Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica

ME-714 Análise de Dados Discretos

Aplicação de clusterização monotética em dados discretos de filmes brasileiros da década de 2000.

Autores:
Gustavo de Almeida Leiva **RA:** 217418
Luana Paias Martínez **RA:** 220559
Marcelo da Silva Lourenço **RA:** 221418
Renan Cyrillo Lima **RA:** 223978

Professor: Hildete Pinheiro

Campinas, julho 2021

1 Introdução

Atualmente, é comum encontrar bancos de dados com muitas informações e buscar identificar características similares e padrões entre os dados afim de obter informações. Existem diversos métodos de análise para realizar isso, sendo um deles a clusterização que, por sua vez, tem como intuito identificar grupos de dados semelhantes no conjunto consolidado. Este presente trabalho utiliza de um método de clusterização divisível monotética para buscar agrupar os filmes brasileiros, produzidos durante a década de 2000, de acordo com seus gêneros.

2 Objetivo

O objetivo do trabalho é construir um algoritmo hierárquico divisível que categoriza grupos de dados discretos a partir da maximização da associação entre as variáveis. Nesse caso, o objetivo principal é clusterizar filmes brasileiros dos anos 2000 a 2010 de acordo com seus gêneros.

3 Metodologia

Algoritmos de clusterização ou agrupamento de dados, são técnicas que vem se popularizando por conta de sua automatização e a necessidade de manipulação de grandes volumes de dados.

De forma simplificada, os algoritmos de agrupamento seguem duas principais vertentes: divisa e aglomerativa. Seu objetivo principal é o de agrupar observações de acordo com algum parâmetro de similaridade e diferem-se quanto ao parâmetro inicial, método de divisão e se há admissão de particionamento recursivo e hierárquico. Este trabalho será limitado apenas a elaborar os cálculos da clusterização monotética (ou Mona, em inglês).

3.1 Clusterização divisiva

O algoritmo monotético utilizado particiona a informação de dados e utiliza de recursão para assim realizar o agrupamento de forma a transformar as características de interesse em regras de decisão binárias.

3.2 MONA

Inicialmente, com as informações dispostas em forma matricial, o primeiro passo é definir a quantidade de clusteres **K** [Tran Brian McGuire]. Valores de **K** elevados podem resultar em clusteres com alto grau de similaridade entre si, já valores pequenos de **K** pode resultar em agrupamentos com variabilidade de certa característica muito discrepante [Milligan e Cooper 1985]. Portanto, é importante que se escolha um valor justo para o conjunto de dados em questão (que não resulte em variabilidade nem similaridade).

Seja y_{iq} a i-ésima observação da variável q (q = 1,...,Q, o nº de variáveis resposta no banco de dados). Queremos definir a quantidade de clusteres mutualmente exclusivos $\mathbf{C}_1, ..., \mathbf{C}_K$ de tal forma que para cada $\mathbf{C}_i$, existam elementos similares entre si e suficientemente diferentes para todo i e j (com $i \neq j$). A partir disso, o algoritmo buscará por divisões binárias em cada variável resposta de tal forma que seja também a melhor divisão das respostas multivariadas em termos do critério de inércia [Olshen]. A inércia (I), interpretada como uma medida de variabilidade euclidiana intra-cluster, pode ser obtida ao se calcular o quadrado da distância euclidiana (d^2_{euc}) entre as observações (em formato matricial), da seguinte forma:

$$I(C_k) = \sum_{i \epsilon C_k} d^2_{euc}(y_i, \bar{y}_{C_k}),$$

Em que $\bar{y}_{C_k}$ é a média de todas as observações no cluster C_k.

Portanto, uma operação de divisão binária sobre um cluster C_k, denominada como $s(C_k)$, resultará em dois subconjuntos menores C_{kL} (apelidado de cluster esquerdo) e C_{kR} (cluster direito). A regra de decisão recai sobre a maximização do decréscimo da inércia da seguinte diferença:

$$\lambda(s, C_k) = I(C_k) - I(C_{kL}) - I(C_{kR})$$

E a maximização de s^* fica como se segue:

$$s^*(C_k = argmax\lambda(s, C_k)$$

O algoritmo é então reaplicado iterativamente até atingir algum critério de parada pré-definido (em nosso caso, o número K de clusteres).

