2
Processo de Agrupamentos

A análise de agrupamentos pode ser definida como o processo de determinação de k grupos em um conjunto de dados. Para entender o que isso significa, observe-se a Figura 1.

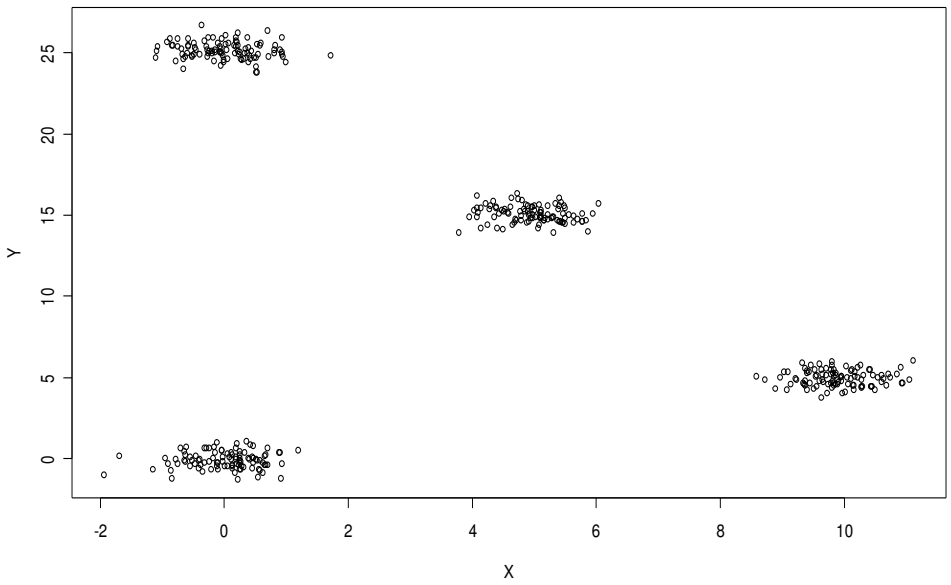

Figura 1:Gráfico ilustrativo de dados agrupados em quatro grupos.

Pode-se observar na distribuição acima a presença de quatro grupos distintos. O processo automático de descobrimento desses grupos é o principal objetivo da análise de agrupamentos, onde basicamente se buscam grupos de objetos (dados) similares entre si.

O processo de agrupamento de dados pode ser divido em três etapas que serão abordados nas seções subseqüentes:

- Seleção e tratamento de dados
- Agrupamento de dados
- Análise dos resultados

2.1.
Seleção e Tratamento de dados

Na seleção de dados o objetivo é extrair do total de variáveis da base de dados apenas os atributos que possuam maior relevância ao processo de agrupamento de dados, eliminando atributos irrelevantes ou redundantes. Esse passo é importante para diminuir o tempo de processamento do processo de agrupamento de dados como também para evitar que ele seja prejudicado por atributos irrelevantes.

No tratamento dos dados o objetivo é preparar esses dados de modo a assegurar sua qualidade e eficiência no processo de agrupamento. Os itens mais importantes para o tratamento dos dados são (Doni, 2004):

- *Eliminação de dados duplicados ou corrompidos* – dados duplicados ou corrompidos são removidos.
- *Tratamento de outliers* – dados com valores inválidos significantemente fora do esperado para uma variável são removidos.
- *Valores faltantes ou inválidos* – dados faltando valor ou com valores inválidos são removidos do conjunto selecionado.
- *Transformação dos dados* – essa etapa pode ser subdividida em duas tarefas:
  - *Tratamento de Atributos* - adequar os diferentes tipos de atributos para o processo de agrupamento. Essa tarefa é descrita em mais detalhes no item *Tratamento de Atributos*.
  - *Normalização* - tratar dados com atributos de diferentes dimensões, quando se pretende que eles tenham a mesma influência no processo. Essa tarefa é descrita em mais detalhes no item *Normalização*.

2.1.1.
Tratamento de Atributos

O primeiro objetivo dessa etapa consiste em transformar os dados de maneira que seja possível realizar o agrupamento de dados de forma adequada.

Uma base de dados pode conter dados numéricos ou categóricos sendo necessário saber lidar adequadamente com cada um destes casos.

Em (Novaes, 2002) o autor divide os tipos de atributos em 6 classes. Entretanto, para um problema real de agrupamento de dados podem ser considerados apenas duas grandes classes de atributos, a saber:

- *Atributos Quantitativos*

  Expressam numericamente a medida de uma dada variável. Estes podem ser de dois tipos:

  - Contínuos: atributos assumem valores reais. Exemplo: salário, renda familiar, altura, etc.
  - Discretas: atributos assumem valores inteiros. Exemplo: idade, número de empregados, cpf, etc.

