3
Métodos de Agrupamento de Dados

Esse é o item de maior destaque no processo de agrupamento de dados, pois ele é o responsável pelo agrupamento propriamente dito.

Conforme visto anteriormente, os métodos de agrupamento de dados podem ser divididos em duas grandes categorias, cada uma delas compreendendo diferentes tipos de algoritmos:

- *Métodos Hierárquicos*
  - Algoritmos Aglomerativos
  - Algoritmos Divisivos
- *Métodos Particionais*
  - Algoritmos Exclusivos
  - Algoritmos Não-exclusivos

A seguir são descritas em detalhes as características de cada um desses conjuntos de métodos, bem como os métodos mais conhecidos ou que tenham alguma relevância no processo de agrupamento.

3.1.
Métodos Hierárquicos

Os métodos hierárquicos são técnicas simples onde os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos (Everitt, 2001). Essa representação facilita a visualização sobre a formação dos agrupamentos em cada estágio onde ela ocorreu e com que grau de semelhança entre eles.

Os métodos hierárquicos não requerem que seja definido um número *a priori* de agrupamentos. Através da análise do dendograma (diagrama que mostra a hierarquia e a relação dos agrupamentos em uma estrutura) pode-se inferir no

número de agrupamentos adequados. O gráfico do dendograma é abordado com mais detalhes nas seções 3.1.1.1 e 3.1.2.1.

Os métodos hierárquicos requerem uma matriz contendo as métricas de distância entre os agrupamentos em cada estágio do algoritmo. Essa matriz é conhecida como matriz de similaridades entre agrupamentos. Dessa forma, imaginando um estágio do algoritmo onde o número de agrupamentos corrente é três (G1, G2, G3), pode-se supor a seguinte matriz de similaridades entre os agrupamentos:

Tabela 2: Tabela ilustrativa da Matriz de Similaridades entre Grupos.

	G1	**G2**	**G3**
G1	0	0,1	0,3
G2	0,1	0	0,4
G3	0,3	0,4	0

Pela tabela ilustrativa acima se pode observar que *G1* e *G2* são os agrupamentos mais similares, enquanto que *G2* e *G3* são os menos similares.

São utilizados os métodos de distância entre grupos para o cálculo dos valores de proximidade entre os agrupamentos. Os métodos de distância entre grupos são descritos na seção 3.1.3.

Os métodos hierárquicos são subdivididos em Métodos Aglomerativos e Métodos Divisivos.

3.1.1. Métodos Aglomerativos

Os métodos aglomerativos são os mais comuns entre os métodos hierárquicos.

Nesse tipo de método inicia-se com cada padrão formando seu próprio agrupamento e gradualmente os grupos são unidos até que um único agrupamento contendo todos os dados seja gerado (Silva, 2005). Logo no início do processo, os agrupamentos são pequenos e os elementos de cada grupo possuem um alto grau de similaridade. Ao final do processo, têm-se poucos agrupamentos, cada um podendo conter muitos elementos e menos similares entre si.

O primeiro passo é criar uma matriz de similaridades entre os agrupamentos, lembrando que, no início do algoritmo, cada padrão é um

agrupamento. O grande problema dos métodos hierárquicos reside nessa matriz de similaridades (Viana, 2004). Considerando-se N dados, tem-se, no início do algoritmo aglomerativo, uma combinação de N por 2. Para uma base de dados contendo 1000 elementos isto resulta em cerca de 500 mil combinações, o que significa que no primeiro momento o algoritmo calculará 500 mil medidas de similaridades. Para uma base de dados contendo 2000 elementos, esse número sobe para 2 milhões!

Depois de criada a matriz de similaridades, o próximo passo é encontrar o menor valor na matriz de similaridades. Esse valor identifica os dois agrupamentos mais similares entre si. Feito isso, esses dois agrupamentos identificados são agrupados, formando assim um novo agrupamento. Logo em seguida, a matriz de similaridades é atualizada, contendo agora um agrupamento a menos. Esse procedimento é feito até restar apenas um único agrupamento.

O procedimento geral pode ser descrito em poucos passos (Matteucci, 2005):

1. Início: Cada agrupamento contém um único padrão.
2. Calcular a matriz de similaridades.
3. Forma-se um novo agrupamento pela união dos agrupamentos com maior grau de similaridade.
4. Os passos 2 e 3 são executados (N-1) vezes, até que todos os objetos estejam em um único agrupamento.

As principais desvantagens dos métodos hierárquicos aglomerativos são (Yuras, 2004):

- Os agrupamentos não podem ser corrigidos, ou seja, os padrões de um determinado agrupamento permanecerão nesse agrupamento até o final da execução do algoritmo;
- Requerem muito espaço de memória e tempo de processamento devido às matrizes de similaridade.

3.1.1.1. Dendograma

É a representação gráfica em forma de árvore sobre a estrutura dos agrupamentos. Nos métodos hierárquicos aglomerativos, o dendograma representa a ordem em que os dados foram agrupados, como mostra a Figura 4.

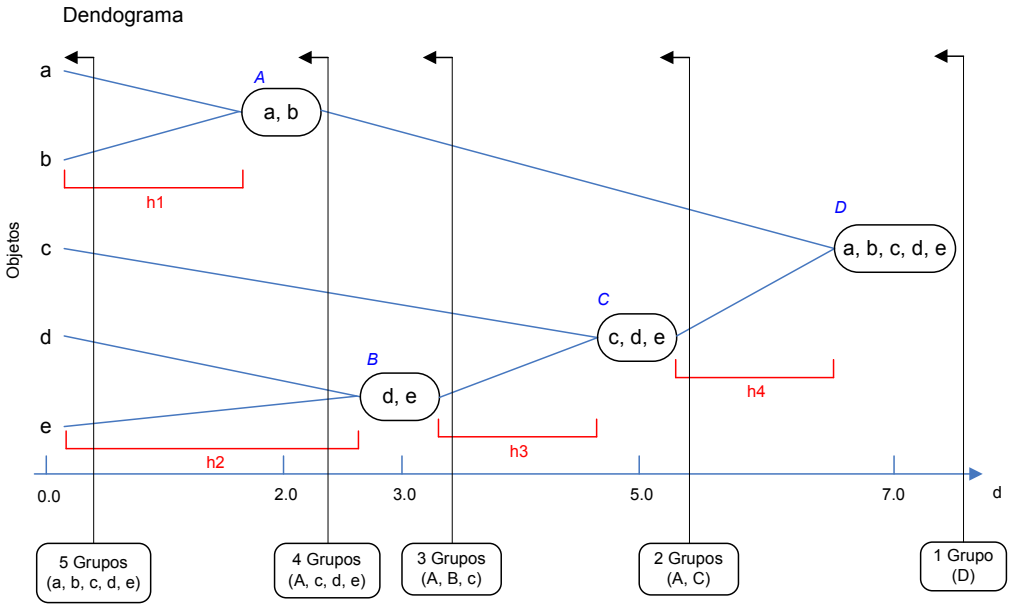

Figura 4: Método Hierárquico Aglomerativo – Dendograma.

Analisando o gráfico acima existem inicialmente cinco agrupamentos (*a, b, c, d, e*). Em seguida os agrupamentos *a* e *b* são agrupados, gerando o agrupamento *A*. A medida de similaridade entre *a* e *b* é definido pela altura *h1*. O mesmo ocorre com os agrupamentos *d* e *e*; eles são agrupados formando o agrupamento *B*. A medida de similaridade entre *d* e *e* é definida pela altura *h2*. Pode-se observar que, nesse momento, existem três agrupamentos (*A*, *B* e *c*). No passo seguinte, os agrupamentos *c* e *B* são agrupados formando o agrupamento *C*. A medida de similaridade entre *c* e *B* é definida pela altura *h3*. Pode-se observar que agora existem apenas os agrupamentos *A* e *C*. No último passo os agrupamentos *A* e *C* são agrupados, formando um único agrupamento *D*. A medida de similaridade entre *A* e *C* é definido pela altura *h4*.

3.1.1.2. Coeficiente Aglomerativo (CA)

Mede a qualidade de um agrupamento aglomerativo (Matlab® Reference, 2005). Para cada objeto *i*, *d(i)* é sua dissimilaridade em relação ao primeiro agrupamento em que foi inserido dividido pela dissimilaridade na etapa final do algoritmo. O coeficiente é então definido como:

$$CA = \frac{1}{n}\sum_{i}^{n} 1 - d(i) \qquad (3.1)$$

onde *n* é o número total de objetos do conjunto de dados.

Os valores do coeficiente variam entre 0 e 1. Valores baixos do coeficiente correspondem a estruturas ruins, onde nenhum agrupamento foi encontrado. Por outro lado, valores próximos a 1 representam que estruturas muito claras foram identificadas.