Com isso definido, é necessário, antes de iniciar a clusterização, algum método de estimação para o número ótimo de clusteres. O método escolhido foi a validação cruzada M-fold ([Tran Brian McGuire]) que particiona aleatoriamente os dados em **M** grupos de igual tamanho. Então **M** - 1 grupos são utilizados como conjunto de treinamento

para assim criar uma árvore de regressão [Olshen] e o grupo restante é utilizado para predição e é avaliado a qualidade da assertividade. O processo é repetido **M** vezes e é calculado a Diferença Quadrática Média (MSE_m):

$$MSE_m = \frac{1}{n_m} \sum_{q=1}^{Q} \sum_{i \in m} d^2_{euc}(y_{iq}, \hat{y}_{(-i)q})$$

Com $\hat{y}_{(-i)q}$ sendo a média do cluster (criado pelo treinamento) sobre a variável q e y_{iq} o valor observado. A partir de **M** repetições feitas, sua média é a estimativa do Erro Quadrático Médio baseado em sua validação cruzada predizendo uma nova obervação, conforme segue:

$$CV_K = \overline{MSE} = \frac{1}{M} \sum_{m=1}^{M} MSE_m$$

O propósito de toda essa operação é de encontrar qual o 'melhor' erro de previsão para novas observações. Ao final, a literatura sugere que seja escolhido o menor CV_k. No entanto, essa solução pode favorecer a escolha de um número muito elevado de clusteres devido a queda retilínea do valor dos erros. Para contornar essa situação, utiliza-se um desvio-padrão abaixo do valor mínimo de erro estimado (com 95% de significância):

$$SE(\overline{MSE}) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (MSE_m - \overline{MSE})}.$$

Uma vez definido o número de clusters do algoritmo, pode-se avançar para a análise de dados. Sabe-se que diversos pesquisadores tem interesse em analisar variáveis binárias, principalmente os da área de sociais. Essas variáveis possuem apenas 2 valores, 0 ou 1, dependendo da presença ou ausência de certo atributo. Existem diversos algoritmos para clusterizar essas variáveis, sendo um deles o MONA.

MONA ou Monothetic Analysis é um algoritmo de clusterização divisível (como explicado anteriormente) para variáveis binárias (ou categóricas). Resumidamente, a ideia do algoritmo é selecionar uma variável (no nosso caso, gênero do filme) e dividir o conjunto de objetos (filmes) em objetos que contém ou não contém um atributo correspondente. Em cada um desses subconjuntos, uma das variáveis remanescentes é selecionada e o mesmo método é aplicado para dividir esse subconjunto em dois grupos menores. O processo é repetido até que o subconjunto contenha apenas um único objeto ou até que as variáveis restantes não puderem mais ser separadas em subconjuntos. Essa última situação só acontece quando cada uma das variáveis permanece constante para todos objetos do subconjunto. Por exemplo, observando a Tabela 1, nota-se que os objetos CCC e DDD não podem ser separados.

Tabela 1: Exemplo de conjunto de dados com variáveis binárias

	Variáveis					
Objetos	1	2	3	4	5	6
A A A	1	1	0	1	1	0
B B B	1	1	0	0	0	1
C C C	1	1	1	1	1	0
D D D	1	1	1	1	1	0
E E E	0	0	0	1	0	1
F F F	0	0	0	0	0	0
G G G	0	0	1	1	0	1
H H H	0	0	1	1	1	0

Por conta do fato de que o conjunto de objetos é dividido em subconjuntos e esse processo é repetido dentro de cada subconjunto, o algoritmo é chamado de hierárquico. Mais especificamente, é divisivo. Além disso, devido a separação ser realizada utilizando apenas uma variável, ela é chamada monotética.

