EletroRevista
Revista Científica e Tecnológica
Institutional Business Consultoria Internacional
ISSN Nº: 1983-2168

TÉCNICAS DE AGRUPAMENTO CLUSTERING

Autores: **Istvan Karoly Kasznar, *PhD***
Professor Titular da FGV e Presidente da IBCI

Bento Mario Lages Gonçalves, *MSc*
Consultor Senior da IBCI

CLUSTERING

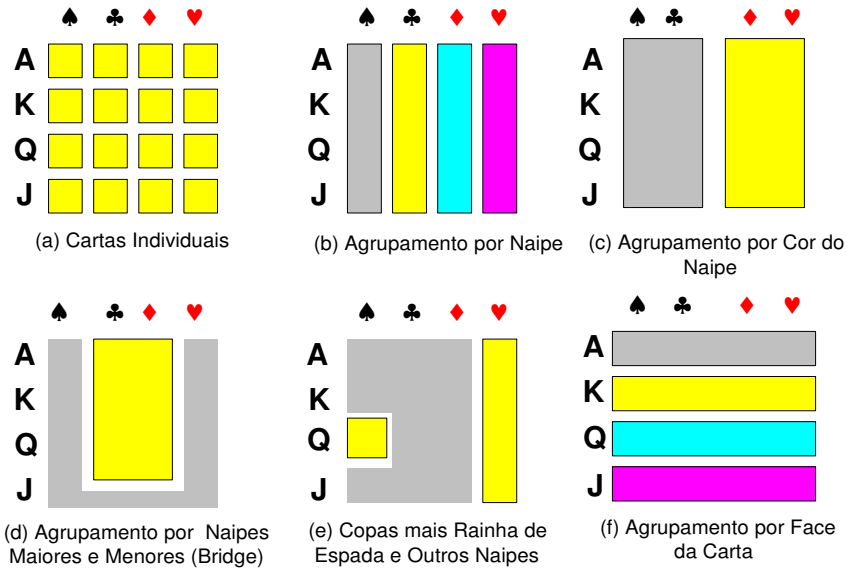
1- Introdução

Os procedimentos explanatórios são de grande ajuda na compreensão da natureza complexa das relações multivariadas. Conforme enfatizamos anteriormente a plotagem dos dados (observações) é um instrumento que permite melhor visualização do universo amostral com o objetivo do desenvolvimento de metodologia de classificação. Neste Apêndice serão discutidas técnicas de plotagem dos dados e metodologias *step by step* (passo a passo) ou algorítmos para o agrupamento de objetos (variáveis ou ítens). A busca de dados para a estruturação de agrupamentos "naturais" é uma técnica explanatória importante. Agrupamentos podem prover meios informacionais para avaliar a dimensionalidade, identificar exclusões grupais e sugerir hipóteses referentes ao interrelacionamento das variáveis grupais.

O agrupamento, ou *clustering*, difere das metodologias de classificação previamente discutidas como a análise discriminante múltipla e a análise canônica. A classificação é pertinente a um número conhecido de grupos e seu objetivo operacional é "enquadrar" novas observações a um destes grupos. A análise de Cluster é uma técnica mais primitiva uma vez que nenhum pressuposto é assumido no que tange ao número de grupos ou a sua estruturação. O agrupamento é realizado a partir de similaridades ou distâncias entre seus componentes (dissimilaridades). Os únicos pré-requisitos são medidas de similaridade ou dados sob os quais possam ser calculadas similaridades.

Para ilustrar a natureza da dificuldade na definição de grupos naturais, vamos considerar a ordenação de 16 cartas figuradas de um baralho convencional em clusters ou objetos similares. Alguns agrupamentos são realizados na *Figura I* a seguir. Fica bastante claro que partições significativas dependem da definição de similaridade.

Figura I – Agrupamentos de Cartas Figuradas



Na maioria das aplicações práticas da análise de cluster o pesquisador tem conhecimento suficiente para distinguir "bons" agrupamentos de "maus" agrupamentos. Por que não enumerar todas as possibilidades de agrupamento e selecionar as "melhores" para estudo posterior ?

Para o exemplo das cartas do baralho, existe uma maneira de formar um *único* grupo de 16 cartas figuradas; existem *32.767* maneiras de particionar as cartas figuradas em *dois* grupos (de tamanhos variados); existem *7.141.686* maneiras de ordenar as cartas figuradas em *três* grupos (de tamanhos variados), e assim por diante [1]. Evidentemente as limitações de tempo tornam possível a determinação dos melhores agrupamentos de objetos similares a partir de uma lista com todas as estruturas possíveis. A evolução da capacidade de processamento dos computadores vem permitindo a manipulação de um número cada vez maior de casos (variáveis), de tal forma que algorítmos vem sendo desenvolvidos na busca de uma boa, talvez não a melhor, forma de agrupamento.

