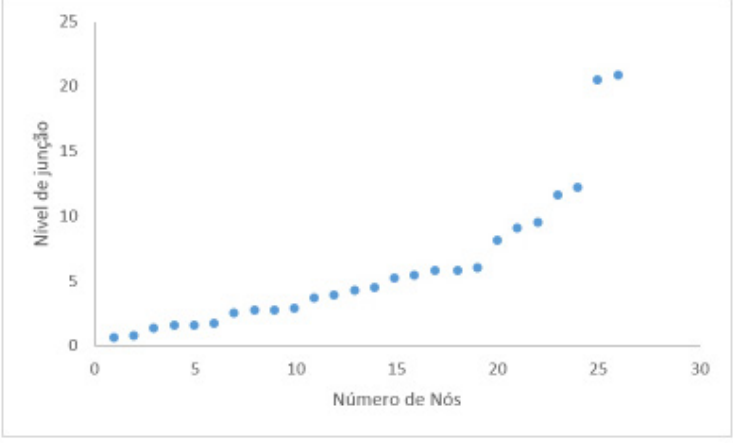


Gráfico 5.2: Representação Gráfica da Evolução do Valor do Nível de Junção dos Objetos ao Longo dos Nós



Do dendrograma apresentado acima podemos ver que existem pelo menos dois objetos (sardinha/E e coração) que são discrepantes ("outiliers") dos demais objetos. Observando a Tabela 5.3, podemos verificar que os valores 46% de cálcio para o alimento sardinha/E e 37% de proteína e 59% de ferro para o alimento coração, devem ser os responsáveis por isto.

Além da possibilidade de detecção de "outliers", podemos também fazer uso do dendrograma para escolher o número apropriado de conglomerados a ser utilizado para agrupar os alimentos. Cortes no dendrograma correspondentes aos valores 6,04, 9,38 e 12,09 do nível de junção dos objetos nos fornecem a indicação de 8, 5 ou 3 conglomerados respectivamente. Este fato pode ser melhor visualizado no gráfico 5.2 a partir dos saltos que ocorrem nas passagens dos níveis 19 para 20, 21 e para 22, 23 e 24.

Método das *K*-Médias

O método das K-médias é um método de partição e, como referido anteriormente, fornece indicações mais precisas sobre o número de conglomerados a ser formado. As Tabelas 5.6, 5.7 e 5.8 apresentam Tabelas de análises de variância efetuadas para cada um dos atributos a partir da aplicação deste método para 2, 3 e 4 conglomerados.

Deste modo, podemos ver na Tabela 5.6, que para 2 conglomerados, os atributos proteína e cálcio não contribuem na discriminação entre os dois conglomerados (níveis descritivos 0,544 e 0,51) enquanto que o atributo ferro é o que mais contribui para esta discriminação (nível descritivo .000). Para 3 ou 4 conglomerados o atributo proteína continua não contribuindo na discriminação enquanto que o atributo cálcio passa a ter um papel importante nesta discriminação (ver Tabela 5.7 e 5.8).

Tabela 5.6: Análise de Variância para 2 Conglomerados

Atributo	Soma Quad. Entre	gl	Soma Quad. dentro	gl	Valor F	Nível Desc.
Energia	77,82	1	196,92	25	9,88	,004
Proteína	14,34	1	945,73	25	,39	,544
Gordura	1.103,67	1	2.191,07	25	12,59	,002
Cálcio	356,89	1	2.129,85	25	4,19	,051
Ferro	2.410,89	1	3.141,18	25	19,19	,000
Total	3.963,61		8.604,75			

Tabela 5.7: Análise de Variância para 3 Conglomerados

Atributo	Soma dos quadrados entre	gl	Soma dos quadrados dentro	Gl	Valor F	Nível Descritivo
Energia	77.87	2	196.88	24	4.75	0.018
Proteína	14.72	2	945.36	24	0.19	0.831
Gordura	1104.34	2	2190.40	24	6.05	0.007
Cálcio	1691.93	2	794.81	24	25.55	0.000
Ferro	2609.26	2	2942.81	24	10.64	0.000
Total	5498,12		7070.26			

Tabela 5.8: Análise de Variância para 4 Conglomerados

Atributo	Soma dos quadrados entre	gl	Soma dos quadrados dentro	gl	F	Nível descritivo
Energia	184,53	3	90,21	23	15,68	,000
Proteína	129,53	3	830,54	23	1,20	,333
Gordura	2.171,16	3	1.123,58	23	14,82	,000
Cálcio	1.691,93	3	710,79	23	19,16	,000
Ferro	2.609,26	3	523,46	23	73,65	,000
Total	5.498,12		3.278,58			

-

Utilizando os resultados da seção 4.4 e o total da soma de quadrados dentro de grupos como estimativa do traço da matriz de dispersão dentro dos grupos podemos construir testes de hipóteses associados à determinação do número de conglomerados. Seja R(k + 1/k) a estatística associada ao teste para se passar do número de conglomerados k para k + 1. Então, como referido na seção 4.4, teremos

R(3/2) = ((8.604,75/7.070,26) – 1)(27 – 2 – 1) = 5,21

R(4/3) = ((7.070,26/3.278,58) – 1)(27 – 3 – 1) = 26,60

R(5/4) = ((3.278,58/2.307,83) – 1)(27 – 4 – 1) = 9,25 e

R(6/4) = ((2.307,83/2.127,34) – 1)(27 – 5 – 1) = 1,78

Os valores para calcular R(3/2) e R(4/3) foram obtidos das Tabelas 5.6, 5.7 e 5.8. Os correspondentes valores para R(5/4) e R(6/5) não estão apresentados aqui.

Assumindo que a estatística R(k + 1/k) tem distribuição F com p e n – k – 1 graus de liberdade, neste exemplo p = 5 e n = 27, o valor do nível descritivo associado. Deste modo, só rejeitaríamos a passagem de 5 para 6 conglomerados.