3.1.1.3. Banner de Dissimilaridade

A hierarquia dos agrupamentos pode ser representada graficamente pelo Banner de Dissimilaridade (Kauffman, 1990). Essa representação mostra as sucessivas uniões entre agrupamentos. A leitura do banner é feita da esquerda para a direita. Os objetos são listados verticalmente. A união de dois agrupamentos é representada por uma barra horizontal que começa na região de dissimilaridades dos agrupamentos envolvidos. O coeficiente aglomerativo (CA) pode ser visto como a largura média do banner. A Figura 5 mostra um exemplo de banner de dissimilaridade para métodos hierárquicos aglomerativos.

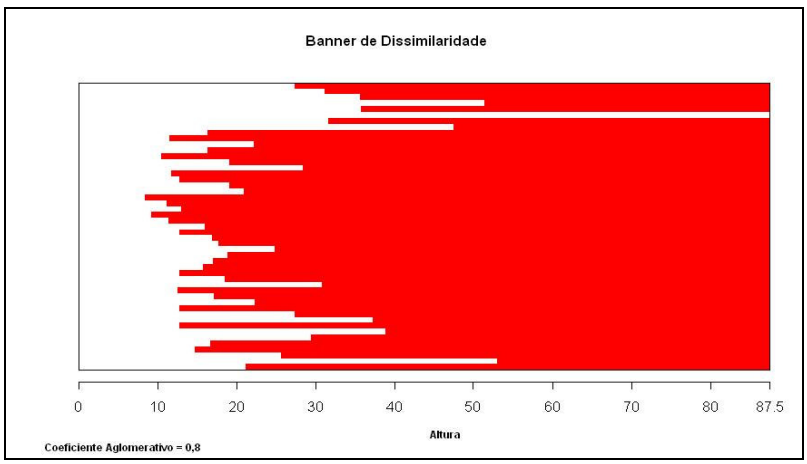

Figura 5: Banner de Dissimilaridade (Métodos Hierárquicos Aglomerativos).

A Figura 6 mostra um exemplo onde se pode ver a relação entre o dendograma e o banner.

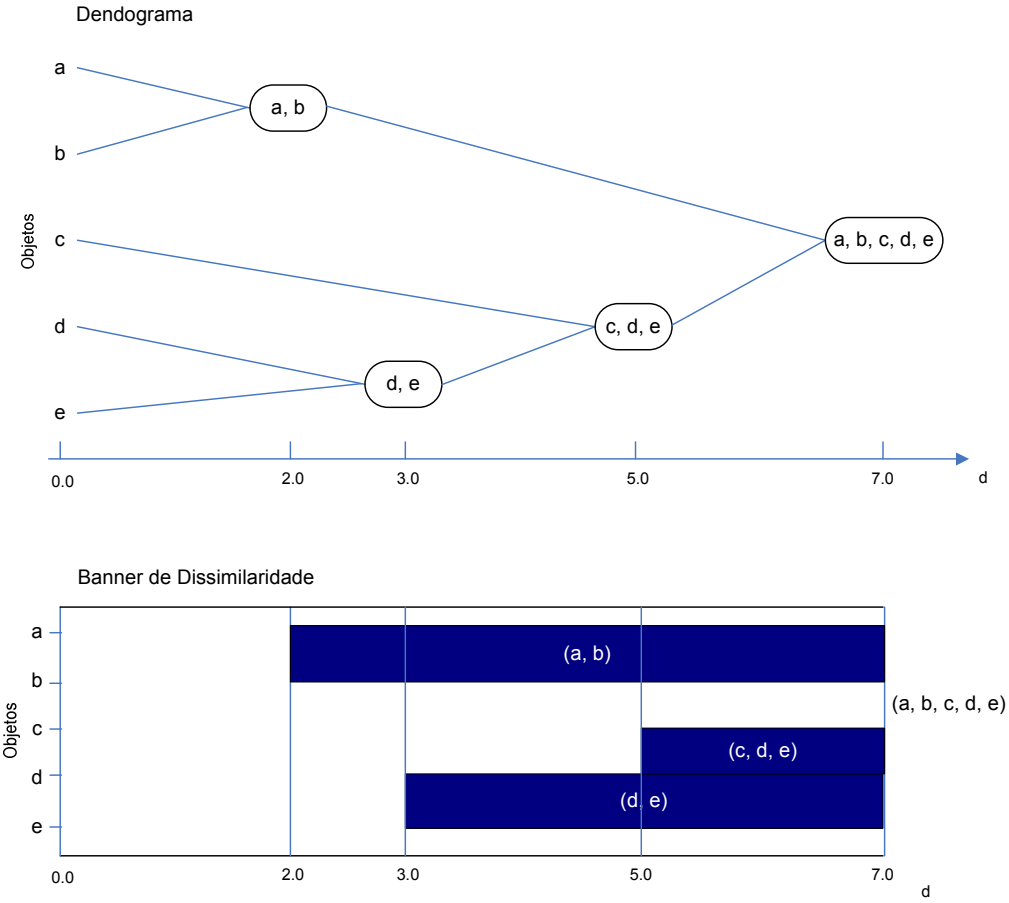

Figura 6: Métodos Hierárquicos Aglomerativos: dendograma e banner.

Observando o dendograma pode-se notar que os objetos *a* e *b* foram agrupados em um novo agrupamento e a medida de similaridade entre eles é 2. Observando agora o banner, na medida de similaridade 2 ocorre o agrupamento entre os objetos *a* e *b*. Essa mesma análise pode ser feita para os demais agrupamentos.

3.1.2.
Métodos Divisivos

Os métodos divisivos são os menos comuns entre os métodos hierárquicos devido a sua ineficiência e exigem uma capacidade computacional maior que os métodos hierárquicos aglomerativos (Costa, 1999).

Esse método começa com um único agrupamento formado por todos os padrões e gradualmente vai dividindo os agrupamentos em agrupamentos menores até que termine com um agrupamento por padrão. A idéia é achar a partição que minimiza a matriz de similaridades.

O procedimento geral pode ser descrito em poucos passos:

1. Início: Um único agrupamento contendo todos os padrões.
2. Calcula-se a matriz de similaridades entre todos os possíveis pares de agrupamentos.
3. Forma-se um novo agrupamento pela divisão dos pares de agrupamentos com menor grau de similaridade.
4. Os passos 2 e 3 são executados até que se tenha um agrupamento por padrão.

O primeiro passo do algoritmo envolve todas as divisões possíveis dos dados em dois agrupamentos, o que o tornaria impraticável para um número grande de elementos, envolvendo, dessa forma, um grande número de combinações (Everitt, 2001).

Os métodos divisivos possuem a vantagem de considerar muitas divisões no primeiro passo, diminuindo a probabilidade de uma decisão errada, sendo assim, mais seguros que os métodos hierárquicos aglomerativos (WinIDAMS, 2005a).

3.1.2.1.
Dendograma

Nos métodos hierárquicos divisivos, o dendograma representa a ordem em que os agrupamentos foram divididos, como mostra a Figura 7.

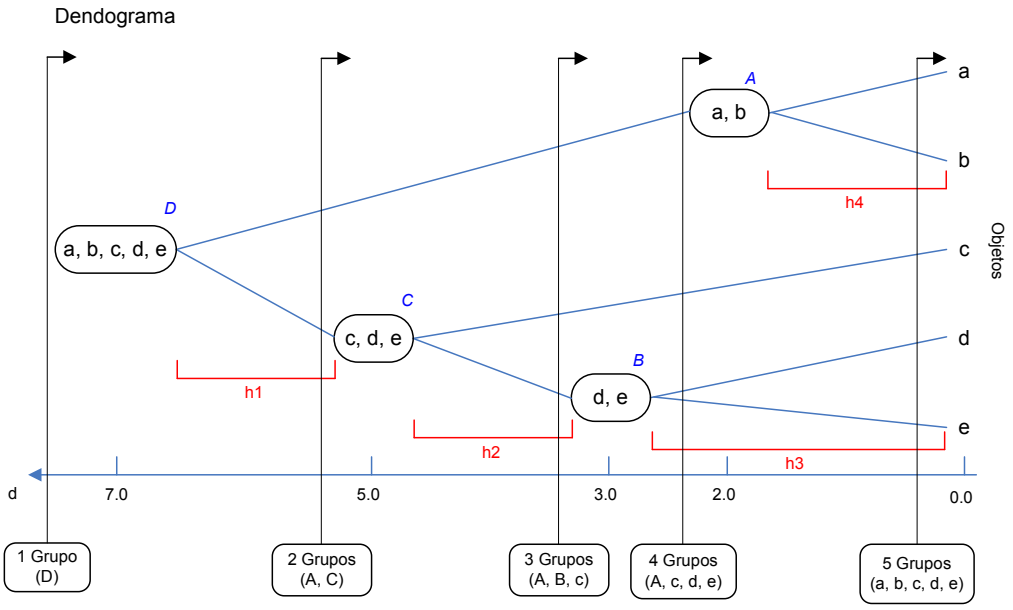

Figura 7: Método Hierárquico Divisivo - Dendograma

Analisando o gráfico existe inicialmente um único agrupamento *D*. Esse agrupamento é dividido em dois agrupamentos *A* e *C*. A medida de similaridade dessa divisão é expressa por *h1*. Nesse momento existem 2 agrupamentos (*C,A*). No passo seguinte, o agrupamento *C* é dividido em dois agrupamentos *c* e *B*. A medida de similaridade dessa divisão é expressa por *h2*. Nesse momento existem 3 agrupamentos (*A, B, c*). O agrupamento *B* é então dividido entre os agrupamentos *d* e *e*. A medida de similaridade dessa divisão é expressa por *h3*. Nesse momento existem 4 agrupamentos (*A, c, d, e*). No último passo, o agrupamento *A* é dividido entre os agrupamentos *a* e *b*. A medida de similaridade dessa divisão é expressa por *h4*. Nesse momento existem 5 agrupamentos (*a, b, c, d, e*).