A seleção de variáveis para separação em subconjunto acontece da seguinte maneira: seleciona-se a variável para a qual a soma das "similaridades"com todas as outras variáveis é a maior possível, ou seja, a variável é que é mais centrada localmente. Essa similaridade entre duas variáveis, no Mona, é definida como o número de objetos para cada combinação que as duas variáveis podem ter. O produto entre o número de objetos que as duas variáveis recebem o valor 0 e o número que eles recebem o valor 1 é calculado. A partir daí o número de objetos que a primeira variável é 0 e a segunda é 1 é multiplicado pelo número de objetos que as duas variáveis recebem os valores opostos. A medida de associação (ou similaridade) é então resultado do valor absoluto da diferença desses dois produtos citados. Considerando os dados da Tabela 1, por exemplo, as variáveis 1 e 2 são idênticas e portanto mostram um valor de associação elevado. Os dois produtos, nesse caso, são 16 (= 4 x 4) e 0 (= 0 x 0). Portanto, a medida de associação resulta em |16 - 0| = 16. Por outro lado, as variáveis 1 e 3 são muito diferentes entre si, e seu valor de associação |(2 x 2) - (2 x 2)| = 0 confirma essa afirmação. A medida de associação entre duas variáveis pode ser calculada também através de tabelas de contingência, conforme pode ser visto nas Tabelas 2 (tabela de contingência genérica), 3 (tabela de contingência para as variáveis 1 e 2) e 4 (tabela de contingência para as variáveis 1 e 3).

Tabela 2: Tabela de Contingência Genérica

	1	0
1	a	b
0	c	d

Tabela 3: Tabela de Contingência para as variáveis 1 e 2

	1	0
1	4	0
0	0	4

Tabela 4: Tabela de Contingência para as variáveis 1 e 3

	1	0
1	2	2
0	2	2

Essa similaridade definida nos últimos parágrafos expressam o quanto duas variáveis provém divisões similares do conjunto de objetos. No exemplo, as variáveis 1 e 3 demonstram informações completamente diferentes e por isso não são nem um pouco similares (o resultado da medida de de associação foi 0). Entretanto, se duas variáveis possuem valores diferentes para todos os objetos da base de dados, elas demonstram informações idênticas e a medida de associação acaba resultando em um valor absoluto grande (como, por exemplo, na Tabela 3). É válido ressaltar que medidas de associação próximas se assemelham a estatística qui-quadrado para tabelas dois a dois.

Como o intuito do método é procurar pela variável que é mais similar para todas as outras variáveis, maximiza-se a soma das associação de todas as outras variáveis. Um exemplo dos resultados do Mona pode ser visto na Figura 1. Observa-se que, no começo do algoritmo (parte esquerda da figura), todos os objetos pertencem a um cluster único. Durante o primeiro passo, a variável 5 é usada para separar esse cluster em dois subconjunto. Os passos de separação se repetem até que o cluster consista de apenas um objeto ou até que as variáveis remanescente não possam ser mais separadas (o que ocorre para o cluster CCC, DDD).

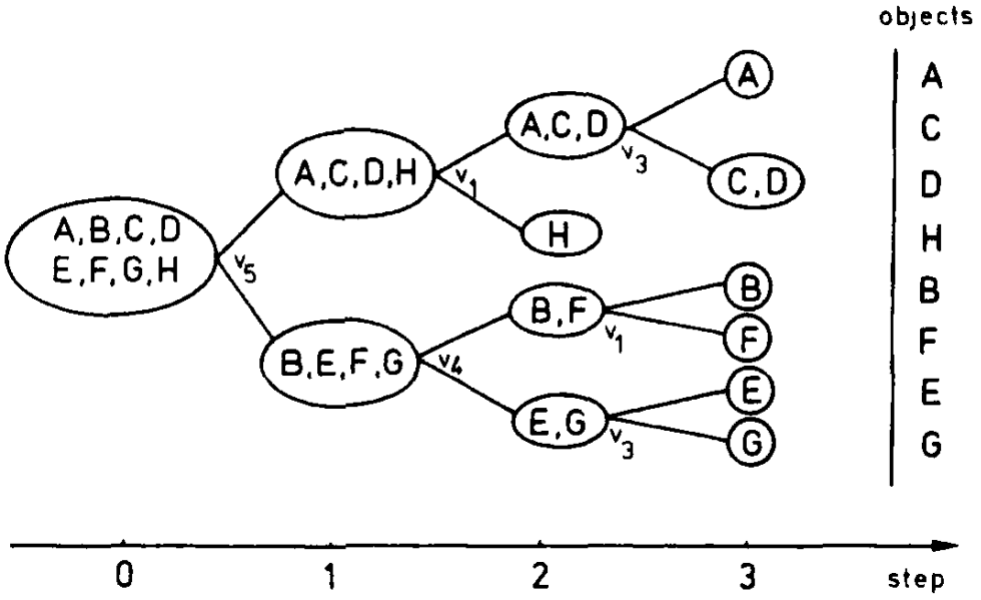

Figura 1: Exemplo de resultados da aplicação do algoritmo Mona, retirado de [Kaufman e Rousseeuw 2005]