Construção dos Conglomerados

A partir dos resultados obtidos acima podemos concluir que 5 é um número razoável de conglomerados a ser considerado neste exemplo. Assim, da aplicação dos 2 métodos propostos obtivemos a seguinte distribuição dos alimentos em 5 conglomerados:

1. Método das Distâncias Médias (M.M.D)

Grupo 1	Grup*o 2*	*Grupo 3*	*Grupo 4*	*Grupo 5*
Bife	Carne/E	Marisco/ CRU	Cavala/E	Sardinha/E
Hambúrguer	Galinha/G	Marisco/E	Salmão/E	
Carne/A	Galinha/E	Camarão/E		
Churrasco	Coração			
Perna Ovelha/A	Língua			
Diant. Ovelha/A	Costela Vitela			
Presunto/D	Bluefish/C			
Porco/A	Siri/E			
Porco/FE	Haddock/F			
	Cavala/G			
	Perca/F			
	Atum/E			

2. Método das K-Médias

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Bife	Hambúrguer	Carne/E	Marisco/CRU	Cavala/E
Carne/A	Perna Ovelha/A	Galinha/G	Marisco/E	Salmão/E
Churrasco	Diant. Ovelha/A	Galinha/E	Siri/E	Sardinha/E
Presunto/D		Coração	Camarão/E	
Porco/A		Língua		
Porco/FE		Costela Vitela		
		Bluefish/C		
		Haddock/F		
		Cavala/G		
		Perca/F		
		Atum/E		

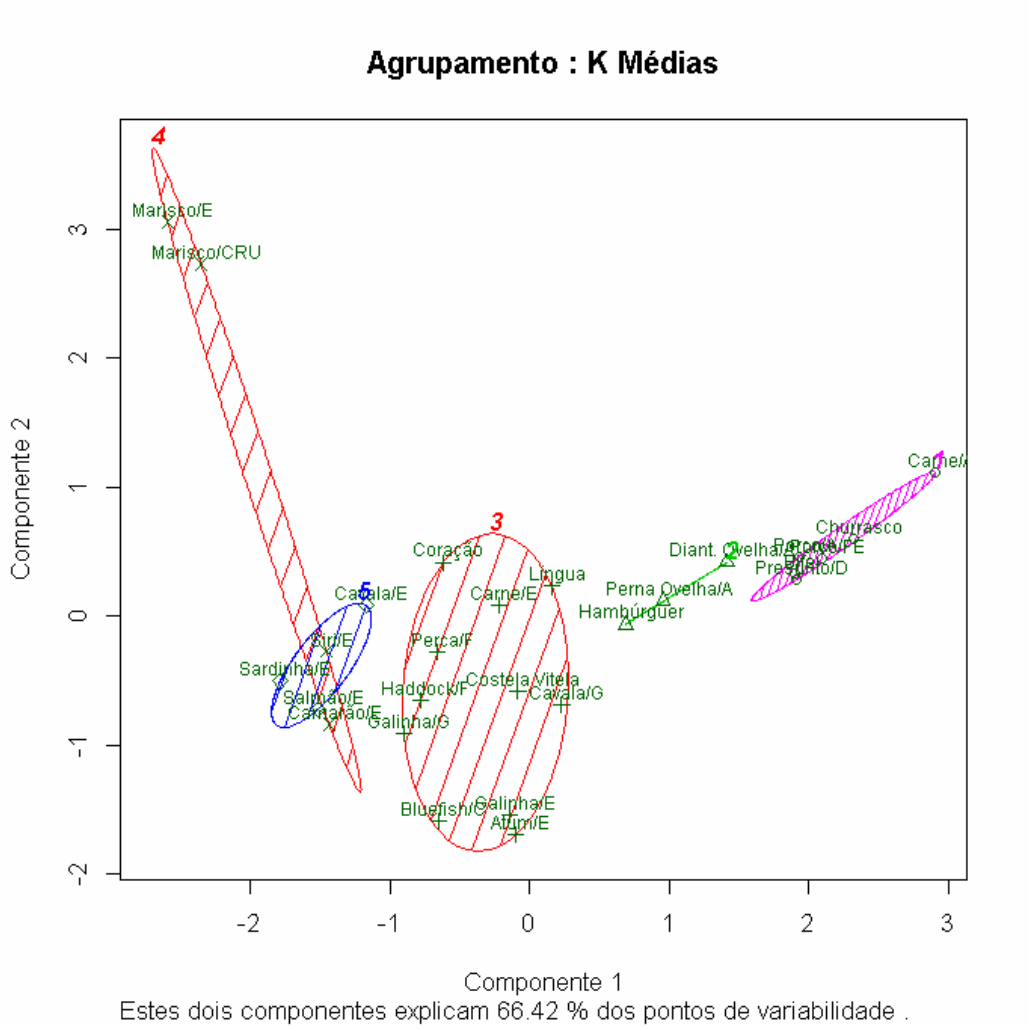
Conglomerado 1: Bife, Presunto/D, Porco/A, Porco/FE, Churrasco, Diant. Ovelha/A, Carne/A, Perna Ovelha/A.

Conglomerado 2: Galinha/G, Galinha/E, Atum/E, Bluefish/C, Perca/F, Cavala/G, Haddock/F, Siri/E.

Conglomerado 3: Hambúrguer, Carne/E, Língua, Costela Vitela, Camarão/E.

Conglomerado 4: Marisco/CRU, Marisco/E, Coração.

Conglomerado 5: Sardinha/E, Salmão/E, Cavala/E.

