**Análise de Agrupamentos Utilizando Métodos Estatísticos**

**Universidade Tecnológica Federal do Paraná - UTFPR**

**Ana Flávia Yanaze Muranobu**

anamur@alunos.utfpr.edu.br

**Eduardo Vanderlei dos Santos Junior**

eduardojunior@alunos.utfpr.edu.br

**Lucas Folmann de Lima**

lucasl@alunos.utfpr.edu.br

Trabalho apresentado à disciplina de Probabilidade e Estatística, professor Luiz Ledo, no primeiro semestre de 2018, como parte da nota para aprovação na disciplina.

Curitiba

2018

**Introdução**

A Análise de Agrupamento é uma técnica utilizada com base na similaridade ou distância dos elementos que, diferentemente da classificação, temos um número de grupos conhecidos e algum(ns) critério(s) mais observatório sobre esses grupos são realizadas. Este artigo tem como objetivo agrupar um conjunto de 31 elementoscom base em sua similaridade e distância, a partir de uma amostra da população que deseja-se estudar, retiramos os dados e com eles, após processamento explicado posteriormente neste mesmo artigo, resultados são gerados.

O conjunto de dados a ser estudado é uma amostra de 31 árvores. Este conjunto de dados possui resumidamente as seguintes características:

|  |  |  |  |
| --- | --- | --- | --- |
| **Dados Iniciais** | Circunferência | Altura | Volume |
| Mínimo | 8.30 | 63 | 10.20 |
| 1 Quadrante | 11.05 | 72 | 19.40 |
| Mediana | 12.90 | 76 | 24.20 |
| Média | 13.25 | 76 | 30.17 |
| 3 Quadrante | 15.25 | 80 | 37.30 |
| Máximo | 20.60 | 87 | 77.00 |
| Variância | 9.847914 | 40.6 | 270.2028 |

**Objetivo**

O estudo relatado neste texto tem como objetivo agrupar **31 elementos** através de um processamento de dados usando como critério a similaridade / distância de um elemento com outro.

**Método - Fundamentação Teórica**

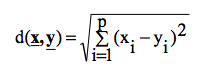
**Medidas de Similaridade e Dissimilaridade**

Geralmente, quando elementos como os de nosso estudo são agrupados, a proximidade é indicada muitas vezes com base no **coeficiente de correlação**.[1] **Similaridade**: quanto maior o valor observado mais parecidos são os objetos. Ex.: o coeficiente de correlação.

**Dissimilaridade**: quanto maior o valor observado menos parecidos (mais dissimilares) serão os objetos. Ex.: distância Euclidiana.

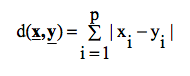
Alguns exemplos de distância podem ser verificados e exemplificados abaixo:

1. Distânicia Euclidiana:  
   É a distância entre dois pontos, que pode ser provada pela aplicação repetida do teorema de Pitágoras. Provavelmente a mais conhecida e mais utilizada como medida de distância, ela é simplesmente a distância geométrica no espaço multidimensional.



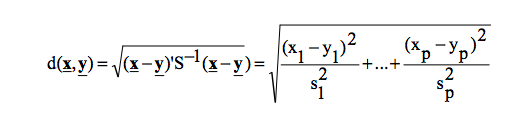
**2)** Distância de Manhattan:

É a distância entre dois pontos num espaço euclidiano com um sistema cartesiano de coordenadas fixo como a soma dos comprimentos da projecção da linha que une os pontos com os eixos das coordenadas.