- *Atributos Categóricos*

  São variáveis não numéricas de valores finitos. Podem assumir dois tipos:

  - Binários: possuem apenas dois tipos de valor. Exemplo: sexo (masculino ou feminino), hipertensão (sim ou não), fumante (sim ou não), etc.
  - Nominais: possuem mais do que dois tipos de valor. Exemplo: escolaridade (analfabeto, primeiro grau, segundo grau ou terceiro grau), cor de pele (branca, preta, outra), situação conjugal (casado, solteiro, separado ou viúvo), etc.

Os atributos categóricos necessitam de uma representação numérica para que o algoritmo de agrupamento de dados consiga operar sobre os seus valores. Dessa forma a variável sexo poderia assumir os valores 1 e 2 para representar respectivamente os valores masculino e feminino.

2.1.2. Normalização dos Atributos

A normalização dos dados é importante para garantir que, ao se efetuar o agrupamento, cada variável tenha o mesmo peso, exercendo a mesma influência

na execução do algoritmo. Essa influência acontece predominantemente ao se calcular as medidas de semelhança ou dessemelhança entre os dados, conhecida como medidas de proximidades que serão descritas em mais detalhes na próxima seção. Sem a normalização, as variáveis com maior escala se tornam dominantes.

Supondo que a medida de proximidade adotada seja a distância euclidiana e o número de variáveis seja igual a dois, a distância entre dois pontos será dada pela fórmula abaixo (Cook, 2004):

$$DE = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2} \qquad (2.1)$$

onde x_i e μ_i são respectivamente o valor do ponto e média da i-ésima variável.

Caso x_1 tenha uma ordem de grandeza muito maior que x_2, a distância será dominada pela primeira variável.

Para evitar esse problema, é aconselhável que se normalize os dados.

As técnicas mais utilizadas são (Reed, 1999):

- z-score

$$x_i^{'} = \frac{\left(x_i - \bar{x}\right)}{\sigma} \qquad (2.2)$$

onde x_i, $x_i^{'}$ são, respectivamente, o valor e o valor normalizado do i-ésimo dado, e $\bar{x}$ e σ são, respectivamente, a média e o desvio padrão dos dados para cada atributo.

Nesse método os dados normalizados possuem média igual a 0 e variância igual a 1. Esse método é útil quando se trabalha com uma distribuição de dados com médias diferentes ou variâncias diferentes.

- min-max cutoff

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (2.3)$$

onde x_i, $x_i^{'}$ são, respectivamente, o valor e o valor normalizado do i-ésimo dado, e x_{min} e x_{max} são, respectivamente, o valor mínimo e o valor máximo dos dados para cada atributo.

Nesse método os dados normalizados estão distribuídos linearmente entre 0 e 1. Esse método é mais simples e útil quando se trabalha com um conjunto de dados com uma distribuição linear.

2.2. Agrupamento de dados

A segunda etapa do processo de análise de agrupamentos é o agrupamento de dados. Esta etapa tem como objetivo dividir um determinado conjunto de dados em grupos com características similares entre si.

Os algoritmos de agrupamento de dados podem ser classificados em duas grandes classes de métodos (Costa, 1999):

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

Os métodos de agrupamento de dados são abordados em detalhes no capítulo 4. Em geral esses algoritmos buscam dados similares entre si através de uma medida de proximidade, conforme visto a seguir.

2.2.1. Medidas de Proximidade

A medida de proximidade pode ser definida como a medida de Similaridade ou Dissimilaridade entre os dados (Koerich, 2005), abordados com mais detalhes nas próximas seções.

Em (Viana, 2004), a Matriz de Similaridades é uma matriz de dimensão n x n contendo as medidas de similaridades/dissimilaridades entre os n objetos. Essa matriz é bastante utilizada em diversos algoritmos de agrupamento de dados.

2.2.1.1. Dissimilaridade

Dissimilaridade é a medida de diferença entre dois objetos. Existem várias maneiras possíveis de se obter essa medida. Métricas muito conhecidas, como a distância Euclidiana e a distância de Manhattan, podem e são utilizadas, mas medidas de dissimilaridade baseadas no Coeficiente de Correlação de Pearson são muito úteis quando o objetivo é o agrupamento de dados, pois ele mede o nível de relacionamento entre duas variáveis (Kauffman, 1990).