4 Aplicação

4.1 Base de dados

Como dito na Introdução, nossa base de dados foi extraída do IMDB, um site que agrega filmes do mundo inteiro e tem diversas características sobre cada um. Os filtros feitos foram:

- País: Brasil;
- Número de avaliações: Mais que 100;
- Ano de estreia: $\geq$ 2000 e $\leq$ 2010.

Também foram feitas manipulações nos dados para que cada coluna seja um gênero de filme diferente e seus respectivos valores sejam 1 ou 0, indicando se, para aquele determinado filme, determinado gênero é presente.Após a realização dos filtros, restaram 297 observações (filmes) no banco de dados.

4.2 Análise Descritiva

Primeiramente, é válido notar que o número máximo de gêneros em um único filme é 3. A Tabela 5 mostra a quantidade de vezes que cada gênero aparece nos filmes do

banco. Nota-se que alguns gêneros, como Drama, Comédia, Romance e Crime possuem frequência maior em comparação aos outros, revelando um possível desbalanceamento na base.

Tabela 5: Quantidade de filmes que possuem cada gênero

Gênero	Qtd. de Filmes	Gênero	Qtd. de Filmes
Drama	198	Animation	9
Comedy	97	Horror	7
Romance	43	Mystery	7
Crime	38	History	6
Adventure	21	Music	6
Family	20	Musical	6
Thriller	19	Sci-Fi	5
Action	17	War	3
Biography	17	Sport	2
Fantasy	12	Western	2

4.3 Clusterização

Para que fosse escolhido o número de clusteres ideal, utilizamos a metodologia descrita em [Chavent 1998]. O gráfico que mostra o comportamento do erro quadrático médio conforme aumenta-se a quantidade de clusteres pode ser observado em Figura 2. De acordo com essa análise, a quantidade de clusteres mais adequada são exatamente 10, o primeiro valor que está dentro do limite dos intervalos de confiança dos erros padrões.