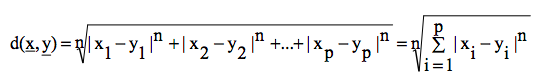
**3)** Distância de Malahnobis

A distância de Mahalanobis é uma métrica que difere da distância Euclidiana por levar em consideração a correlação entre os conjuntos de dados. Sua fórmula é a descrita abaixo:



**4)** Distância de Minkowski

É uma métrica em um vetor de espaço normal que pode ser considerado uma generalização de ambas Distâncias Euclidianas e Distância de Manhattan, mencionadas anteriormente. Sua fórmula pode ser dada da seguinte forma:



**Método de Agrupamento Hierárquico**

Resumidamente, neste método, no início existem tantos grupos quanto objetos (itens). Diversos objetos semelhantes são agrupados primeiro; Estes grupos iniciais são fundidos de acordo com as suas similaridades, eventualmente, relaxando no critério de similaridade os sub-grupos vão se unindo a outros sub-grupos até formar um grupo único.

Os passos para esse método, mais detalhadamente, são os seguintes:

1. Tem-se tanto grupos quanto elementos inicialmente, ou seja, denominaremos essa quantidade de **n.** Calcula-se a **matriz simétrica de distância** n x n, D = d(i,j) onde d(i,j) é a distância ou similaridade entre um objetivo i e j.
2. Na matriz, acha-se o par de distância entre os objetos e seleciona-se a menor das distância entre eles; esses dois elementos serão agrupados.
3. Forma-se um novo grupo com os elementos do passo acima, denominado por A e B. (grupos primitivos). Uma nova matriz de distância é construída com esses grupos e adiciona-se as linhas e colunas dadas pelas distâncias entre (AB) e os outros grupos.
4. Repete-se de 2 e 3 **n-1 vezes**, considerando os novos grupos e tendenciando-se, cada vez mais, a agrupar os elementos similares com pequenas distâncias entre si.

**Ligações**

Descreveu-se, no item anterior, o Método de Agrupamento Hierárquico com o objetivo de aglomerar elementos semelhantes. Esse método, para realizar o tal agrupamento, utiliza-se das distâncias entre os objetos, que são realizadas através de ligações. Essas ligações, geralmente, podem ser as ligações simples, completas, método das médias da distância, método de centróide ou método de Ward.

O método utilizado para o estudo aqui apresentado é a **ligação simples**, também conhecido como vizinho mais próximo. O agrupamento é feito juntando-se dois grupos com as menores distâncias entre as distâncias disponíveis. Uma vez formado o novo grupo, a distância entre (AB) e algum outro grupo C é dado por:



**Dendograma**

A cada formação de grupos, mencionado anteriormente neste texto, os grupos são formados e uma representação gráfica desse processo inteiro pode ser gerada. Essa representação pode ser denominada **dendograma**, proveniente de uma análise estatística dos elementos dispostos.

É através do dendograma que pode-se escolher uma estrutura de corte para definir quais e quantos grupos serão formados. Quanto mais acima do dendograma essa linha de corte ocorrer, menos grupos finais serão apresentados.

**Análises e Resultados**

Este tópico abordará como os procedimentos foram realizados. Para a realização da análise, primeiramente foi carregado os dados com auxílio de funções da biblioteca R. **A ferramenta R foi utilizada em todo o procedimento,** a fim de obter os grupos distintos. Uma análise gráfica também começou a ser gerada. Em seguida, foi criada a matriz de distância para os elementos carregados e o Agrupamento Hierárquico com a função hclust, utilizando os dados e, como critério de ligação entre os elementos, a **ligação simples** previamente explicada foi utilizada.

Depois disso, foi gerado um Dendograma e, em seguida, escolhido três grupos (o corte da árvore)