(1) O número de maneiras de ordenação de n objetos em k grupos não fechados é um número Stirling(Ver (1) Bibliografia) do segundo grau dado por :

$$(1/k!)\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{n}$$

Adicionando tais números para k = 1,2,..., n grupos, obtemos o número total de possibilidades de ordenação de n objetos em grupos.

Em suma, o objetivo básico da análise de cluster é descobrir grupamentos naturais dos ítens (ou variáveis). Desse modo, devemos primeiramente desenvolver uma escala quantitativa de maneira a medir a associação (similaridade) entre os objetos. A seção a seguir é dedicada a discussão das medidas de similaridade. Nas seções seguintes são discutidos os algorítmos mais comuns utilizados na ordenação de objetos em grupos.

2 – Medidas de Similaridade

A maioria dos esforços dispendidos na produção de uma estrutura grupal simples a partir de um conjunto de dados complexos requer medidas de "proximidade" ou "similaridade". Existe sempre um elevado grau de subjetividade no que tange a escolha de uma medida de similaridade. Considerações importantes como a natureza das variáveis (discreta, contínua, binária), as escalas de medida (nominal, ordinal, intervalo, quosciente) e o conhecimento específico do assunto em tela; devem ser ativadas.

Quando *itens* (unidades ou casos) são clusterizados, sua proximidade é indicada por algum tipo de distância. Por outro lado, as variáveis são agrupadas baseadas no seu coeficiente de correlação ou outras medidas estatísticas de associação.

2.1 – Distâncias e Coeficientes de Similaridade para Pares de Ítens

A noção de distância advém da discussão relativa às medidas de dispersão estatística. Relembrando a distância Euclideana (linha reta) entre duas observações *p*-dimensionais $X = [x_1, x_2, ..., x_p]$ e $Y = [y_1, y_2, ..., y_p]$ é dada por :

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} = \sqrt{(x-y)/(x-y)}$$

A distância estatística entre as mesmas observações é dada por :

$$d(x,y) = \sqrt{(x-y)/A(x-y)}$$

Onde $A = S^{-1}$, e S contém as variâncias simples e as covariâncias. Contudo sem o conhecimento prévio dos grupos distintos, estas quantidades amostrais não podem ser computadas. Por esta razão a distância Euclideana é frequentemente preferida na análise de cluster.

Uma outra medida de distância é a métrica de Minkowsky, que é dada por :

$$d(x,y) = \left[\sum_{i=1}^{p} (x_i - y_i)^m \right]^{1/m}$$

Para *m=1*, *d(x,y)* mede a distância em bloco de dois pontos em *p* dimensões. Para *m=2*, *d(x,y)* se torna a distância Euclideana. De uma maneira geral, a variação de *m* determina o peso dado para grandes e pequenas diferenças de distância.

Bibliografia

Arley, N. e Buch, K. R.; Introduction to then theory of probability and Statistics; Wiley and Sons Publishers; New York; USA; 1.950.

Cramér, Harald; Random Variables and probability Distributions; Cambridge University Press; Cambridge; 1.937.

Doob. J. L.; Stochastic Processes; Wiley and Sons Publishers; New York; USA; 1.953.

Cramér, Harald; Elementos da Teoria da Probabilidade e Algumas de suas aplicações; Editora Mestre Jou; São Paulo; SP; 1.973.

Gnanadesikan, R.; Methods for Statistical Data Analysis of Multivariate Observations; John Wiley; New York; USA; 1.977.

Haavelmo, T.; The Statistical Implications of a System of Simultaneous Equations; Econometrica; volume 11; january, 1.943.

Kendall, M. G.; The advanced Theory of Statistics; volumes I and II; London; Griffin; 1.959.

Lévy, P.; Théorie de l'addition des variables aléatoires; Paris, Gauthier, Vilars; France; 1.977.

Lipschutz, Seymour; Probabilidade; Coleção Schaum; Editora McGraw – Hill do Brasil; 2@ edição revisada; São Paulo; SP; 1.974.

Spiegel, Murray R.; Estatística; Coleção Schaum; Editora McGraw – Hill do Brasil; São Paulo, SP; 1.976.

Wonnacott, Ronald, J.; Wonnacott, Thomas, H.; Econometria; Livros Técnicos e Científicos Editora; Rio de Janeiro; RJ; 1.978.