Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Mestrado em Estatística

Métodos de agrupamento na análise de dados de expressão gênica

Fabiene Silva Rodrigues

São Carlos - São Paulo
24 de fevereiro de 2009

Universidade Federal de São Carlos

Centro de ciências exatas e de tecnologia

Departamento de Estatística

Mestrado em Estatística

Métodos de agrupamento na análise de dados de expressão gênica

Fabiene Silva Rodrigues

Dissertação submetida ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade Federal de São Carlos, como parte dos requisitos necessários para obtenção do grau de Mestre em Estatística.

Orientador: Prof. Dr. Luís A. Millan

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

R696ma

Rodrigues, Fabiene Silva.
Métodos de agrupamento na análise de dados de expressão gênica / Fabiene Silva Rodrigues. -- São Carlos : UFSCar, 2009.
93 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2009.

1. Análise multivariada. 2. Método de agrupamento. 3. Fator de Bayes. 4. Modelos com misturas. I. Título.

CDD: 519.535 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luis. Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-829213351-8241 - FAX: (016) 3351-8243
13565-905 - SÃO CARLOS-SP-BRASIL
www.ufscar.br/~des ppgest@power.ufscar.br

DECLARAÇÃO

Declaramos, para os devidos fins, que Fabiene Silva Rodrigues defendeu sua Dissertação de Mestrado no dia 11/02/2008, tendo sido **aprovada.** A aluna deverá apresentar a versão final da dissertação (com as correções e sugestões da Banca, e a ficha catalográfica anexada), e a Certidão Negativa da Biblioteca Comunitária, para formação do processo de homologação e emissão do Diploma do Titulo.

Igualmente, o aluno deverá apresentar a documentação da pesquisa (rotinas, arquivos em LaTeX, resultados complementares etc.) ao seu orientador, visando facilitar a confecção de relatórios técnicos que condensarão os resultados obtidos.

Essa declaração é válida pelo período de 30 dias.

UNIVERSIDADE FEDERAL - SÃO CARLOS - PPG-Es

Prof. Dr. Francisco Louzada Neto
Coordenador – PPG-Es / UFSCar

Agradecimentos

Agradecer a todos que de forma direta ou indireta contribuíram com este trabalho, não é tarefa fácil. Quero agradecer não só aos que me ajudaram efetivamente na construção dessa Dissertação, mas aos amigos e colegas que partilharam comigo idéias, fomentaram discussões, e muitas vezes pérolas um tanto quanto poéticas ou pitorescas. Obrigada aos que tornaram as longas madrugadas de estudos mais alegres e produtivas.

Aos amigos de alojamento pelas brincadeiras e companheirismo, que de formas descontraídas ajudaram a amenizar a falta da família e das belas praias cearenses. Em especial, os amigos Tenylson, Luciano, Juliano, Roberta, Flávia e Douglas. Que sempre tentavam rasurar minhas pontinhas de desânimo ou tristeza com sorrisos, músicas, brincadeiras, filmes e longas caminhadas até o carrinho do sorvete aos domingos. Não saberia agradecer, senão oferecendo-lhes essas linhas.

Ao professor Louzada Neto pelo incentivo e apoio em momentos difíceis. Ao meu orientador professor Millan, pela boa vontade ao marcarmos reuniões aos finais de semana e pela sua contribuição na escolha dos assuntos a serem estudados e escritos.

Quero agradecer imensamente ao professores presentes em minha banca, que muito contribuíram com o meu trabalho. Agradeço a generosidade e disponibilidade com que leram e interagiram com o meu texto.

Aos meus pais, José e Eneuma, por terem me ensinado a lutar e nunca desistir de meus projetos profissionais e pessoais. Obrigada aos meus irmãos, Wagner e Fagner, pela força e apoio, e aos demais familiares e amigos de Fortaleza que mesmo de longe enviavam votos de sucesso.

A Deus pela força e perseverança que me foi concedida nesses dois anos de dedicação.