3.1.2.2. Coeficiente Divisivo (CD)

Mede a qualidade de um agrupamento divisivo de dados (Kauffman, 1990). Para cada objeto *i*, *d(i)* é o diâmetro do último agrupamento ao qual o objeto pertenceu (antes de ser dividido em um agrupamento de um único objeto), dividido pelo diâmetro de todo o conjunto de dados. O coeficiente é então definido como:

$$CD = \frac{1}{n}\sum_{i}^{n} d(i) \qquad (3.2)$$

onde *n* é o número total de objetos do conjunto de dados.

Os valores do coeficiente variam entre 0 e 1. Valores baixos do coeficiente correspondem a estruturas ruins, onde nenhum agrupamento foi encontrado. Por outro lado, valores próximos a 1 representam que estruturas muito claras foram identificadas.

3.1.2.3. Banner de Dissimilaridade

Em (Kauffman, 1990), a hierarquia dos agrupamentos pode ser representada graficamente pelo Banner de Dissimilaridade. Essa representação mostra as sucessivas divisões entre agrupamentos. A leitura do banner é feita da esquerda para a direita. Os objetos são listados verticalmente. A divisão de dois agrupamentos é representada por uma barra horizontal que termina na região de dissimilaridades dos agrupamentos envolvidos. O coeficiente divisivo (CD) pode ser visto como a largura média do banner. A Figura 8 mostra um exemplo de um banner de dissimilaridade para métodos hierárquicos divisivos.

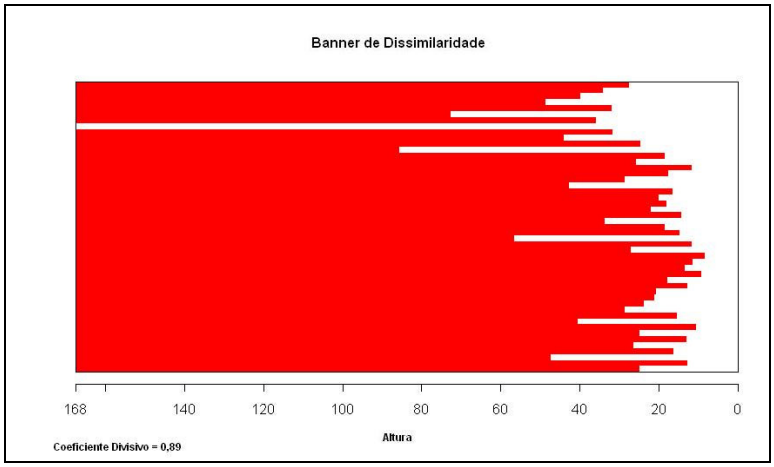

Figura 8: Banner de Dissimilaridade (Métodos Hierárquicos Divisivos).

A Figura 9 mostra um exemplo onde se pode ver a relação entre o dendograma e o banner.

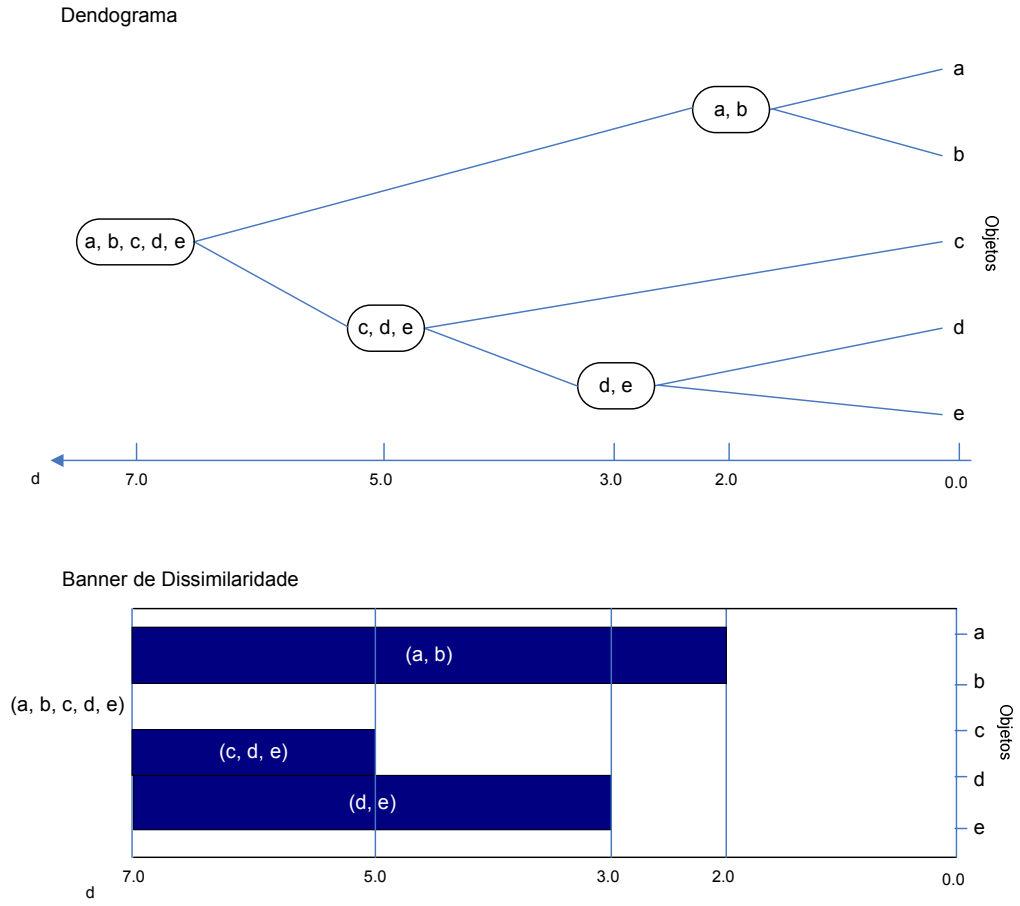

Figura 9: Métodos Hierárquicos Divisivos: dendograma e banner.

No exemplo acima, observando o dendograma se pode notar que o agrupamento *{c, d, e}* foi dividido em dois agrupamentos *c* e *{d, e}* e a similaridades entre eles é 5. Observando agora o banner, na medida de similaridade 5 ocorre a divisão entre os agrupamentos *c* e *{d, e}*. Essa mesma análise pode ser feita para os demais agrupamentos.

3.1.3. Métodos de Distância entre Grupos

Como foi dito anteriormente, no início do algoritmo aglomerativo cada agrupamento é formado por um único padrão. Sendo assim, o grau de similaridade entre os agrupamentos se resume ao grau de similaridades entre os elementos, que nesse caso, pode ser calculado através das medidas de distância vistas anteriormente (por exemplo, a distância euclidiana). Acontece que nos passos seguintes do algoritmo, os agrupamentos conterão mais de um padrão.

Existem vários métodos para medir a distância entre grupos, dentre as quais as mais importantes são (Bao, 2004; Simon, 2004):

- *Ligação Simples*

  A distância entre dois agrupamentos é dada pela distância entre os seus padrões mais similares.

$$D(C_1, C_2) = \min_{\substack{i \in C_1 \\ j \in C_2}} (d(i, j)) \qquad (3.3)$$

  onde *i* e *j* são respectivamente os padrões dos agrupamentos C_1 e C_2 e $d(i, j)$ é a distância entre os objetos *i* e *j*.

  Algumas das características desse método são (Viana, 2004):

  - Em geral grupos muito próximos podem não ser identificados;
  - Permite detectar grupos de formas não-elípticas;
  - Apresenta pouca tolerância a ruído, pois tem a tendência a incorporar os ruídos em um grupo já existente;
  - Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
  - Tendência a formar longas cadeias (encadeamento): um primeiro grupo de um ou mais elementos passa a incorporar, a cada iteração, um grupo de apenas um elemento. Assim, é formada uma longa cadeia, onde se torna difícil definir um nível de corte para classificar os elementos em grupos.

- *Ligação Completa*

  A distância entre dois agrupamentos é dado pela distância entre os seus padrões menos similares.

$$D(C_1, C_2) = \max_{\substack{i \in C_1 \\ j \in C_2}} (d(i, j)) \qquad (3.4)$$

  onde *i* e *j* são respectivamente os padrões dos agrupamentos C_1 e C_2 e $d(i, j)$ é a distância entre os objetos *i* e *j*.

  Algumas das características desse método são (Viana, 2004):

  - Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;

- Tendência a formar grupos compactos;
- Os ruídos demoram a serem incorporados ao grupo.