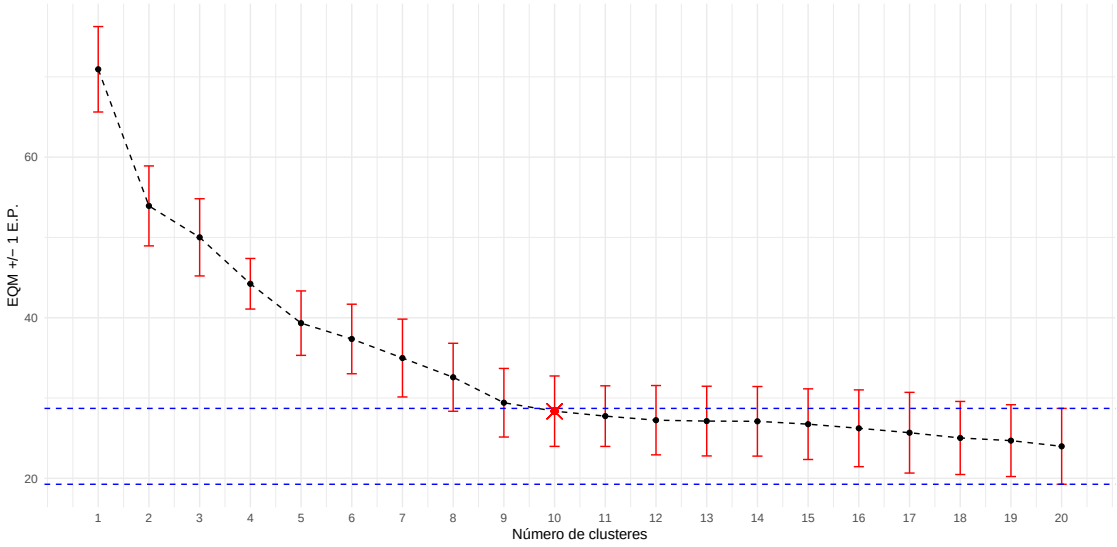

Figura 2: EQM para validação cruzada da clusterização monotética

A Figura 3 mostra a árvore de partição da nossa base de dados para 10 clusteres. Podemos ver que a primeira partição foi entre filmes que continham ($>= 0.5$) ou não continham (< 0.5) drama dentro de sua lista de gêneros. A próxima partição em ambos os lados foi dividindo os filmes que tinham ou não comédia dentro de sua lista. Para a esquerda, vimos que os filmes que NÃO continha drama e NÃO continham comédia foram levados para uma partição que separava entre ter ou não o gênero família. Voltando um partição atrás, vemos que os filmes que NÃO continham drama e continham comédia foram levados para uma partição que divide entre ter ou não romance. Olhando para o lado direito da árvore, vemos a divisão entre ter drama e ter ou não comédia. Os que tem comédia, acabam em um cluster ali mesmo com 32 membros. Os que não tem comédia, são levados para várias outras divisões que, claro, sempre começam tendo drama e NÃO tendo comédia. Depois, passam por vários filtros dizendo se o filme contém crime, biografia, romance ou suspense. É interessante notar que não são todos os gêneros de filmes que são analisados para essa clusterização, apenas alguns dos gêneros mais comuns como visto em 5.

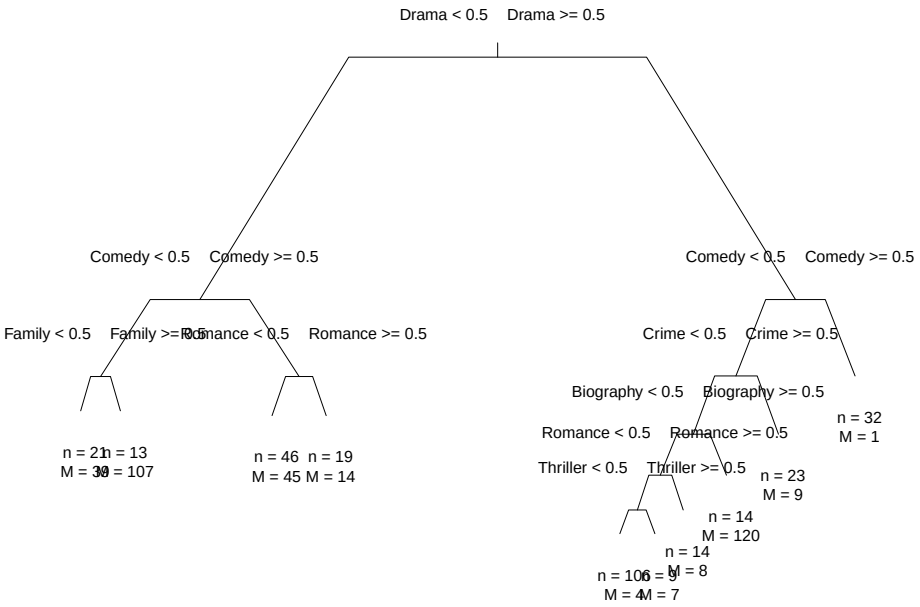

Figura 3: Árvore de partição binária com 9 partições e 10 clusteres

A Figura 4 mostra o comportamento dos gêneros que não aparecem na Figura 3 separados por cluster. Lembrando que o número do cluster é referente a ordem que ele aparece da esquerda para a direita na Figura 3. Portanto, além de cada gênero visto na Figura 4, esses clusteres contém os gêneros de acordo com a árvore anterior. O Cluster 1 possui vários filmes que contém ação e horror, também mistério e ficção científica. O Cluster 2 possui um incidência considerável de filmes de aventura e fantasia. Já o Cluster 3 possui aventura e animação. O Cluster 8 possui a combinação de filmes com história e música, portanto devem ter algumas "biopics"(filmes biográficos) de músicos ou bandas. O Cluster 9 possui praticamente só ação (além de crime e drama, como visto na Figura 3).

Apesar de parecer ter alguns padrões nesse gráfico, é importante dizer que só o que foi considerado na montagem dos clusteres foram os gêneros presentes na Figura 3, portanto as "coincidências"que vemos aqui originaram de gêneros de filmes que são comumente associados com os gêneros principais vistos no primeiro gráfico.