O Coeficiente de Correlação de Pearson *R(x, y)* é dado por:

$$R(x,y) = \frac{\sum_{i=1}^{p}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{p}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{p}(y_i - \mu_y)^2}} \qquad (2.4)$$

onde x e y e x_i e y_i são, respectivamente, os valores dos i-ésimos atributos dos dados x e y, e μ_x e μ_y são, respectivamente, os valores de média dos dados x e y.

Os valores do coeficiente de correlação de Pearson estão no intervalo $-1 \leq R \leq 1$. Quanto mais próximo de 1, mais correlacionados estão os dados, da mesma forma que, quanto mais próximos de -1, menos correlacionados estão os dados.

As Medidas de Dissimilaridade baseadas no Coeficiente de Correlação de Pearson são (Kauffman, 1990):

$$d(x,y) = \frac{(1 - R(x,y))}{2} \qquad (2.5)$$

$$d(x,y) = 1 - |R(x,y)| \qquad (2.6)$$

onde *d(x, y)* é a medida de dissimilaridade entre os dados x e y.

Com a expressão 2.5, variáveis com uma correlação positiva alta terão um coeficiente de dissimilaridade perto de zero, enquanto que variáveis com uma correlação negativa forte serão consideradas muito dissimilares. Em outras aplicações pode ser preferível usar a expressão 2.6 (Kauffman, 1990), onde variáveis com uma correlação negativa ou positiva alta receberão um coeficiente de dissimilaridade perto de zero, e serão considerados muito dissimilares quando o coeficiente de correlação for próximo de zero.

A escolha de uma das duas expressões depende muito do problema e do entendimento que se tem sobre a base de dados. Pode ser conveniente usar a

expressão 2.6 quando se deseja que variáveis muito correlacionadas ou pouco correlacionadas sejam similares. Já na expressão 2.5, a relação entre a correlação e a dissimilaridade é linear.

Comparações entre as duas equações sobre dados reais mostraram que a equação 2.5 apresentou resultados bem melhores, embora a equação 2.6 tenha apresentado resultados relativamente bons (Kauffman, 1990).

2.2.1.2. Similaridade

Similaridade é a medida de igualdade entre dois objetos. Tipicamente a similaridade *s* entre dois objetos *x* e *y* assume valores entre 0 e 1, onde 0 expressa que os dois objetos não são similares, enquanto que 1 expressa máxima similaridade. Geralmente são consideradas as seguintes condições para se definir similaridade:

- $0 \leq s(x, y) \leq 1$
- $s(x, x) = 1$
- $s(x, y) = s(y, x)$

Como as medidas de similaridades não podem ser calculadas diretamente através do coeficiente de correlação de Pearson, é necessário efetuar algumas transformações a fim de se respeitar as condições de similaridade (Kauffman, 1990). Existem essencialmente duas maneiras para isso, dependendo do significado dos dados e do propósito da aplicação. Supondo que variáveis com uma correlação negativa forte são variáveis muito diferentes, considere-se a expressão de similaridade abaixo.

$$s(x, y) = \frac{1 + R(x, y)}{2} \qquad (2.7)$$

onde *x* e *y* são objetos em um conjunto de dados, e *s(x, y)* e *R(x, y)* são, respectivamente, a similaridade e o coeficiente de correlação entre *x* e *y*.

A expressão 2.7 indica que sempre que o coeficiente de correlação for próximo de -1, a similaridade será próxima de 0, enquanto que valores de correlação próximos de 1 representam valores de similaridade próximos de 1.

De acordo com Kauffman & Rousseeuw (1990), existem situações onde variáveis com uma correlação positiva ou negativa forte representam

essencialmente o mesmo significado. Para esses casos pode ser usada a expressão abaixo:

$$s(x, y) = |R(x, y)| \qquad (2.8)$$

onde *x* e *y* são objetos em um conjunto de dados, *s(x, y)* e *R(x, y)* são, respectivamente, a similaridade e o coeficiente de correlação entre *x* e *y*.

A expressão 2.8 indica que para coeficiente de correlação próximo de 1 ou -1, a similaridade será próxima de 1, enquanto que valores de correlação próximos de 0 representam objetos pouco similares, com valores de similaridade próximos de 0.

2.2.2. Métricas Comuns em Medidas de Proximidade

As métricas mais comuns e utilizadas na prática são a distância Euclidiana, distância de Mahalanobis e distância de Manhattan (Novaes, 2002; WinIDAMS, 2005a). A Figura 2 mostra as superfícies observadas por cada uma das métricas aqui abordadas.