Os métodos ligação simples e ligação completa trabalham em direções opostas. Se eles apresentam resultados semelhantes, significa que o grupo está bem definido no espaço, ou seja, o grupo é real. Mas se ocorre o contrário, os grupos provavelmente não existem (Viana, 2004).

- *Centróide*

A distância entre dois agrupamentos é dado pela distância entre os centróides. O centróide é a média dos padrões do agrupamento.

$$D(C_1, C_2) = d(\mu_1, \mu_2) \qquad (3.5)$$

onde μ_1 e μ_2 são respectivamente os centróides dos agrupamentos C_1 e C_2 e $d(\mu_1, \mu_2)$ é a distância entre eles.

Algumas das características desse método são:

- Ao se usar uma medida de distância *d* na equação 3.5 que não seja a distância euclidiana, podem levar a resultados estranhos, e por isso não é recomendada (Kauffman, 1990);
- Robustez à presença de ruídos (Doni, 2004).

- *Média das Ligações*

A distância entre dois agrupamentos é dada pela média das distâncias entre cada padrão de um agrupamento em relação aos padrões do outro agrupamento.

$$D(C_1, C_2) = \frac{N_1 \underset{i \in C_1}{d}(i, C_1) + N_2 \underset{j \in C_2}{d}(j, C_2)}{N_1 + N_2} \qquad (3.6)$$

onde N_1 e N_2 são respectivamente os números de objetos dos agrupamentos C_1 e C_2 e i e j são respectivamente os padrões das classes C_1 e C_2.

Algumas das características desse método são (Viana, 2004):

- Menor sensibilidade a ruídos que o os métodos ligação simples e ligação completa;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Tendência a formar grupos com número de elementos similares.

- *Média do Grupo das Ligações*

  A distância entre dois agrupamentos é dada pela média das distâncias entre os padrões da união dos dois agrupamentos relacionados.

$$D(C_1, C_2) = \frac{1}{N_1 N_2} \sum_{\substack{i \in C_1 \\ j \in C_2}} d(i, j) \quad (3.7)$$

  onde N_1 e N_2 são respectivamente os números de objetos dos agrupamentos C_1 e C_2 e i e j são respectivamente os padrões das classes C_1 e C_2.

  Algumas das características desse método são (Everitt, 2001):

  - As dissimilaridades entre os agrupamentos são estatisticamente consistentes;
  - Apresenta bons resultados na prática;
  - Pode ser aplicada em dados que não estão restritos a distância euclidiana.

- *Ward*

  O método Ward procura por partições que minimizem a perda associada a cada agrupamento (Ward, 1963). Essa perda é quantificada pela diferença entre a soma dos erros quadráticos de cada padrão e a média da partição em que está contido. A soma dos erros quadráticos para cada agrupamento é definida como:

$$ESS_k = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \quad (3.8)$$

  onde k é o agrupamento em questão, n é o número total de objetos do agrupamento k e x_i é o i-ésimo objeto do agrupamento k.

  Algumas das características desse método são (Viana, 2004):

- Apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias;
- Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual;
- Tem tendência a combinar grupos com poucos elementos;
- Sensível à presença de *outliers*.

3.1.4. Métodos Hierárquicos Conhecidos

3.1.4.1. Agglomerative Nesting (AGNES)

Agnes é um método hierárquico aglomerativo. A união entre agrupamentos é feita entre os agrupamentos com a menor dissimilaridade entre si.

Suponha que dois agrupamentos *A* e *B* tenham sido agrupados em um novo agrupamento *R*, no passo seguinte terá que ser calculada a dissimilaridade do agrupamento *R* em relação aos demais agrupamentos.

De acordo com Kauffman & Rousseeuw (1990), supondo:

- *Q* um agrupamento qualquer;
- *R* o agrupamento resultante da união entre os agrupamentos *A* e *B*;
- |*A*|, |*B*|, |*R*| e |*Q*| os números de objetos dos agrupamentos *A*, *B*, *R* e *Q* respectivamente;
- *d(., .)* a dissimilaridade entre os objetos relacionados.

A dissimilaridade entre dois agrupamentos *R* e *Q* será definida como:

$$d(R,Q) = \frac{1}{|R|.|Q|}.\sum_{\substack{i\in R\\ j\in Q}} d(i,j)$$

$$= \frac{1}{|R|.|Q|}.\sum_{\substack{i\in A\\ j\in Q}} d(i,j) + \frac{1}{|R|.|Q|}.\sum_{\substack{i\in B\\ j\in Q}} d(i,j)$$

$$= \frac{|A|}{|R|}\left(\frac{1}{|A|.|Q|}.\sum_{\substack{i\in A\\ j\in Q}} d(i,j)\right) + \frac{|B|}{|R|}\left(\frac{1}{|B|.|Q|}.\sum_{\substack{i\in B\\ j\in Q}} d(i,j)\right)$$

$$d(R,Q) = \frac{|A|}{|R|}d(A,Q) + \frac{|B|}{|R|}d(B,Q) \qquad (3.9)$$

Pode-se notar que as dissimilaridades *d(A, Q)* e *d(B, Q)* estão disponíveis na matriz de similaridades no passo anterior à formação do agrupamento *R*. Com isto, é necessário guardar apenas uma matriz de similaridades durante todo o processo de agrupamento.

As principais vantagens desse método residem na simplicidade da expressão 3.9 e no tempo de computação (Kauffman, 1990).

3.1.4.2. Divisive Analysis (DIANA)

DIANA é um método hierárquico divisivo. O algoritmo consiste de *n-1* divisões sucessivas, onde *n* é o número de dados do conjunto de dados. Em cada passo é selecionado o agrupamento C com o maior diâmetro, definido como:

$$diam(C) = \max_{i,j \in C} d(i, j) \qquad (3.10)$$

onde *d(i, j)* é a dissimilaridade entre os objetos *i* e *j* pertencentes ao agrupamento *C*.

Considerando *diam(C)>0*, o agrupamento *C* pode ser dividido em dois agrupamentos *A* e *B* (Struyf, 1996):

1. Primeiramente considera-se que o agrupamento $A = C$ e $B=\varnothing$
2. Mover um objeto de *A* para *B*

   Para cada objeto $i \in A$, é calculada a dissimilaridade média do objeto i para todos os objetos de *A* denotado por *a(i)*. O objeto *m* de *A* para o qual *a(m)* é o maior, é movido para *B*.
3. Mover outros objetos de *A* para *B* segundo a regra abaixo:

   Se total de objetos de $A = 1 \Rightarrow$ Pare

   Senão para $i \in A$ calcule a dissimilaridade média de *i* para todos os objetos pertencentes a *A* denotado por *a(i)* e a dissimilaridade média de *i* para todos os objetos pertencentes a *B* denotado por *d(i,B)*. Selecione o objeto $h \in A$ para o qual:

   $$a(h) - d(h, B) = \max_{i \in A} (a(i) - d(i, B)) \qquad (3.11)$$

   onde *d(h, B)* é a dissimilaridade média entre *h* e *B*.
4. Se $a(h) - d(h,B) > 0 \Rightarrow$ mover *h* de *A* para *B* e repetir o passo 2.

5. Se $a(h) - d(h,B) \leq 0 \Rightarrow$ Manter A e B e parar o processo.

3.1.4.3. Monothetic Analysis (MONA)

O método MONA é destinado exclusivamente a dados do tipo binário. Apesar do algoritmo ser hierárquico divisivo, ele não usa dissimilaridades entre objetos, e por isso a matriz de dissimilaridades não é computada (Kauffman, 1990). A divisão em agrupamentos utiliza as variáveis diretamente.

De acordo com Kauffman & Rousseeuw (1990), o primeiro passo do algoritmo consiste em selecionar uma das variáveis da matriz de dados e agrupar os dados cujo valor é igual a 1 e os dados cujo valor é igual a 0. No passo seguinte, cada agrupamento obtido no passo anterior é dividido mais adiante, usando a mesma lógica do primeiro passo. O processo continua até que cada agrupamento contenha apenas 1 objeto. O exemplo abaixo demonstra o uso desta técnica em uma base bem simples contendo três variáveis.

Variável 1	Variável 2	Variável 3
1	1	1
0	0	1
1	0	0

1) Variável selecionada: Variável 2

Grupo 1:

Variável 1	Variável 2	Variável 3
1	1	1

Grupo 2:

Variável 1	Variável 2	Variável 3
0	0	1
1	0	0

2) Variável selecionada: Variável 1

Grupo 1:

Variável 1	Variável 2	Variável 3
1	1	1

Grupo 2:

Variável 1	Variável 2	Variável 3
0	0	1

Grupo 3:

Variável 1	Variável 2	Variável 3
1	0	0

Figura 10: Exemplo de agrupamento de dados binário usando MONA.