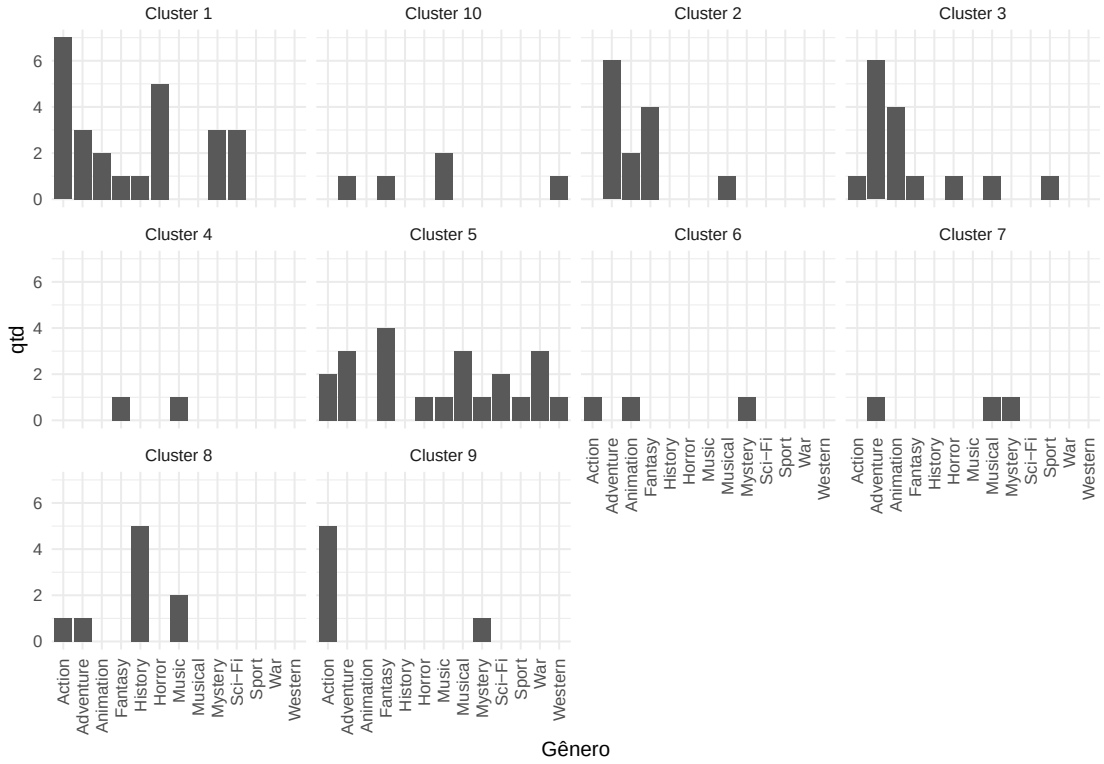

Figura 4: Quantidade de vezes que cada gênero faltante na árvore aparece por cluster

4.4 Conclusões

A metodologia que utilizamos é bastante interessante mas aparenta funcionar melhor com menos variáveis, já que ela consiste em encontrar padrões por meio de uma árvore de partições. Desta forma, no nosso estudo de filmes, vimos que apenas alguns gêneros foram considerados para montagem dos clusteres. Mesmo assim, foi possível observar padrões interessantes dentre os grupos. Clicando aqui, podemos ver quais filmes estão em cada cluster. A base de dados completa utilizada (com a divisão de clusteres) pode ser encontrada aqui.

Referências

CHAVENT, M. A monothetic clustering method. *Pattern Recognition Letters*, v. 19, n. 11, p. 989–996, set. 1998. ISSN 0167-8655.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis: 603*. Illustrated edição. [S.l.]: Wiley-Interscience, 2005. ISBN 978-0-471-73578-6.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, v. 50, n. 2, p. 159–179, jun. 1985. ISSN 1860-0980. Disponível em: <https://doi.org/10.1007/BF02294245>.

OLSHEN, B. F. S. *Classification and Regression Trees*.

TRAN BRIAN MCGUIRE, M. G. T. *Monothetic Clustering*. Disponível em: <https://cran.r-project.org/web/packages/monoClust/vignettes/monoclust.html>.