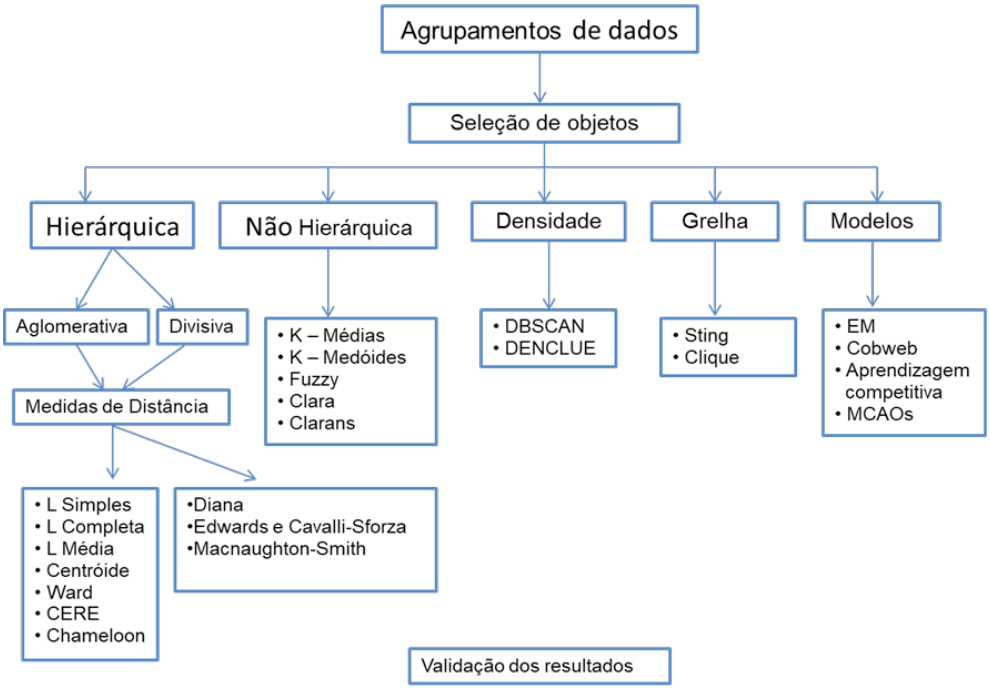


A Tabela 5.9, mostra como os objetos alocados por um dos métodos foram distribuídos por outro método de Análise de Agrupamento nos 5 conglomerados.

Tabela 5.9: Frequência da Distribuição Conjunta dos Objetos nos 5 Conglomerados pelos 2 Métodos

K-médias	1	2	M.M.D 3	4	5	Total
1	8	0	0	0	0	8
2	0	8	0	0	0	8
3	4	1	0	0	0	5
4	0	0	1	2	0	3
5	0	2	0	0	1	3
Total	12	1 1	1	2	1	27

Da tabela acima podemos ver que 19 dos 27 objetos (70%) foram alocados nos mesmos conglomerados pelos 2 métodos. Medidas de concordância poderiam ser utilizadas para melhor estudar esta tabela. Distâncias de cada um dos objetos ao centro do conglomerado ao qual ele foi alocado e ao centro dos outros conglomerados, bem como outras estatísticas, podem ser úteis para verificar se um ou mais objetos foram indevidamente alocados.



CAPÍTULO 6

Comandos em R

Em R, essas medidas estão disponíveis na função dist(). Por exemplo, usando o data set “USArrest” (pacote datasets), # podemos calcular a distância euclidiana das primeiras cinco.

Observações multivariadas da seguinte forma:

```
#Matriz de distâncias
d<-dist(USArrests[1:5,], method = “euclidean”)
d
ou
dist(USArrests[1:7,], method = “euclidean”); round(d,2) #duas classes decimais
```

```
Alabama    Alaska   Arizona  Arkansas
Alaska     37.17701
Arizona    63.00833  46.59249
Arkansas   46.92814  77.19741 108.85192
California 55.52477  45.10222  23.19418  97.58202
```

#Podemos observar que, em termos de indicadores de violência (variáveis), #os Estados do Arkansas e do Arizona são os mais distintos dos cinco avaliados.

A distância Mahalanobis pode estar disponível no pacote biotools,

#install.packages("biotools", por meio da função D2.dist().

Utilizando o exemplo anterior.

```
#install.packages("biotools"),
library(biotools)
S<-cov(USArrests)
S
```

```
 Murder   Assault   UrbanPop     Rape
Murder   18.970465  291.0624   4.386204  22.99141
Assault  291.062367 6945.1657 312.275102 519.26906
UrbanPop   4.386204  312.2751 209.518776  55.76808
Rape     22.991412  519.2691  55.768082  87.72916
```

```
D2.dist(data=USArrests[1:5,],cov=S); round(D2,2)
```

```
Alabama Alaska Arizona Arkansas
Alaska      19.33
Arizona      9.97  15.01
Arkansas     2.01  12.29    7.29
California  12.42  12.67    3.04    10.61
```

#Observa-se agora que, como as variáveis têm variância bem diferentes

e ainda estão correlacionadas, os Estados correlacionados, os Estados mais distintos são Alaska e Alabama.

```
# Utilizando a matriz de distância de Mahalanobis, podemos realizar o
# agrupamento de Tocher com os seguintes comandos:
 D2<-D2.dist(data=USArrests[1:5,],cov=S)
 tocher(D2, algorithm = "original")
```

```
Tocher's Clustering
Call: tocher.dist(d = D2, algorithm = "original")
Cluster algorithm: original
Number of objects: 5
Number of clusters: 2
Most contrasting clusters: cluster 1 and cluster 2, with
  average intercluster distance: 14.82547
$`cluster 1`
[1] Alabama   Arkansas  Arizona   California
$`cluster 2`
[1] Alaska
```

Métodos Hierárquicos

Utilizando o objeto D2, podemos realizar agrupamento hierárquicos com os quatro métodos da seguinte forma:

```
require(cluster)    # require=library
hc1<- hclust(D2, method = “single”); plot(hc1,hang=-1) # Método ligação mais próximo
hc2<- hclust(D2, method = “complete”); plot(hc2,hang=-1) # Método ligação mais distante
hc3<- hclust(D2, method = “average”); plot(hc3,hang=-1)# # Método ligação média
hc4<- hclust(D2, method = “ward.D2”); plot(hc4,hang=-1)# # Método Ward
```

#Para ajustar o comprimento dos ramos, devemos utilizar o argumento “hang -1”:

plot(cluster,hang=-1)

Observação:

Versos do R > 3.0.3 apresentam dois algoritmos para agrupamento de **Ward**. A opção **ward.D** equivale à antiga **ward** e não corresponde ao método original (WARD, 1963) que foi recentemente implementado na opção **ward.D2**.