O critério de seleção da variável da matriz de dados em cada passo é pela variável que tenha um grau de associação mais forte em relação as demais. A medida de associação entre duas variáveis (f e g) é definida da seguinte maneira:

$$A_{fg} = \left| a_{fg} d_{fg} - b_{fg} c_{fg} \right| \qquad (3.12)$$

onde A_{fg} é a medida de associação entre *f* e *g*, a_{fg} é o número de objetos *i* onde $x_{if}=x_{ig}=0$, d_{fg} é o número de objetos *i* onde $x_{if}=x_{ig}=1$, b_{fg} é o número de objetos *i* onde $x_{if}=0$ e $x_{ig}=1$ e c_{fg} é o número de objetos *i* onde $x_{if}=1$ e $x_{ig}=0$. Considere x_{if} e x_{ig} respectivamente como os valores do dado *i* da variável *f* e *g*.

A medida A_{fg} expressa se as variáveis *f* e *g* fornecem divisões similares em um conjunto de objetos, e por isso pode ser considerada como um tipo de similaridade entre variáveis.

A medida total de associação de uma variável em relação a todas as outras é dada pela fórmula abaixo.

$$A_f = \sum_{g \neq f} A_{fg} \qquad (3.13)$$

Essa é a medida usada como critério de escolha de uma variável em cada divisão, onde é selecionado a variável com a maior medida.

3.2. Métodos Particionais

Os métodos particionais são métodos baseados na minimização de uma função de custo, onde os padrões são agrupados em um número *k* de agrupamentos escolhido a priori. Cada padrão é agrupado na classe em que essa função de custo é minimizada.

Uma das principais vantagens dos métodos particionais em relação aos métodos hierárquicos é a possibilidade de um padrão poder mudar de agrupamento com a evolução do algoritmo e a possibilidade de se operar com bases de dados maiores. Os métodos particionais são extremamente mais rápidos que os métodos hierárquicos (Fung, 2001).

As principais desvantagens dos métodos particionais estão no fato do número de agrupamentos ter que ser escolhido a priori, o que poderá sugerir

interpretações erradas sobre a estrutura dos dados caso o número de agrupamentos não seja o ideal e no fato de que o algoritmo é em geral sensível às condições iniciais, podendo gerar resultados diferentes a cada rodada (Fung, 2001). O problema quando se escolhe erroneamente o número de agrupamentos é que o método irá impor uma estrutura aos dados, no lugar de buscar a estrutura inerente a estes (Kainulainen, 2002).

3.2.1. Métodos Não-Exclusivos

A segmentação de dados numéricos forma a base de muitos algoritmos de classificação. O seu propósito é identificar agrupamentos naturais de dados para produzir uma representação concisa do comportamento do sistema.

É comum em um processo de agrupamento de dados que cada objeto, representado pelo dado, pertença a um único agrupamento. Os métodos não-exclusivos, conhecidos também como métodos *fuzzy*, permitem alguma ambigüidade entre os dados, o que geralmente acontece (Everitt, 2001).

Esses métodos são técnicas de agrupamento de dados onde cada padrão pertence a um agrupamento com um certo grau de pertinência. Como exemplo, considere-se a Figura 11.

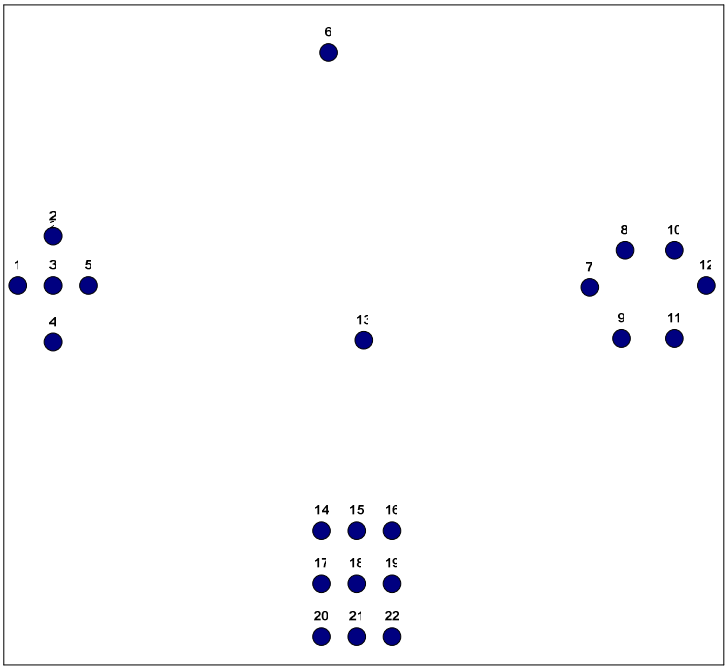

Figura 11: Exemplo de um conjunto de 22 dados.

Podem-se identificar claramente três agrupamentos:

{1, 2, 3, 4, 5},

{7, 8, 9, 10, 11, 12} e

{14, 15, 16, 17, 18, 19, 20, 21, 22}.

Quanto aos pontos 6 e 13, provavelmente seria necessário efetuar uma escolha arbitrária. Os métodos de agrupamento fuzzy lidam melhor com este tipo de situação, pois a cada elemento é atribuído um grau de pertinência a cada agrupamento, permitindo uma análise mais rica da distribuição dos dados nos diversos grupos.

Em (Everitt, 2001) a principal vantagem dos agrupamentos fuzzy em relação aos demais métodos particionais está no fato de representar com muito mais detalhes a informação sobre a estrutura dos dados. Por outro lado, isso poderia ser considerado uma desvantagem, pois a quantidade informação gerada cresce muito rapidamente com o número de objetos e o número de agrupamentos, tornando a compreensão mais difícil. Outra desvantagem é a ausência de objetos representativos dos agrupamentos formados e o fato de que geralmente os algoritmos são mais complicados e consomem um maior tempo de computação. No entanto, os princípios fuzzy são muito interessantes, pois permitem a descrição de algumas incertezas que geralmente estão presentes em dados reais.

3.2.1.1. Coeficiente de DUNN

O coeficiente de DUNN mede o grau de generalização dos agrupamentos, ou seja, o quanto um agrupamento é 'fuzzy' (Everitt, 2001).

$$F_k = \sum_{i=1}^{N} \sum_{c=1}^{k} \frac{u_{ic}^2}{N} \qquad (3.14)$$

onde N representa o total de objetos em um conjunto de dados, k é o número de agrupamentos e u_{ic} é o grau de pertinência do objeto i em relação ao agrupamento c.

Os valores acima podem variar de $1/k$ até 1 (mais fuzzy).

A versão normalizada do coeficiente de DUNN onde os valores variam de 0 a 1 independentemente do número de agrupamentos k escolhidos é a seguinte:

$$F_k^{'} = \frac{F_k - (1/k)}{1 - (1/k)} = \frac{k \cdot F_k - 1}{k - 1} \qquad (3.15)$$

onde F_k é o coeficiente de DUNN não normalizado e k é o número de agrupamentos.

3.2.2. Métodos Particionais Conhecidos

3.2.2.1. K-Means

O método *k-means* é um dos métodos mais populares das técnicas particionais (Fung, 2001). O método particiona os dados em k agrupamentos mutuamente exclusivos. Diferentemente dos métodos hierárquicos, o *k-means* não cria uma estrutura em árvore para descrever o agrupamento dos dados e é mais adequado para uma grande quantidade de dados.

O algoritmo procura, dentro do possível, a partição em que os padrões de cada agrupamento estão mais próximos entre si e mais distantes dos padrões dos outros agrupamentos. O *k-means* é um algoritmo iterativo que minimiza a soma das distâncias de cada padrão ao centróide de cada agrupamento, sobre todos os agrupamentos. Este algoritmo move padrões entre os agrupamentos até que a função objetivo não se altere ou se altere muito pouco, ou até que o número de iterações máximo pré-determinado tenha sido alcançado. O resultado é um conjunto de agrupamentos compactos e bem separados tanto quando possível.

Em resumo, cada agrupamento é representado pelo centro do grupo e cada padrão é atribuído ao agrupamento que está mais próximo.

O procedimento geral pode ser descrito em poucos passos (Fung, 2001):

1. Inicializar as médias das k partições.
2. Para cada padrão determinar a partição mais próxima.
3. Calcular a média de cada partição.
4. Se houver mudança na média das partições, voltar ao passo 2.
5. Resultado: as médias das k partições.

Como muitos outros problemas de minimização, a solução encontrada pelo *k-means* geralmente depende do ponto de partida, mas em geral o algoritmo encontra um mínimo local. Trata-se de um método prático e computacionalmente eficiente, embora seja sensível a ruído e outliers e não é aplicável para agrupamentos não-convexo.

O resultado deste método pode, em muitos casos, ser drasticamente afetado pela escolha das condições iniciais (Kainulainen, 2002). Entretanto, em bases de dados bem estruturadas, em geral, espera-se a convergência para um mínimo global. Comportamentos como convergência lenta e resultado de agrupamentos bastante diferentes para diferentes configurações inicias pode indicar que o número de agrupamentos escolhido esteja errado, ou que os dados não possuam estrutura de agrupamentos.