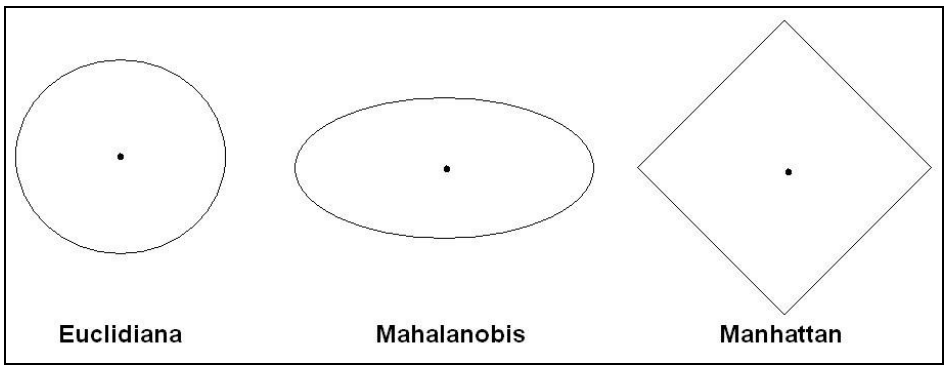

Figura 2: Superfícies observadas pelas distâncias Euclidiana, Mahalanobis e Manhattan.

A seguir são apresentadas as principais métricas de distância *d* entre os dados *x* e *y*.

- *Distância Euclidiana*

  É uma medida invariante a translações, porém assume covariâncias iguais entre as classes e em geral não é invariante a transformações lineares (Costa, 1999). É a métrica mais utilizada na prática (Almeida, 2004).

$$d(x, y) = \|x - y\| = \sqrt{(x - y)^T (x - y)} \qquad (2.9)$$

- *Distância de Manhattan ou 'city-block'*

  A distância de Manhattan é uma simplificação da distância Euclidiana, e, por isso, é uma medida mais simples e de fácil implementação. É mais eficiente para aplicações em tempo real devido a sua simplicidade (Kugler, 2003).

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \qquad (2.10)$$

  onde x_i e y_i são, respectivamente, os valores do i-ésimo atributo para os dados x e y.

- *Distância de Mahalanobis*

  Essa métrica considera que as superfícies de cada classe são elipsóides centradas na média. No caso especial em que a covariância é zero e a variância é a mesma para todas as variáveis, as superfícies serão esferas, e a distância de Mahalanobis fica equivalente à distância Euclidiana.

  Essa métrica supre muita das limitações da distância Euclidiana, porém pode ser bastante difícil determinar precisamente as matrizes de covariância, e o custo computacional cresce muito com o número de variáveis envolvidas (Costa, 1999).

  Por estas razões, em geral prefere-se usar a distância Euclidiana.

$$d(x^T, y^T) = \sqrt{(x - y) S^{-1} (x - y)^T} \qquad (2.11)$$

  onde S é a matriz de covariância dos dados que pode ser definida como (Costa, 1999):

$$S = matriz(s_{kj}) = \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \overline{x_k})(x_{ij} - \overline{x_j}) \qquad (2.12)$$

  onde $k, j = 1 \ldots p$, p é o número de atributos, x_{ik} e x_{ij} são, respectivamente, os valores do dado da linha i e coluna k e j, e $\overline{x_k}$ e

$\overline{x_j}$ são, respectivamente, as médias correspondentes dos dados ao longo das colunas *k* e *j*.

2.3. Análise dos resultados

A análise dos resultados compreende primeiramente a avaliação da qualidade dos agrupamentos, o que pode ser observado através do gráfico da silhueta, explicado no próximo item. É importante ressaltar que alguns métodos possuem métricas específicas para cálculo da qualidade do agrupamento.

O passo seguinte é a compreensão e interpretação dos agrupamentos gerados, a fim de inferir regras ou características que expliquem cada grupo.

2.3.1. Gráfico da Silhueta

O gráfico da Silhueta mede a qualidade de um agrupamento (Kauffman, 1990). Sendo *A* o agrupamento ao qual o objeto *i* pertence, a dissimilaridade média do objeto *i* em relação a todos os outros objetos de *A* é dada por:

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \qquad (2.13)$$

onde |*A*| representa o total de objetos presentes no agrupamento e *d(i,j)* representa a dissimilaridade entre os objetos *i* e *j*.

Considere-se agora qualquer agrupamento *C* diferente de *A*. A dissimilaridade média do objeto *i* para todos os objetos de *C* será dada por:

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j) \qquad (2.14)$$

onde |*C*| representa o total de objetos presentes no agrupamento *C* e *d(i,j)* representa a dissimilaridade entre os objetos *i* e *j*.