Resumo

As técnicas de agrupamento (*clustering*) vêm sendo utilizadas com freqüência na literatura para a solução de vários problemas de aplicações práticas em diversas áreas do conhecimento. O principal objetivo deste trabalho é estudar tais técnicas. Mais especificamente, estudamos os algoritmos *Self Organizing Map* (SOM), *k-means, k-medoids, Expectation-Maximization* (EM). Estes algoritmos foram aplicados a dados de expressão gênica. A análise de expressão gênica visa, entre outras possibilidades, a identificação de quais genes estão diferentemente expressos na sintetização de proteínas associados a tecidos normais e doentes.

O objetivo deste trabalho é comparar estes métodos no que se refere à eficiência dos mesmos na identificação de grupos de elementos similares, ressaltando vantagens e desvantagens de cada um. Os métodos foram testados por simulação e depois aplicamos as metodologias a um conjunto de dados reais.

Abstract

The clustering techniques have frequently been used in literature to the analyse data in several fields of application. The main objective of this work is to study such techniques. There is a large number of clustering techniques in literature. In this work we concentrate on Self Organizing Map (SOM), k-means, k-medoids and Expectation-Maximization (EM) algorithms. These algorithms are applied to gene expression data. The analisys of gene expression, among other possibilities, identifies which genes are differently expressed in synthesis of proteins associated to normal and sick tissues.

The purpose is to do a comparing of these metods, sticking out advantages and disadvantages of such. The metods were tested for simulation and after we apply them to a real data set.

Palavras chave: Self Organizing Map (SOM), k-means, k-medoids, algoritmo EM, expressão gênica, seleção de modelos e fator de Bayes.

Sumário

Resumo

As técnicas de agrupamento (*clustering*) vêm sendo utilizadas com freqüência na literatura para a solução de vários problemas de aplicações práticas em diversas áreas do conhecimento. O principal objetivo deste trabalho é estudar tais técnicas. Mais especificamente, estudamos os algoritmos *Self Organizing Map* (SOM), *k-means, k-medoids, Expectation-Maximization* (EM). Estes algoritmos foram aplicados a dados de expressão gênica. A análise de expressão gênica visa, entre outras possibilidades, a identificação de quais genes estão diferentemente expressos na sintetização de proteínas associados a tecidos normais e doentes.

O objetivo deste trabalho é comparar estes métodos no que se refere à eficiência dos mesmos na identificação de grupos de elementos similares, ressaltando vantagens e desvantagens de cada um. Os métodos foram testados por simulação e depois aplicamos as metodologias a um conjunto de dados reais.

Abstract

The clustering techniques have frequently been used in literature to the analyse data in several fields of application. The main objective of this work is to study such techniques. There is a large number of clustering techniques in literature. In this work we concentrate on Self Organizing Map (SOM), k-means, k-medoids and Expectation-Maximization (EM) algorithms. These algorithms are applied to gene expression data. The analisys of gene expression, among other possibilities, identifies which genes are differently expressed in synthesis of proteins associated to normal and sick tissues.

The purpose is to do a comparing of these metods, sticking out advantages and disadvantages of such. The metods were tested for simulation and after we apply them to a real data set.

Palavras chave: Self Organizing Map (SOM), k-means, k-medoids, algoritmo EM, expressão gênica, seleção de modelos e fator de Bayes.

1. Introdução

Um dos objetivos da Biologia Molecular é identificar o funcionamento de expressões gênicas e as causas de importantes fenômenos biológicos, como o crescimento e o ciclo celular, a diferenciação e o desenvolvimento, bem como as patologias decorrentes de falhas no funcionamento desses genes. Mais concretamente, a Biologia Molecular investiga as interações entre os diversos sistemas celulares incluindo a relação entre DNA, RNA e síntese de protéica[1]. Um objetivo tão amplo que envolve inúmeras etapas, que geralmente dependem do desenvolvimento de novas tecnologias.