O método apresenta bons resultados apenas quando os agrupamentos são hiperesféricos e possuem aproximadamente o mesmo número de padrões em cada agrupamento (Costa, 1999). O bom desempenho do algoritmo depende muito também da escolha adequada da medida de distância e do ponto inicial de partida do algoritmo.

3.2.2.2. Fuzzy C-Means

Fuzzy c-means (FCM) é uma técnica de agrupamento de dados fuzzy. Esta técnica foi originalmente introduzida por Jim Bezdek em 1981 (Matlab® Reference, 2005) como uma evolução das técnicas de agrupamento de dados mais recentes e fornece um método que mostra como agrupar padrões que pertencem a um espaço multidimensional em um número específico de diferentes agrupamentos.

O método começa com uma suposição inicial sobre os centros de cada agrupamento. Essa suposição inicial é na maioria das vezes incorreta. Para cada padrão é assinalado um grau de pertinência para cada agrupamento.

Iterativamente os centros de cada agrupamento, bem como os graus de pertinência de cada padrão são atualizados. Essa iteração tem como objetivo minimizar uma função objetivo que representa a distância de cada padrão em relação ao centro de cada agrupamento ponderado pelo grau de pertinência do

padrão. As pertinências dos padrões aos grupos podem assumir qualquer valor real no intervalo [0, 1].

Apesar dos padrões poderem pertencer a mais de um agrupamento, geralmente restringe-se a função de pertinência de forma que a soma das pertinências de um padrão seja igual a 1 nos K número escolhido de agrupamentos.

$$\sum_{k=1}^{K} u_{ik} = 1 \qquad (3.16)$$

onde i é um objeto do conjunto de dados, K é o número de agrupamentos, u_{ik} é valor de pertinência do objeto i em relação ao agrupamento k.

Dado os agrupamentos $\{C_1, C_2, ..., C_k\}$, os K centros dos agrupamentos $(v_1, v_2, ..., v_k)$ associados a cada agrupamento são obtidos por:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m . x_k}{\sum_{k=1}^{n} u_{ik}^m} \qquad (3.17)$$

onde v_i é o centro do agrupamento i, m é o coeficiente fuzzy, u_{ik} é valor de pertinência do objeto i em relação ao agrupamento k e x_k é o objeto k do conjunto de dados.

O coeficiente fuzzy $m \in (1, \infty)$ é um número real que controla a influência dos graus de pertinência. O vetor v_i, centro do agrupamento da classe fuzzy C_i, pode ser visto como uma média ponderada dos dados em C_i. Quando o parâmetro $m \to 1$, o *fuzzy c-means* converge para o método *k-means*, enquanto que quando $m \to \infty$, os centros dos agrupamentos ficam mais próximos do centróide do conjunto de dados (Fung, 2001) e a variância de cada agrupamento se torna maior, tornando os agrupamentos mais fuzzy.

O índice de desempenho é definido em termos dos centros dos agrupamentos por (Fung, 2001):

$$J_m(P) = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^m . \| x_i - v_k \|^2 \qquad (3.18)$$

onde J é o índice de desempenho do agrupamento P, n é o número total de objetos do conjunto de dados, x_i é o objeto i desse conjunto de dados, v_k é o centro do agrupamento k e u_{ik} é valor de pertinência do objeto i em relação ao agrupamento k.

O índice de desempenho mede a soma das distâncias ponderadas entre o padrão ao centro do agrupamento fuzzy. O objetivo desse método é minimizar o índice de desempenho.

Para executar o algoritmo é necessário definir o critério de distância, o número de agrupamentos K, o valor m $\in$ (1, ∞) e um valor ε como critério de parada.

O procedimento geral pode ser descrito em poucos passos (Fung, 2001):

1. Inicializar: K (número de agrupamentos), ε (critério de parada), m (coeficiente fuzzy), U^0 (matriz inicial com os graus de pertinência);
2. Calcular o centro dos k agrupamentos ($\boldsymbol{v_1^{(t)}}$, $\boldsymbol{v_2^{(t)}}$,..., $\boldsymbol{v_k^{(t)}}$), onde t é a iteração corrente.
3. Atualizar a matriz U^{t+1} segundo a equação abaixo:

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{K} \left(\frac{\left\| x_i - v_k^t \right\|^2}{\left\| x_i - v_j^t \right\|^2} \right)^{1/m-1} \right]^{-1} \qquad (3.19)$$

   onde $u_{ik}^{(t+1)}$ é o grau de pertinência do objeto i ao agrupamento k na iteração $t+1$ e x_i é o objeto i do conjunto de dados.

4. Calcular $\nabla = \left\| U^{t+1} - U^t \right\|$. Se $\nabla < \varepsilon \Rightarrow$ fim do algoritmo, senão, $t = t+1$ e voltar ao passo 2.

A Figura 12 mostra a dissimilaridade média entre os agrupamentos para diferentes valores do coeficiente fuzzy *m*. Ou seja, agrupamentos muito parecidos terão valores de dissimilaridade próximos a 0, enquanto que agrupamentos muito diferentes terão valores de dissimilaridade próximos de 1. Em agrupamentos fuzzy é esperada alguma interseção entre os agrupamentos, enquanto que em agrupamentos não-fuzzy não há interseção entre os agrupamentos.

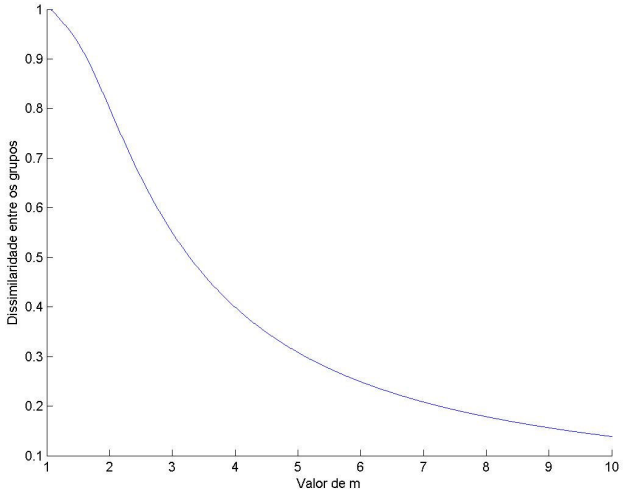

Figura 12: Gráfico Dissimilaridade entre os grupos X Valor de m.

Na Figura 12 pode-se notar que, quando o valor de *m* tende a 1, o algoritmo tende ao método *k-means*, onde a dissimilaridade média entre os grupos é 1 (um padrão pertence exclusivamente a um grupo), enquanto que com o aumento do valor de *m* os agrupamentos se tornam mais fuzzy. O parâmetro *m* é selecionado de forma ad hoc (Costa, 1999). O valor comumente usado é o valor 2 (Rashid, 2005), quando se pode verificar alguma similaridade entre os agrupamentos, mantendo ainda as suas características próprias. A escolha desse valor é muito subjetiva, dependendo exclusivamente de como se queira tratar o problema.

3.2.2.3. Fuzzy Analysis (FANNY)

Fuzzy Analysis (FANNY) é uma técnica de agrupamento de dados fuzzy proposta por Kauffman (1990). Assim como o *FCM*, este método atribui a cada padrão um grau de pertinência aos agrupamentos envolvidos, com a vantagem de ser mais robusto (Neville, 2003). O algoritmo roda iterativamente e pára quando a função objetivo converge, gerando assim estimativas para os *k* agrupamentos.

Implicitamente, no *FCM* considera-se que cada objeto, representado como o centro de cada agrupamento, é dado pela média das coordenadas em um espaço p-dimensional. O método *FANNY* não possui representações de tais objetos, sendo necessárias apenas as distâncias ou as dissimilaridades entre os dados.

O método *FANNY* pode operar com matrizes de dados onde estes são medidas de observações (idade, sexo, altura, renda familiar, etc.), ou com dados que são medidas de similaridade (Simões, 2004).

É possível fazer uma comparação direta entre os dois métodos quando os dados são medidas de observações (Kauffman, 1990). Nesse caso, a função objetivo se torna exatamente a mesma, demonstrando a equivalência dos métodos para esse tipo de situação. Em termos de desempenho, o método *FANNY* é mais lento do que o *FCM*.

O algoritmo tem como objetivo minimizar a função objetivo:

$$C = \sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2} \qquad (3.20)$$

onde k é o número de agrupamentos, n é o número de objetos do conjunto de dados, e u_{iv} e u_{jv} são respectivamente os graus de pertinência do objeto i e j do conjunto de dados em relação ao agrupamento v.