A menor distância de dissimilaridade entre o objeto *i* a um dado agrupamento *A* será dado por:

$$b(i) = \min_{C \neq A} d(i, C) \qquad (2.15)$$

Considere-se como *B* o agrupamento *C* que contém a menor distância dada acima. Esse agrupamento é chamado de vizinho do objeto i e é o segundo melhor agrupamento para esse objeto.

O valor de silhueta do objeto i é definido como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (2.16)$$

O valor de s(i) está entre -1 e 1 e pode ser interpretado da seguinte forma:

s(i) ≈ 1 $\Rightarrow$ objeto *i* bem classificado no agrupamento *A*.

s(i) ≈ 0 $\Rightarrow$ objeto *i* está entre os agrupamentos *A* e *B*.

s(i) ≈ -1 $\Rightarrow$ objeto *i* mal classificado no agrupamento *A*. Está mais perto do agrupamento *B* do que do *A*.

O gráfico da silhueta do agrupamento *A* é dado pelo gráfico da silhueta de todos os objetos pertencentes ao agrupamento *A* em ordem decrescente. Quanto mais próximo de 1, melhor é a qualidade desse agrupamento.

Os valores da silhueta podem ser interpretados como se segue na tabela abaixo:

Tabela 1: Valores da Silhueta.

s(i)	**Descrição**
0,71 – 1,00	Uma estrutura forte foi encontrada.
0,51 – 0,70	Uma estrutura razoável for encontrada.
0,26 – 0,50	A estrutura é fraca e pode ser superficial. É aconselhável o uso de outros métodos para esses dados.
≤ 0,25	Nenhuma estrutura substancial foi encontrada.

A Figura 3 mostra um exemplo de um gráfico da silhueta para dados agrupados em 4 grupos.

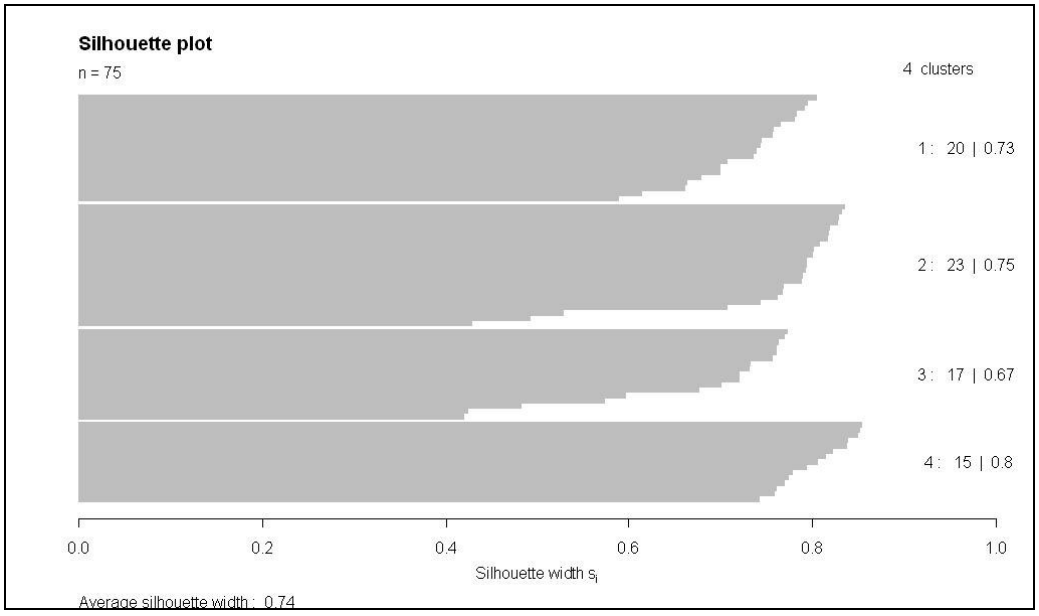

Figura 3: Gráfico da Silhueta.

O eixo vertical é o eixo que representa os n dados enquanto que o eixo horizontal é o eixo dos valores da silhueta. O gráfico representa os valores de silhueta dos dados, onde os dados estão divididos por agrupamento e ordenados em ordem decrescente de valor de silhueta. Dessa forma, pode-se observar 4 agrupamentos e os seguintes valores:

- Total de dados: 75
- Agrupamento 1: 20 dados / valor médio de silhueta: 0,73
- Agrupamento 2: 23 dados / valor médio de silhueta: 0,75
- Agrupamento 3: 17 dados / valor médio de silhueta: 0,67
- Agrupamento 4: 15 dados / valor médio de silhueta: 0,8

De acordo com a tabela 1, todos os quatro agrupamentos apresentam uma boa estrutura.