Para entendermos como os genes determinam os fenômenos acima, é importante estudar as diferenças de expressões gênicas entre determinadas amostras. Os dados analisados neste trabalho são provenientes de organismos eucariontes[2], a informação genética está organizada da seguinte forma: o DNA (ácido desoxirribonucléico), localizado no núcleo da célula, contém todas as unidades de informação (genes) que determinam as características de um indivíduo. No núcleo é produzida uma cópia complementar desta informação, conhecida como mRNA (ácido ribonucléico mensageiro), etapa denominada transcrição (Lewin, 2000). O mRNA obtido é transportado do núcleo para o citoplasma depois para o ribossomo, sendo que no ribossomo ocorre a tradução da molécula de mRNA em proteína, etapa esta chamada de "tradução". Este processo é esquematizado na Figura 1. Uma maneira de se medir diferenças de expressão gênica é quantificando os mRNAs (Revista Brasileira de Zootecnia, Julho de 2007) na célula ou em tecidos para diferentes estados biológicos.

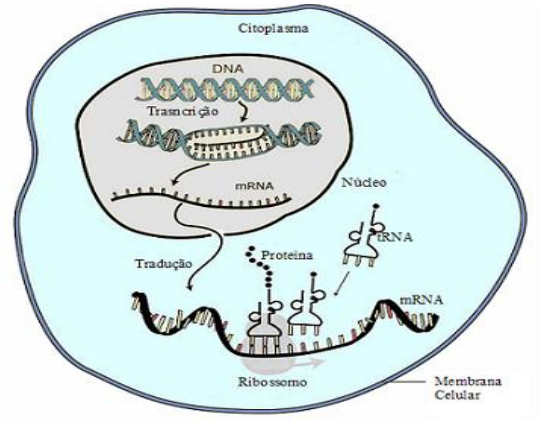


Figura 1: Processos de transcrição e tradução na célula (*Modeling Transcriptional Control in Gene Networks-Methods*, 2000).

[1]A síntese protéica é um fenômeno relativamente rápido e muito complexo, que ocorre no interior das células. Este processo tem duas fases: *transcrição* e a *tradução*.

[2]Organismos eucariontes são seres vivos formados por células eucariontes. Que por sua vez são células que possuem o núcleo individualizado, limitado pela membrana carioteca.

A possibilidade de medir de forma rápida e precisa as expressões dos genes de um conjunto de células tem também importantes aplicações na medicina como, por exemplo, em diagnóstico e desenvolvimento de drogas. Em algumas situações o médico pode não distinguir baseado nos dados histopatológicos usuais, tipos de tumores e, que resultam em diferentes evoluções e respostas a um determinado tratamento. Atualmente é possível identificar o código de algumas patologias pela análise de expressão gênica ("Differentially expressed genes in gastric tumors identified by cDNA array". Cancer Letters, 2003).

Em experimentos de expressão gênica, a idéia, em geral, é mensurar comparativamente os níveis de expressão para um determinado conjunto de genes, os quais podem ser escolhidos com base em conhecimento prévio destes. O experimento começa com a obtenção de amostras de tecidos de uma determinada região avaliada nos pacientes. A etapa seguinte é a extração de mRNA destas amostras. Para a realização deste tipo de experimento, duas das técnicas mais usadas são *Microarray* e SAGE (http://www.sagepub.com). Neste trabalho abordamos a primeira delas, que foi utilizada na geração dos dados analisados neste trabalho, conforme ilustra a Figura 2.

A pesquisa genômica[3] e, sobretudo, a pesquisa pós-genômica é, pois de fundamental importância para todas as áreas que se preocupam com os mecanismos biológicos básicos, e isso cobre campos de estudo que vão da medicina a aplicações de interesse industrial. O sucesso do Projeto genoma foi consolidado ao produzir uma seqüência final de alta qualidade do genoma do Humano. Anunciado em 2003, durante o 50º aniversário da descoberta da dupla hélice do DNA, representa o coroamento de um esforço internacional para se compreender melhor as bases moleculares da vida, e tem impulsionado inúmeros outros projetos de seqüenciamento em vários países, inclusive no Brasil.