O procedimento geral pode ser descrito pelos passos:

1. Inicializar: k (número de agrupamentos), ε (critério de parada), U^0 (matriz inicial com os graus de pertinência);
2. Calcular para cada i=1,...,n os seguintes valores:
   2.1. Calcular para cada $v=1,...,k$

$$
\begin{aligned}
{}^{m}a_{iv} = {} & \frac{2\left(\sum_{j=1}^{i-1} {}^{m+1}u_{jv}^2 d(i,j) + \sum_{j=i}^{n} {}^{m}u_{jv}^2 d(i,j)\right)}{\sum_{j=1}^{i-1} {}^{m+1}u_{jv}^2 + \sum_{j=i}^{n} {}^{m}u_{jv}^2} \\
& - \frac{\begin{array}{l}\sum_{j=1}^{i-1}\sum_{h=1}^{i-1} {}^{m+1}u_{jv}^2 {}^{m+1}u_{hv}^2 d(i,j) + \sum_{j=1}^{i-1}\sum_{h=i}^{n} {}^{m+1}u_{jv}^2 {}^{m}u_{hv}^2 d(i,j) \\ + \sum_{j=i}^{n}\sum_{h=1}^{i-1} {}^{m}u_{jv}^2 {}^{m+1}u_{hv}^2 d(i,j) + \sum_{j=i}^{n}\sum_{h=i}^{n} {}^{m}u_{jv}^2 {}^{m}u_{hv}^2 d(i,j)\end{array}}{\sum_{j=1}^{i-1} {}^{m+1}u_{jv}^2 + \sum_{j=i}^{n} {}^{m}u_{jv}^2}
\end{aligned}
\qquad (3.21)
$$

   onde a_{iv} é a medida referente ao objeto i do conjunto de dados no agrupamento v, u_{jv} e u_{hv} são respectivamente os graus de pertinência do objeto j e h do conjunto de dados em relação ao agrupamento v, $d(i, j)$ é a dissimilaridade entre os objetos i e j do conjunto de dados e m e $m+1$ são apenas índices indicando valor corrente e valor seguinte.

   2.2. Calcular para cada $v=1,...,k$:

$$A_v = \frac{1/a_{iv}^{m}}{\sum_{w=1}^{k}\left(1/a_{iw}^{m}\right)} \qquad (3.22)$$

2.2.1. Se $A_v \leq 0 \Rightarrow V- = V- \cup \{v\}$

onde $V-$ é o conjunto contendo os agrupamentos v onde A_v é menor ou igual a 0.

2.2.2. Se $A_v > 0 \Rightarrow V+ = V+ \cup \{v\}$

onde $V+$ é o conjunto contendo os agrupamentos v onde A_v é maior que 0.

2.3. Para todos $v \in V-$ => ${}^{m+1}u_{iv} = 0$

2.4. Calcular para todos $v \in V+$

$${}^{m+1}u_{iv} = \frac{1/a_{iv}^{m}}{\sum_{w \in V+}\left(1/a_{iw}^{m}\right)} \qquad (3.23)$$

2.5. Reinicializar os conjuntos $V-$ e $V+$ => $V- = V+ = \varnothing$

2.6. Voltar ao passo 2.1 com o próximo valor de i

3. Calcular a nova função objetivo ${}^{m+1}C$. Se $({}^{m}C/{}^{m+1}C\text{-}1) < \varepsilon$ então voltar para o passo 2, senão fim do algoritmo.

3.2.2.4. Gustafson-Kessel

Esse método é uma extensão do método *FCM* e tem o objetivo de detectar agrupamentos de formas geométricas diferentes, se adaptando, dessa forma, a diferentes estruturas (Balasko, 2005).

O método de Gustafson-Kessel pode ser definido com um método adaptativo. A sua principal motivação é encontrar agrupamentos não-esféricos.

A idéia básica do algoritmo está na utilização de medidas não-esféricas de distância específicas para cada agrupamento. Essas medidas de distância evoluem com o tempo.

Para $i=1,...,K$ e $j=1,...,n$, onde K é o número de agrupamentos e n o número de objetos do conjunto de dados, a medida não-esférica de distância para cada agrupamento pode ser definida por:

$$d_i(x_j, v_i) = \sqrt{(x_j - v_i)^T . A_i . (x_j - v_i)} \qquad (3.24)$$

onde x_j é o objeto j do conjunto de dados, v_i é o centro do agrupamento i e

$$A_i = (\det(S_i))^{1/n} . S_i^{-1} \qquad (3.25)$$

e

$$S_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^m} . \sum_{k=1}^{n} u_{ik}^m . (x_k - v_i)(x_k - v_i)^T \qquad (3.26)$$

onde u_{ik} é o grau de pertinência do objeto i do conjunto de dados em relação ao agrupamento k e m é o coeficiente fuzzy.

Assim como o *FCM*, é necessário definir o número de agrupamentos K, o valor $m \in (1, \infty)$ e um valor ε como critério de parada.

O procedimento geral pode ser descrito em poucos passos (Balasko, 2005):

1. Inicializar: k (número de agrupamentos), ε (critério de parada), m (coeficiente fuzzy), U^0 (matriz inicial com os graus de pertinência).
2. Calcular o centro dos k agrupamentos ($\boldsymbol{v_1^{(t)}}$, $\boldsymbol{v_2^{(t)}}$,..., $\boldsymbol{v_k^{(t)}}$) descrito no método *FCM*.
3. Calcular as distâncias $d_i(x_j, v_i)$ para $i=1,...,K$ e $j=1,...,n$.
4. Atualizar a matriz U^{t+1} segundo a equação abaixo:

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{K} \left(\frac{d_j(x_k, v_i)^2}{d_j(x_k, v_j)^2} \right)^{1/m-1} \right]^{-1} \qquad (3.27)$$

   onde u_{ik} é o grau de pertinência do objeto i do conjunto de dados em relação ao agrupamento k.

5. Calcular $\nabla = \left\| U^{t+1} - U^t \right\|$. Se $\nabla < \varepsilon \Rightarrow$ fim do algoritmo, senão, $t = t+1$ e voltar ao passo 2.

3.2.2.5. Gath-Geva

Gath-Geva é uma técnica de agrupamento de dados fuzzy que tem como propósito detectar agrupamentos de formas, tamanhos e densidades diferentes (Balasko, 2005). O algoritmo contém apenas o parâmetro de fuzzificação m como

entrada para o algoritmo e utiliza uma medida de distância baseada na estimativa de máxima probabilidade fuzzy.

Para *i=1,...,K* e *j=1,...,n*, onde *K* é o número de agrupamentos e *n* o número de objetos do conjunto de dados, a medida de distância pode ser definida como:

$$d_i(x_j, v_i) = \frac{\sqrt{\det(S_i)}}{\frac{1}{n}\sum_{k=1}^{n} u_{ik}} \exp\left(\frac{1}{2}(x_j - v_i)^T S_i^{-1}(x_j - v_i)\right) \qquad (3.28)$$

onde x_j é o objeto *j* do conjunto de dados, v_i é o centro do agrupamento *i*, u_{ik} é o grau de pertinência do objeto *i* do conjunto de dados em relação ao agrupamento *k* e

$$S_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^m} . \sum_{k=1}^{n} u_{ik}^m . (x_k - v_i)(x_k - v_i)^T \qquad (3.29)$$

onde u_{ik} é o grau de pertinência do objeto *i* do conjunto de dados em relação ao agrupamento *k* e *m* é o coeficiente fuzzy.

Este método é pouco robusto, pois depende da inicialização de sua matriz de pertinências já que, devido ao fato da medida de distância ser exponencial, o algoritmo converge para um mínimo local próximo (Balasko, 2005). Esse problema pode ser resolvido ou minimizado se for utilizado o resultado do método *FCM* para inicializar a matriz de pertinências.

Assim como o *FCM*, é necessário definir o número de agrupamentos *K*, o valor $m \in (1, \infty)$ e um valor ε como critério de parada.

O procedimento geral pode ser descrito em poucos passos (Balasko, 2005):

1. Inicializar: *k* (número de agrupamentos), ε (critério de parada), *m* (coeficiente fuzzy), U^0 (matriz inicial com os graus de pertinência).
2. Calcular o centro dos *k* agrupamentos ($\boldsymbol{v_1}^{(t)}$, $\boldsymbol{v_2}^{(t)}$,..., $\boldsymbol{v_k}^{(t)}$) descrito no método *FCM*.
3. Calcular as distâncias $d_i(x_j, v_i)$ para *i=1,...,K* e *j=1,...,n*.
4. Atualizar a matriz U^{t+1} segundo a equação abaixo:

$$u_{ik}^{(t+1)} = \left[\sum_{j=1}^{K}\left(\frac{d_j(x_k, v_i)^2}{d_j(x_k, v_j)^2}\right)^{\frac{1}{m-1}}\right]^{-1} \qquad (3.30)$$

onde u_{ik} é o grau de pertinência do objeto *i* do conjunto de dados em relação ao agrupamento *k* e *m* é o coeficiente fuzzy..

5. Calcular $\nabla = \left\| U^{t+1} - U^{t} \right\|$. Se $\nabla < \varepsilon \Rightarrow$ fim do algoritmo, se não, $t = t+1$ e voltar ao passo 2.

3.2.2.6. Mistura de Densidades

Nesse método os dados são vistos como provenientes de uma função de densidade de probabilidade, cada função representando um agrupamento diferente. Dessa forma, os agrupamentos podem ser consideradas como uma soma de gaussianas ponderadas pela probabilidade *a priori* de cada agrupamento (Hamerly, 2003; Duda, 2001).