Os dendrogramas podem ser obtidos fazendo:

```
par(mfrow = c(2,2))
 plot(hc1)
 plot(hc2)
 plot(hc3)
 plot(hc4)
```

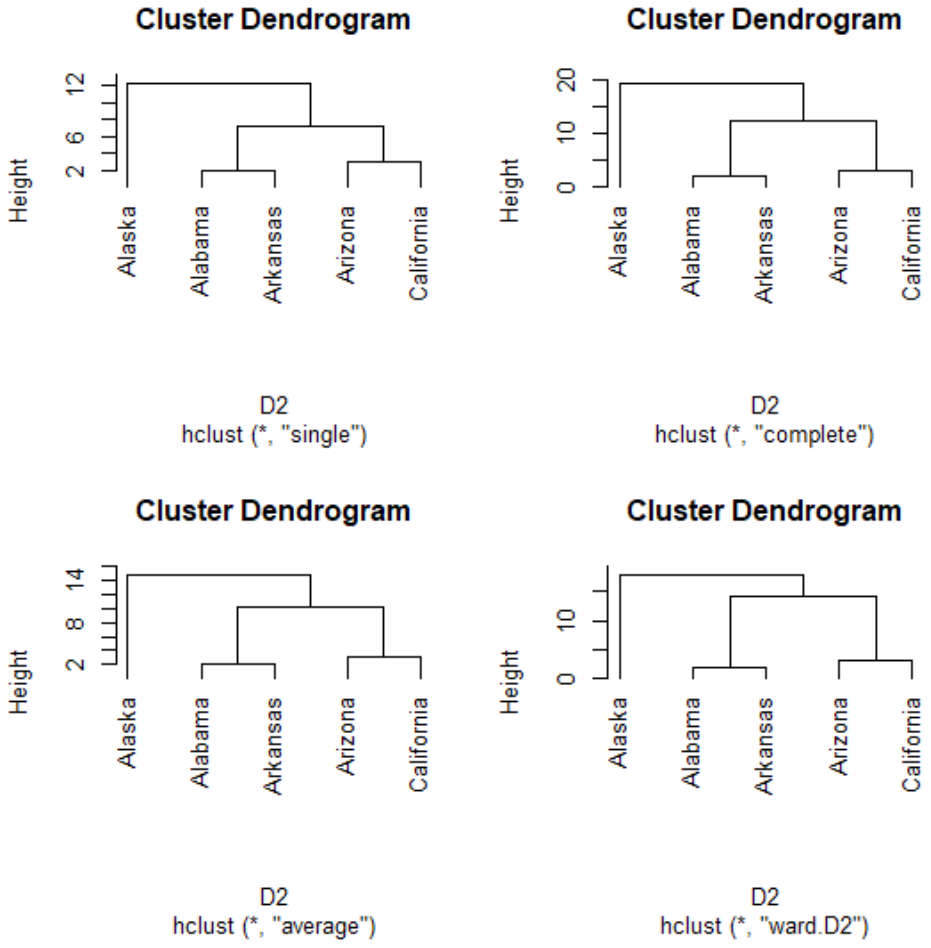


Fig.: Dendrogramas obtidos com métodos: ligação simples, completa, média e Ward, com base na distância Mahalanobis dos cinco primeiros Estados dos dados "USArrets".

Correlação Cofenética

Apesar de ser largamente utilizada para avaliar resultados de agrupamentos hierárquicos, foi proposto por Silva e Dias (2013) um algoritmo para obter a correlação cofenética para o método de Tocher, no qual as distâncias coféneticas são baseadas nas distâncias médias intra e intergrupos.

Exemplo de Agrupamento Tocher

Utilizando a matriz de distância (objeto D2), podemos calcular a correlação cofenética do agrupamento de Tocher da seguinte forma:

```
#agrupamento
toc<- tocher(D2)
# matriz cofenética
dist.tocher<-cophenetic(toc)
dist.tocher
```

```
> toc<- tocher(D2) #agrupamento
> dist.tocher<-cophenetic(toc) # matriz cofenética
> dist.tocher
Alabama    Alaska   Arizona  Arkansas
Alaska     14.825474
Arizona    7.557493 14.825474
Arkansas   7.557493 14.825474  7.557493
California 7.557493 14.825474  7.557493  7.557493
Ou com duas casas decimais
round(dist.tocher,2)
Alabama Alaska Arizona Arkansas
Alaska      14.83
Arizona      7.56  14.83
Arkansas     7.56  14.83   7.56
California   7.56  14.83   7.56    7.56
```

Correlação Cofenética do Agrupamento

Em agrupamento hierárquicos, a matriz cofenética é obtida com a função cophenetic(). As correlações coféneticas para os quatro métodos apresentados e utilizados.

```
dist.hc1<-cophenetic (hc1)
dist.hc2<-cophenetic (hc2)
dist.hc3<-cophenetic (hc3)
dist.hc4<-cophenetic (hc4)
cor(D2, dist.hc1)
cor(D2, dist.hc2)
cor(D2, dist.hc3)
cor(D2, dist.hc4)
```

```
> dist.hc1<-cophenetic (hc1)
> dist.hc2<-cophenetic (hc2)
> dist.hc3<-cophenetic (hc3)
> dist.hc4<-cophenetic (hc4)
> cor(D2, dist.hc1)
[1] 0.8926901
> cor(D2, dist.hc2)
[1] 0.9042972
> cor(D2, dist.hc3)
[1] 0.9048841
> cor(D2, dist.hc4)
[1] 0.8871881
```

Observamos então que os métodos de ligação simples, ligação completa, ligação médias apresentaram resultados com valores altos, produzindo dendrogramas com representação alta. Já o método de Ward, por ser baseado na minimização da soma de quadrados dentro dos grupos, é esperado que apresente correlação cofenética inferior aos demais.