A produção em massa de dados genômicos produz também um grande desafio: como se pode rapidamente extrair informações biologicamente relevantes das seqüências de adenina, timina, guanina e citosina, e do genoma como um todo, e transformá-la em produtos que beneficiem a humanidade? Entre as principais questões pós-genômicas estão a integração multidisciplinar necessária à quantificação, a modelagem estatística desses dados, e o reconhecimento das estruturas e inter-relações produzidas pelos genes. Como resposta a esses desafios, pesquisadores de diversas áreas (Matemática, Estatística, Engenharia, Ciência da Computação, entre outros), têm proposto metodologias para análise desse tipo de dados. Estas propostas trazem consigo a diversidade de linguagens de seus propositores. Este processo ao mesmo tempo em que enriquece a solução, dada a diversidade de propostas, traz dificuldades de identificar as melhores soluções, dado o grande número de linguagens empregadas.

Então, o objetivo deste trabalho é fazer uma revisão dos métodos de agrupamento (parte deles de fato) utilizados no agrupamento de genes, a partir da expressão de cada gene, obtidos em experimentos de microarranjos (*microarrays*) e fazer uma comparação dos mesmos. Um conjunto de dados será simulado e então aplicaremos os métodos abordados e compararemos a eficiência com que cada um classificará esses dados. Para estes dados simulados, tentaremos manter as características encontradas em dados de expressão gênica, como por exemplo, alta variabilidade.

Vale ressaltar que o desenvolvimento da pesquisa aqui relatada difere do tratamento usual dado em mestrados, onde o foco central se dá em torno de uma metodologia pré-estabelecida. Aqui o eixo central, em torno do qual gravitam as discussões e os temas abordados é o problema a ser resolvido, no caso agrupamento de genes, e por esta razão mais que uma metodologia é abordada e, em geral, o aprofundamento, em cada uma, não é possível dada a limitação de tempo.

Esta dissertação esta organizada de forma que no Capítulo 2 descrevemos a tecnologia de micro arranjos, a seguir nos Capítulos 3 e 4 tratamos dos métodos de agrupamento, no Capítulo 5 fazemos uma breve introdução sobre redes neurais e o método SOM. Nos Capítulos 6, 7 e 8, falamos de agrupamentos baseados em Modelos de Mistura, no Capítulo 9 comentamos algumas vantagens e desvantagens de cada método e por fim no Capítulo 10 é feito uma comparação, em termos de acerto no agrupamento, dos métodos abordados via simulação. Ao final da simulação elegemos um único método, os Modelos de Mistura, que na maioria das vezes apresentou melhores resultados, isso é feito no capitulo 11.

2. Tecnologia de *Microarray*

Após o seqüenciamento de um genoma, a técnica de *Microarray* tem sido uma das mais usadas na geração de conjuntos de dados referentes à expressão gênica. O *Microarray* é uma metodologia utilizada para comparar a expressão de um grande número de genes, simultaneamente. Esta técnica emprega arranjos (*array*), que contêm um grande número de genes distribuídos de forma ordenada sobre placas de vidro.

[3]Genômica: Ciência que aborda todos os genes de um organismo como um todo.