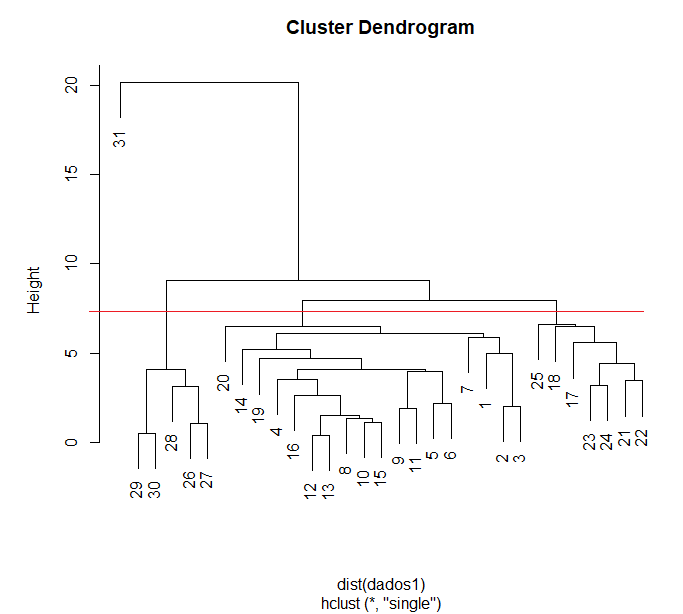


Figura 1 - Dendograma

Conforme demonstrado na figura 1, após a formação dos grupos, foi feito um corte no Height de valor aproximadamente 7 (demonstrado pela linha vermelha da figura), separando os dados em 3 grupos. Após realizar o corte, os resultados obtidos para as variáveis circunferência, altura e volume de cada grupo foram:

**No grupo 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Grupo 1** | Circunferência | Altura | Volume |
| Mínimo | 8.30 | 63.00 | 10.20 |
| 1 Quadrante | 10.72 | 69.25 | 16.85 |
| Mediana | 11.15 | 74.50 | 19.80 |
| Média | 11.12 | 73.00 | 18.99 |
| 3 Quadrante | 11.62 | 76.00 | 22.00 |
| Máximo | 13.80 | 83.00 | 25.70 |

**No grupo 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Grupo 2** | Circunferência | Altura | Volume |
| Mínimo | 12.90 | 72.00 | 27.40 |
| 1 Quadrante | 13.65 | 75.50 | 32.75 |
| Mediana | 14.20 | 78.00 | 34.50 |
| Média | 14.46 | 78.86 | 34.94 |
| 3 Quadrante | 15.25 | 82.50 | 37.30 |
| Máximo | 16.30 | 86.00 | 42.60 |

**No grupo 3**

|  |  |  |  |
| --- | --- | --- | --- |
| **Grupo 3** | Circunferência | Altura | Volume |
| Mínimo | 17.30 | 80.00 | 51.00 |
| 1 Quadrante | 17.60 | 80.00 | 52.48 |
| Mediana | 17.95 | 80.50 | 55.55 |
| Média | 18.22 | 81.67 | 58.15 |
| 3 Quadrante | 18.00 | 81.75 | 57.65 |
| Máximo | 20.60 | 87.00 | 77.00 |

Observando os resultados destes três grupos, é possível perceber que as árvores da amostra foram separadas em pequenas (grupo 1), médias (grupo 2) e grandes (grupo 3).

**Conclusão**

Este estudo explicou como é realizada as agrupações através do método de agrupamento hierárquico e mostrou a eficiência deste, pois a similaridade dos elementos agrupados é muito maior do que a similaridade entre dois elementos que não fazem parte do mesmo grupo, visando a ligação que foi utilizada, independente do método de distância adotado.

Assim, pôde-se observar que como em muitos casos no mundo real, em diversas ocasiões classificatórias é utilizado três “divisores” com características semelhantes entre si para se distinguirem dos outros dois; como por exemplo uma peça de roupa, ela pode ser Pequena e atender à certas atribuições em intervalos de determinadas medidas. O método é muito útil na realidade contemporânea em várias áreas.

Também foi realizado um *overview* (uma visão geral) dos métodos de distância e comprovada a eficiência da ferramenta R para realização de análises estatísticas, visto que a mesma contém inúmeras funções já implementadas e algoritmos prontos para auxiliar a execução de procedimentos estatísticos, tais como o *hclust* ou *plot*.

**Referência Bibliográfica**

[1] MARQUES, Jair Mendes: Análise Multivariada Aplicada à Pesquisa;