O resultado do método se assemelha ao Teorema de Fourier onde qualquer função pode ser aproximada por uma soma de cossenoides. Neste caso, qualquer função de densidade de probabilidade pode ser aproximada por uma soma de gaussianas com parâmetros desconhecidos a serem estimados (Duda, 2001).

Existem duas formas de estimar esses parâmetros a partir dos dados (Duda, 2001):

- Máxima Verossimilhança
- Aprendizado Bayesiano

- **Máxima Verossimilhança**

Considerando θ como a matriz contendo os parâmetros a serem estimados, n o número total de dados, a verossimilhança de um conjunto de dados $\mathcal{X} = \{x_1,..., x_n\}$ pode ser definido como se segue abaixo.

$$p(\mathcal{X} \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta) \tag{3.31}$$

onde $p(\mathcal{X}|\theta)$ e $p(x_k|\theta)$ são respectivamente as probabilidades de $\mathcal{X}$ e x_k dado a matriz θ.

A estimativa de máxima verossimilhança $\hat{\theta}$ é o valor de θ que maximiza $p(\mathcal{X}|\theta)$ (Duda, 2001). Esse valor pode ser calculado através do gradiente do logaritmo da expressão acima e pode ser escrito como:

$$\nabla_{\theta_i} l = \sum_{k=1}^{n} P(w_i \mid x_k, \theta) \nabla_{\theta_i} \log\left(p(x_k \mid w_i, \theta_i)\right) \tag{3.32}$$

onde θ_i é o parâmetro i a ser estimado, l é o logaritmo da verossimilhança, w_i é o agrupamento i, x_k é o objeto k do conjunto de dados, ∇ é o gradiente e $P(w_i|x_k, \theta)$ é a probabilidade de w_i dado x_k e θ.

Os parâmetros estimados terão que satisfazer as condições (Duda, 2001):

$$\sum_{k=1}^{n} P\left(w_i \mid x_k, \hat{\theta}\right) \nabla_{\theta_i} \log\left(p\left(x_k \mid w_i, \hat{\theta}_i\right)\right) = 0 \tag{3.33}$$

para $i=1,...,K$

- **Aprendizado Bayesiano Não Supervisionado**

Esse método utiliza a regra de Bayes para estimar a função de densidade de probabilidade dos agrupamentos com base na distribuição *a priori* e as informações contidas nos dados (Duda, 2001).

Considerando:

- θ - matriz contendo os parâmetros a serem estimados;
- n - número total de dados;
- $\mathcal{X}$ - conjunto de dados $\{x_1,..., x_n\}$;
- w_i - agrupamento i;
- $P(w_i)$ – probabilidade *a priori* do agrupamento i;
- $p(x|\theta, w_i)$ – gaussiana onde θ é desconhecido.

  Exemplo: $p(x|\theta, w_i) \sim N(\mu_i, \sigma_i)$ – parâmetros desconhecidos μ_i e σ_i.

O procedimento geral pode ser descrito em poucos passos (Duda, 2001):

1. Calcular iterativamente $p(\theta|\mathcal{X})$ segundo a expressão abaixo para os n dados.

$$p\left(\theta \mid \chi^n\right) = \frac{p\left(\chi_n \mid \theta\right) p\left(\theta \mid \chi^{n-1}\right)}{\int p\left(\chi_n \mid \theta\right) p\left(\theta \mid \chi^{n-1}\right) d\theta} \tag{3.34}$$

Exemplo para os primeiros termos da iteração:

$$p\left(\theta \mid x_1\right) = \frac{p\left(x_1 \mid \theta\right) p(\theta)}{\int p\left(x_1 \mid \theta\right) p(\theta) d\theta}$$

e

$$p(\theta \mid x_1, x_2) = \frac{p(x_2 \mid \theta) p(\theta \mid x_1)}{\int p(x_2 \mid \theta) p(\theta \mid x_1) d\theta}$$

2. Calcular $p(\theta|w_i, \chi)$ segundo a expressão abaixo.

$$p(\theta \mid w_i, \chi) = \int p(x \mid \theta_i, w_i) p(\theta \mid \chi) d\theta \qquad (3.35)$$

3. Pela regra de Bayes, calcular $p(w_i|x)$ – probabilidade a posteriori de uma padrão x pertencer ao agrupamento w_i.

$$p(w_i \mid x) = \frac{p(x \mid \theta_i, w_i) P(w_i)}{\sum_{j=1}^{k} p(x \mid \theta_j, w_j) P(w_j)} \qquad (3.36)$$

4. O padrão x pertencerá ao agrupamento onde $p(w_i|x)$ for maior.

3.2.2.7. Partitioning Around Medoids (PAM)

Esse algoritmo procura por k objetos chamados de medóides (WinIDAMS, 2005a). Os medóides são objetos representativos de cada agrupamento e contêm os padrões onde a dissimilaridade média dos padrões pertencentes a um dado agrupamento é mínima. Em outras palavras, esse algoritmo minimiza a soma das dissimilaridades.

O algoritmo possui duas fases:

- *Construção*

  Essa é a fase onde os medóides são construídos. Eles são obtidos através de k seleções de objetos representativos. O primeiro objeto corresponde ao padrão onde a soma das dissimilaridades entre todos os padrões é mínima. Os objetos subseqüentes são selecionados de forma a minimizar a função objetivo o máximo possível.

  A função objetivo é definida como a soma das dissimilaridades de todos os objetos ao medóide mais próximo, conforme a expressão abaixo:

$$\sum_{i=1}^{N} d(i, m(i)) \qquad (3.37)$$

  onde N é o total de dados, i é o objeto do conjunto de dados, $m(i)$ é o medóide mais próximo do objeto i e $d(i,m(i))$ é a dissimilaridade entre i e $m(i)$.

De acordo com Kauffman & Rousseeuw (1990), os objetos podem ser encontrados de acordo com os passos abaixo.

1. Considere-se um objeto *i* que não tenha sido selecionado ainda;
2. Considere-se um objeto *j* não selecionado e calcule a diferença entre a sua dissimilaridade em relação ao último objeto selecionado (D_j) com a dissimilaridade do objeto *i* selecionado no passo anterior ($d(j,i)$);
3. Se a diferença for positiva, o objeto *j* irá contribuir com a decisão de se selecionar o objeto *i*. Calcule:
   $$C_{ji} = \max\left(D_j - d(j,i), 0\right) \qquad (3.38)$$
4. Calcule o total obtido por selecionar o objeto *i*:
   $$\sum_j C_{ji} \qquad (3.39)$$
5. É selecionado o objeto i que maximize a expressão 3.39.

- *Troca*

  Nesta fase tenta-se melhorar o conjunto de medóides através da troca de objetos entre eles. Dessa forma, faz-se a troca entre objetos pertencentes a agrupamentos diferentes e computa-se o novo valor da função objetivo. Quando econômica, ou seja, quando a troca feita se mostra proveitosa, diminuindo o valor da função objetivo, ela é mantida; caso contrário ela é desfeita.

A medida do resultado final do agrupamento de dados é definida como *Distância Média Final*, e é definida como (WinIDAMS, 2005a):

$$\frac{1}{N}\sum_{i=1}^{N} d\left(i, m(i)\right) \qquad (3.40)$$

onde *N* é o total de dados, *i* é o objeto do conjunto de dados, *m(i)* é o medóide mais próximo do objeto *i* e *d(i,m(i))* é a dissimilaridade entre *i* e *m(i)*.

No método *k-means*, o centro de cada agrupamento é definido como a média dos objetos pertencentes ao agrupamento. Dessa forma, o *k-means* necessita de todos os dados quando se quer definir o centro do agrupamento, enquanto o método *PAM* necessita apenas do objeto que define o medóide.

O *k-means* pressupõe uma distribuição esférica dos dados, já que ele visa minimizar a soma quadrática das distâncias euclidianas entre os objetos. O método *PAM* é mais robusto porque minimiza a soma das dissimilaridades e não depende de uma suposição inicial para os centros dos agrupamentos, como acontece com o *k-means* (Struyf, 1996).

3.2.2.8. Clustering Large Applications (CLARA)

Esse método é uma adaptação do método *PAM* para um conjunto grande de dados (Júnior, 2002).

O método *PAM* trabalha com uma matriz de dissimilaridades contendo todos os *n* dados, consumindo, dessa forma, muito espaço de memória para um conjunto grande de dados (WinIDAMS, 2005b). Por esse motivo se torna impraticável o uso desse método. Já o método *CLARA* não trabalha com toda a matriz de dissimilaridades de uma só vez. Ele trabalha com subconjuntos de tamanhos previamente definidos por vez.

O método consiste em usar o método *PAM* para um subconjunto dos dados. A seguir, cada objeto não pertencente ao subconjunto é assinalado o agrupamento com o medóide mais próximo.

A qualidade dos agrupamentos gerados nesse passo é definida pela distância média dos objetos do agrupamento ao medóide.

Todo esse procedimento é repetido para outros subconjuntos dos dados (geralmente 5) e são selecionados os agrupamentos gerados pelo subconjunto com a menor distância média.

Os cálculos são os mesmos usados no método *PAM*.