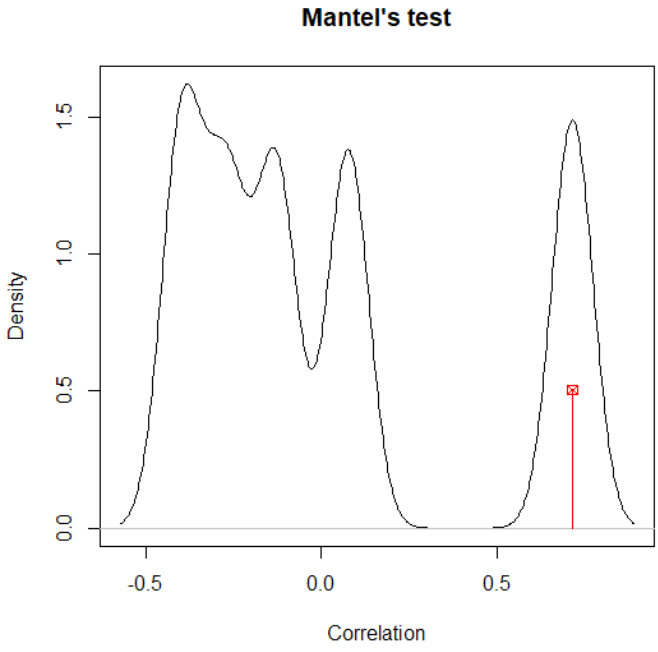
Teste de Mantel

As hipóteses $H_0 : \square_{cof} = 0$
$H_0 : \square_{cof} > 0$

Podem ser avaliadas pelo teste da aleatorização de Mantel (1967). O teste é baseado na distribuição empírica da correlação cofenética obtida a partir de aleatorizações de uma das duas matrizes de distâncias envolvidas. Em geral, o teste é unilateral à direita, pois é esperado que as matrizes de distância sejam positivamente correlacionadas.

No pacote **biotools** há uma implantação desse teste. Por exemplo: a correlação cofenética obtida com o agrupamento de Tocher (0.71) pode ser testada por meio de:

```
#O teste de Mantel
 mantelTest(D2, dist.tocher, nperm =999)
```



```
mantelTest(D2, dist.tocher, nperm =999)
Mantel's permutation test
Correlation: 0.7146411
p-value: 0.218, based on 999 matrix permutations
Alternative hypothesis: true correlation is greater than 0
```

Podemos concluir que a correlação cofenética obtida com o método de Tocher não difere de zero (p = 0.218), de acordo com o teste de Mantel, com base em 999 permutações.

Silva et al. (2015) apresentam uma alternativa simples e rápida para análise do poder do teste de Mantel. A metodologia está implementada na função **mantelPower** do pacote **biotools.**

```
mt1 <- mantelTest(D2, dist.tocher, xlim = c(-1, 1))
mantelPower(mt1, effect.size = 0.3)
```

Outros Pacotes

```
####################
Install.packages("cluster"); (vegan); (ecodist); (MASS); (pvclust);  (fpc)
library(cluster); library(vegan)
library(ecodist);  library(MASS)
library(pvclust);  library(fpc)
options(digits=3)
```

Exemplo: Dados do atlas de desenvolvimento humano (2010).

```
IDMH=c( 0.577, 0.800, 0.712,
        0.559, 0.777, 0.671,
       0.561, 0.805, 0.677,
```

```
0.628, 0.809, 0.695,
0.528, 0.789, 0.646,
0.629, 0.813, 0.694,
0.624, 0.793, 0.690,
0.562, 0.757, 0.612,
0.547, 0.777, 0.635,
0.615, 0.793, 0.651,
0.597, 0.792, 0.678,
0.555, 0.783, 0.656,
0.574, 0.789, 0.673,
0.520, 0.755, 0.641,
0.560, 0.781, 0.672,
0.555, 0.783, 0.663,
0.638, 0.838, 0.730,
0.653, 0.835, 0.743,
0.675, 0.835, 0.782,
0.719, 0.845, 0.789,
0.668, 0.830, 0.757,
0.697, 0.860, 0.773,
0.642, 0.840, 0.769,
0.629, 0.833, 0.740,
0.635, 0.821, 0.732,
0.646, 0.827, 0.742,
0.742, 0.873, 0.863)

IDMH=matrix(IDMH,nrow=3,ncol=27,byrow=TRUE)
y=c("RO","AC","AM","RR","PA","AP","TO","MA","PI","CE",
"RN","PB","PE","AL","SE","BA","MG","ES","RJ","SP","PR",
 "SC","RS","MS","MT","GO","DF")
```

```
 colnames(IDMH) <- y
IDMH=t(IDMH)
#####################################
IDMH=t(IDMH)
## escolher qual a distância que vai utilizar #euclidean bray-curtis man-hattan mahalanobis  jaccard  simple difference sorensen Partial
## escolher qual a método que vai utilizar
# “ward.D”, “single”, “complete”, “average”, “mcquitty”, “median” or “centroid”
 dist1<-distance(IDMH,”euclidean”)
 clust1<-hclust(dist1,”single”)
 plot(clust1, hang=-0.5, main=”,xlab=”distâncias”)
 y<-rect.hclust(clust1, k=5, border=”black”);y  ##ou red=vermelho k= número de grupos
#Matriz Cofenética
 #options(digits=3)
 #cophenetic(clust1)
cor(dist1,cophenetic(clust1))
var(cophenetic(clust1))
```

Algumas funções úteis, em se tratando de agrupamento hierárquicos (função hclust()), são:

rect.hclust()

identify

cutree()

as.dendrogram()

cut()

heatmap()

Outras formas de visualização de dendrograma estão disponíveis no pacote **ape** (PARADIS, 2004). Veja alguns exemplos na Figura cujos códigos necessários são apresentados a seguir:

```
D2<-D2.dist(data=USArrests[1:12,], cov=S)
hc1<- hclust(D2, method = "ward.D2")
grupos<-cutree(hc1, k=3)
hc1.phylo<-as.phylo(hc1)
par(mfrow=c(3,2))
plot(hc1.phylo,type="phylogram", tip.color=grupos)
plot(hc1.phylo,type="phylogram", tip.color=grupos)
plot(hc1.phylo,type="cladogram", tip.color=grupos)
plot(hc1.phylo,type="fan", tip.color=grupos)
plot(hc1.phylo,type="unrooted", tip.color=grupos)
plot(hc1.phylo,type="radial", tip.color=grupos)
```

Para Programa Não Hierárquico I

```
#pacotes para realizar Análise de Agrupamento, método não hierárquico
#install.packages("MVA) # instalar o pacote
library(MVA)
library(cluster)
x = c( 0.88  , 0.90  , 0.90, 0.87 , 0.93 , 0.89 , 0.88 , 0.81 , 0.82 , 0.85, 0.77 ,0.71, 0.75, 0.70, 0.44,
0.47, 0.23, 0.34, 0.31, 0.24, 0.76, 0.99  , 0.99  , 0.98, 0.98 , 0.93 , 0.97 , 0.87 , 0.92 , 0.92 , 0.90,
0.85 ,0.83, 0.83, 0.62, 0.58, 0.37, 0.33, 0.36, 0.35, 0.37, 0.80, 0.91  , 0.93  , 0.94, 0.97 , 0.93 , 0.92 ,
0.91 , 0.80 , 0.75 , 0.64, 0.69 ,0.72, 0.63, 0.60, 0.37, 0.45, 0.27, 0.51, 0.32, 0.36, 0.61, 1.10  , 1.26  ,
1.24, 1.18 , 1.20 , 1.24 , 1.41 , 0.55 , 1.05 , 0.07,-1.36 ,0.47,-0.87, 0.21,-1.36,-0.68,-1.26,-1.98,-0.55,
0.20, 0.39 )
x=matrix(x,nrow=4,ncol=21,byrow=TRUE)
y=c("1ReinoUnido","2Austrália","3Canadá","4EUnidos","5Japão","6França","7Cingapura",
"8Argentina","9Uruguai","10Cuba","11Colômbia", "12Brasil", "13Paraguai", "14Egito",
"15Nigéria", "16Senegal", "17Serra Leoa","18Angola", "19Etiópia", "20Moçambique", "21China"
)
colnames(x) <- y
x=t(x)
print(x) # mostrar tabela
x=matrix(x,nrow=4,ncol=21,byrow=TRUE)
y=c("ReinoUnido","Austrália","Canadá","EUnidos","Japão","França","Cingapura","Argentina",
"Uruguai","Cuba","Colômbia", "Brasil", "Paraguai", "Egito", "Nigéria", "Senegal",
"Serra Leoa","Angola", "Etiópia", "Moçambique", "China" )
colnames(x) <- y
x=t(x)
matriz <- as.matrix(x)
#Particionar os cluster
#usando a matriz de dissimilaridade
#"pam" minimiza a soma
IndicedeDesenvolvimento <- daisy(x)
x.clus <- pam(IndicedeDesenvolvimento, 4, diss = TRUE)$clustering
#Gráfico dos cluster identificado
#interatividade dos grupos desejados
# número de grupos adotado 4
if(interactive())
clusplot(IndicedeDesenvolvimento, x.clus, lines=5,diss = TRUE,
color=TRUE,col.p="black",labels=3,)
clusplot
```

Para Programa não hierárquico II

```
x = c( 0.88  , 0.90  , 0.90,  0.87 , 0.93 , 0.89 , 0.88 , 0.81 , 0.82 , 0.85, 0.77 ,0.71, 0.75,
0.70, 0.44, 0.47, 0.23, 0.34, 0.31, 0.24, 0.76, 0.99  , 0.99  ,  0.98, 0.98 , 0.93 , 0.97 ,
0.87 , 0.92 , 0.92 , 0.90, 0.85 ,0.83, 0.83, 0.62, 0.58, 0.37, 0.33, 0.36, 0.35, 0.37, 0.80,
0.91  , 0.93  ,  0.94, 0.97 , 0.93 , 0.92 , 0.91 , 0.80 , 0.75 , 0.64, 0.69 ,0.72, 0.63, 0.60,
0.37, 0.45, 0.27, 0.51, 0.32, 0.36, 0.61, 1.10  , 1.26  ,  1.24, 1.18 , 1.20 , 1.24 , 1.41 ,
0.55 , 1.05 , 0.07, -1.36,0.47,-0.87, 0.21,-1.36,-0.68,-1.26,-1.98,-0.55, 0.20, 0.39 )
x=matrix(x,nrow=4,ncol=21,byrow=TRUE)
y=c("ReinoUnido","Austrália","Canadá","EUnidos","Japão","França","Cingapura"
,"Argentina","Uruguai","Cuba","Colômbia", "Brasil", "Paraguai", "Egito", "Nigéria",
"Senegal", "Serra Leoa","Angola", "Etiópia", "Moçambique", "China" )
colnames(x) <- y
x=t(x)
matriz <- as.matrix(x)
matriz
matriz <- as.matrix(x)
ANH1 <- kmeans(matriz,centers = 4, algorithm = c("Hartigan-Wong"))
ANH2 <- kmeans(matriz,centers = 4, algorithm = c("Lloyd"))
ANH3 <- kmeans(matriz,centers = 5, algorithm = c("Forgy"))
ANH4 <- kmeans(matriz,centers = 5, algorithm = c("MacQueen"))

par(mfrow=c(2,2))
plot(matriz, col = ANH1$cluster, main = "Hartigan-Wong")
points(ANH1$centers, col=1:5,pch=2)

plot(matriz, col = ANH2$cluster, main = "Lloyd")
points(ANH2$centers, col=1:4, pch = 8)

plot(matriz, col = ANH3$cluster, main = "Forgy")
points(ANH3$centers, col=1:5, pch = 8)

plot(matriz, col = ANH4$cluster, main = "MacQueen")
points(ANH4$centers, col=1:5, pch = 8)
```

Exercícios

1 - Defina Análise de Agrupamento.

2 - Quais os objetivos da Análise de Agrupamento?

3 - Os resultados da Análise de Agrupamento indicam categoricamente os agrupamentos existentes numa determinada amostra ou população, de modo a permitir ao pesquisador uma conclusão definitiva e única quanto aos mesmos.

Comente e critique essa afirmativa.

4 - Quais são os pressupostos para a aplicação da Análise de Agrupamento?

5 - O que significa padronizar os dados e qual a sua importância dentro da Análise de Agrupamento?

6 - O que deve orientar o pesquisador na escolha do número de agrupamentos na solução final da Análise de Agrupamento, ou seja, quantos grupos devem ser formados?

7 - A Tabela a seguir apresenta o resultado de um questionário aplicado em uma amostra de dez alunos para saber o grau de satisfação destes sobre o curso de administração. Os quesitos avaliados foram conhecimento adquirido, avaliação do corpo docente e mercado de trabalho. As notas são atribuídas em uma escala quantitativa de zero a dez para cada questão.

Tabela 6.1. Grau de Satisfação Deste Sobre o Curso de Administração

Aluno	Conhecimento	Docentes	Mercado
A	9	8	9
B	9	7	6
C	8	7	8
D	2	3	2
E	6	5	4
F	6	7	8
G	4	2	3
H	3	2	4
I	6	4	5
J	3	1	3

Fundamentado nesta base de dados e por meio da aplicação da técnica de Análise de Agrupamento hierárquicos com a distância euclidiana com a utilização do método da ligação simples, identifique o número de agrupamentos e interprete as saídas geradas.

8- A base de dados a seguir apresenta características de 25 empresas varejistas (itens no sortimento, número de lojas com mais de 1.000 m^2 e faturamento mensal em R$). Por meio da elaboração da Análise de Agrupamento, interprete todas as saídas do processamento (hierárquico e k-means), após padronização das variáveis pelo método Z scores, utilizando a distância euclidiana quadrada e o método entre grupos.

Tabela 6.2 - A Base de Dados a Seguir Apresenta Características de 25 Empresas Varejistas (itens no sortimento, número de lojas com mais de 1.000 m^2 e faturamento mensal em R$).

Empresa	Itens no sortimento	Número de lojas com mais de 1.000m^2	Faturamento mensal
A	2500	3	250000
B	2700	3	240000
C	4000	4	310000
D	3700	3	390000
E	9850	5	540000
F	17000	7	740000
G	25000	8	850000
H	3600	3	290000
I	4500	4	350000
J	6900	4	450000
K	12800	8	650000
L	32000	16	980000
M	1000	1	120000
N	1200	2	190000
O	1450	3	190000
P	14500	6	695000
Q	8500	5	490000
R	9800	3	570000
S	72000	18	1250000
T	4500	3	290000
U	35000	10	1000000
V	17000	8	820000
X	7000	4	410000
Y	79000	24	1950000
Z	7000	3	390000

9 - A Tabela 7.3 - apresenta dados relativos a 21 países, de acordo com o banco de dados da ONU (2012) disponível no site: www.undp.org/hdro. As variáveis aqui analisadas são os índices de: expectativas de vida, educação, renda (PIB) e estabilidade política e de segurança. Estes índices foram construídos por uma metodologia própria da ONU e quanto maiores seus valores, melhor é a qualidade do país.

Tabela 6.3: valores dos índices de desenvolvimento de países

Países	Expectativa de vida	Educação	PIB	Estabilidade política
1. Reino Unido	0,88	0,99	0,91	1,10
2. Austrália	0,90	0,99	0,93	1,26
3. Canadá	0,90	0,98	0,94	1,24
4. Estados Unidos	0,87	0,98	0,97	1,18
5. Japão	0,93	0,93	0,93	1,20
6. França	0,89	0,97	0,92	1,04
7. Cingapura	0,88	0,87	0,91	1,41
8. Argentina	0,81	0,92	0,80	0,55
9. Uruguai	0,82	0,92	0,75	1,05
10. Cuba	0,85	0,90	0,64	0,07
11. Colômbia	0,77	0,85	0,69	-1,36
12. Brasil	0,71	0,83	0,72	0,47
13. Paraguai	0,75	0,83	0,63	-0,87
14. Egito	0,70	0,62	0,60	0,21
15. Nigéria	0,44	0,58	0,37	-1,36
16. Senegal	0,47	0,37	0,45	-0,68
17. Serra Leoa	0,23	0,33	0,27	-1,26
18. Angola	0,34	0,36	0,51	-1,98
19. Etiópia	0,31	0,35	0,32	-0,55
20. Moçambique	0,24	0,37	0,36	0,20
21. China	0,76	0,80	0,61	0,39
Média	0,69	0,75	0,68	0,16
Desvio ádrão	0,24	0,25	0,23	1,06

Fonte: ONU, 2012, site.undp.org/hdro. Relatório de Desenvolvimento Humano.

Calcular o método de ligação simples, completa, média, de Ward e comparar os mesmos.

10 - Para fins de ilustração, apresentamos os resultados de um estudo das percepções e preferências de estudantes sobre 10 diferentes marcas de cerveja. Cada um dos 32 estudantes (um subconjunto escolhido aleatoriamente e sistemático de um estudo maior) classificou sua preferência em escala de 10 pontos para cada uma das seguintes marcas: Skol, Brahma, Kaise, Schin, Boehmia, Devassa, Cristal, Heineken, Stella e Bavária. A tabela abaixo mostra os resultados.

Tabela 6.4 . Preferência Expressas por 32 Estudantes por 10 Marcas Diferentes de Cerveja (medidas em uma escala de 9 pontos).

	Skol	Brahma	Kaiser	Schin	Boehmia	Devassa	Cristal	Heineken	Stella	Bavária
A001	5	9	7	1	7	6	6	5	9	5
A008	7	5	6	8	8	4	8	8	7	7
A015	7	7	5	6	6	1	8	4	7	5
A022	7	7	5	2	5	8	4	6	8	9
A029	9	7	3	1	6	8	2	7	6	8
A036	7	6	4	3	7	6	6	5	4	9
A043	5	5	5	6	6	4	7	5	5	6
A050	5	3	1	5	5	5	3	5	5	9
A057	9	3	2	6	4	6	1	5	3	6
A064	2	6	6	5	6	4	8	4	4	3
A071	7	7	7	5	7	8	6	7	7	8
A078	8	3	3	9	9	2	1	9	7	8
A085	6	5	3	7	6	5	8	6	7	5
A092	5	6	3	8	6	7	6	7	6	7
A099	4	7	2	8	5	9	8	3	8	8
A106	3	3	4	5	6	5	9	7	5	5
A113	2	4	5	7	6	6	8	1	7	4
A120	9	3	7	4	2	4	6	3	8	6
A127	5	3	4	7	7	7	6	6	6	6
A134	2	4	4	8	5	5	5	4	6	6
A141	5	7	6	7	5	8	8	7	5	7
A148	8	9	6	7	7	8	6	8	8	8
A162	5	6	6	7	5	3	7	3	4	3
A169	5	5	6	7	5	4	6	3	7	6
A176	5	5	7	8	7	6	7	5	4	7
A183	3	5	4	7	5	1	2	6	6	5
A190	4	3	6	8	7	1	8	2	7	7
A197	3	8	4	8	7	2	8	4	6	1

A204	3	5	1	5	3	3	4	6	7	5
A211	3	8	5	8	6	5	5	3	7	8
A218	8	8	5	7	6	9	7	7	6	8
A225	7	6	2	2	5	6	2	7	5	5

Usando Análise de Agrupamento para reduzir essa heterogeneidade, dividindo a amostra em subgrupos menores e mais homogêneos. Calcular a técnica hierárquica e não hierárquica e comparar.

REFERÊNCIAS

ALBUQUERQUE, M. A. Análise de Agrupamento hierárquica e incremental - estudo de caso em ciências florestais. Tese de Doutorado. UFRPE, 2013.

ANDERSON, J. J. B. Numeric Examination of Multivariate Soil Samples. Math. Geol., 3, 1971.

BUSSAB, W de O; MIAZAKI, E. S; ANDRADE, D. Introdução à análise de agrupamentos. 9º Simpósio Nacional de Probabilidade e Estatística. São Paulo. Associação Brasileira de Estatística, 1990.

CORMACK, R. M. A Review of Classifications. JRSS, A, 134-321-367, 1971.

DURAN, B.S.; ODELL, P. L. Cluster Analysis: A Survey. Springer-Verlog, Berlin, 1974.

EVERITT, B. Cluster Analysis. Heinemann Educational Books, London, 1974.

FERNANDO, Frei. Introdução à análise de agrupamento: teoria e prática, São Paulo: Editora UNESP, 2006.

GOWER, J. C. Some Distances Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika, 53, 325-338, 1966.

GOWER, J. C. Classification and Geology. Rev. I.S.I., 38, 35-41, 1970.

GUTTMAN, L. A General Nonmetric Technic for Finding the Smallest Coordinate Space for a Configuration of Points. Psychometrika, 33, 469-506, 1968.

HARTIGAN, J. A. Representation of Similarity Matrices by Trees. JASA, 62, 1140-1158, 1967.

HARTIGAN, J. A. Clustering Algorithms. John Wiley & Sons, New York, 1975.

HO, L. L. Alguns Aspectos da Técnica AID para Variável Resposta Categorizada.

Dissertação de Mestrado, IME/USP, 1987.

JARDINE, R. C. The Structure and Construction of Taxionomic Hierarchies. Math. Biosci, 1, 173-179, 1967.

JOHNSON, R. A.; WICHERN, D. W. Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey, 1982.

MANTEL, N. (1967). The detection of disease clustering and generalized regression approach. Cancer Research, 27, 209-220, 1967.

MANZATO, A. J. Análise Hierárquica de Agrupamentos para Distribuições Multinomiais. Dissertação de Mestrado, IME/USP, 1983.

MIAZAKI, E. S. Mistura de Multinomiais como Técnica de Análise de Conglomerados. Dissertação de Mestrado, IME/USP, 1979.

PARADIS, E; CLAUDE, L. STRIMMER, K. APE: Analyses of Phylogenetics and Evolution in R Language Bioinformatics, 20: 289-290, 2004.

ROMESBURG, H. C. Cluster Analysis for Researchers. Lifetime Learning Publications, California, 1984.

SAMMON, J. W. A Nonlinear Mapping for Data Structure Analysis. IEEE Trans. Computer, C18, 401-409, 1969.

SILVA, Anderson Rodrigo da. Métodos de Analise Multivariada em R, 1.ed, FEALQ, 2016.

SILVA, A. R.; DIAS, C. T. S. A cophenetic correlation coefficient for Tocher's method. Pesquisa Agropecuária Brasileira, 48, 589-596, 2013.

SHEPARD, R. N. The Analysis of Proximities Multidimensional Scaling with na Unknown Distance Function. Psychometrika, 27, 125-139, 1962.

SOKAL, R. R. & Rohlf, F. J. The Comparison of Dendrograms by Objectives Methods. Taxon., 11, 33-40, 1962.

SPÄTH, H. Cluster Analysis Algorithms. Ellis Horwood Limited, England, 1980

WOLFE, J. H. A Monte Carlo Study of the Sampling Distribution of the LR for Mixture of Multinormal Distribution Technical Bulletin, Naval Personnal and Trainning Research Laboratory, 1971.

SOBRE O LIVRO

PROJETO GRÁFICO E EDITORAÇÃO	Arão de Azevêdo Souza
MANCHA GRÁFICA	16 x 24cm
TIPOLOGIA UTILIZADA	Faustina 12 pt

Este livro se originou da necessidade de texto básico sobre análise de agrupamento e suas aplicações computacionais por parte de estudantes de graduação e pos-graduação, pesquisadores e profissionais das mais diversas áreas. E também vai ao encontro da considerável expansão e utilização do softwore R. Uma vez que o recurso a software apropriado é atualmente indispensável no tratamento de dados, é feito uma intensa utilização da linguagem R.

eduepb