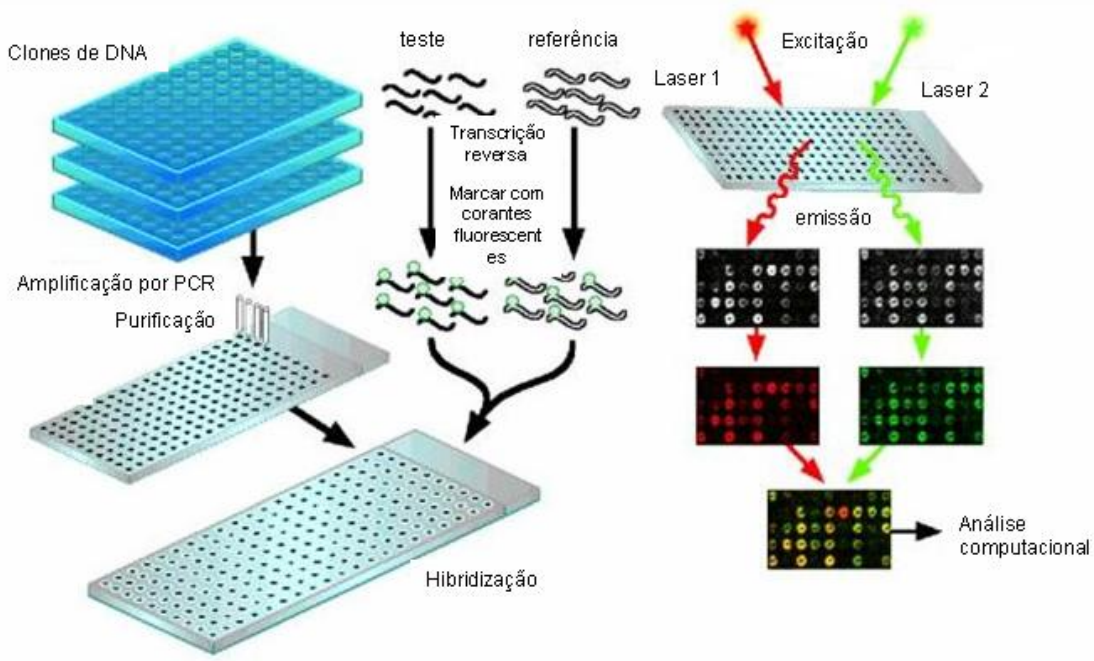


Figura 2: Etapas do processo de *Microarray* (Duggan *et al*., 1999).

A primeira etapa na preparação de um *Microarray* é a obtenção de uma lâmina de vidro, a qual é caracterizada pela sua geometria, definida basicamente pelos seguintes fatores: quantidade de *spots* e quantidade de linhas e colunas de sub-*arrays*, quantidade de linhas e colunas de *spots* por sub-*array*. Desta maneira, a tecnologia de *microarrays* consiste na utilização de um *slide* (lâmina ou microarranjo) no qual as sondas (amostras de DNA) foram imobilizadas em quantidades e posições precisamente definidas (*spots*). Além disso, para uma lâmina específica, é conhecido o cDNA (ácido desoxirribonucléico complementar, seqüência de um determinado gene) fixado em cada uma das posições por um robô que consegue a precisão necessária para separar os *spots*. Estes cDNA são selecionados de um banco de clones, previamente criado. O material fixado na lâmina é conhecido como material fixado ou *probe* (Nature Genetics Supplement Vol 21 no. 1, January 1999).

Numa lâmina são testadas duas amostras de mRNAs, sendo que uma é marcada por fluorescência com Cy3 e a outra com Cy5, as quais fluorescem nas cores verde e vermelho, respectivamente. É comum ter apenas uma amostra sendo testada por lâmina, enquanto que a segunda é uma amostra de referência, comum para todas as lâminas, sendo a amostra de referência geralmente obtida a partir de uma mistura de amostras de mRNAs. Por exemplo, em um experimento testando amostras tumorais, a referência pode ser formada por uma mistura de amostras normais. Depois de ter marcado ambas as amostras, elas são misturadas em proporções iguais, e esta mistura é espalhada pela lâmina. Os resultados são produzidos sob forma de diferentes intensidades de fluorescência, que são captados por

microscopia de fluorescência a laser, em função dos diferentes níveis de expressão de cada gene. A imagem dos pontos fluorescentes é processada por computadores e programas específicos, sendo gerada simultaneamente uma grande quantidade de informações.

Neste processo, conhecido como hibridização, ocorre um pareamento das moléculas complementares. A partir desse pareamento temos em cada um dos *spots* da lâmina, que referencia um determinado gene, as proporções de mRNAs nas duas amostras testadas. O material usado na hibridização é conhecido como material flutuante ou *target*. As proporções poderão ser usadas para estimar e comparar as diferenças nos níveis de expressão dos genes sendo analisados.

Figura 3: Resultado da digitalização da imagem de uma lâmina
(*Gene expression Informatics*, **21**: 5 1-55, 1999).

2.1 Análise Estatística de Dados de *Microarray*

No presente trabalho, focamos nas questões estatísticas após a análise da imagem. Esta etapa consiste na interpretação de toda a coleção de dados numéricos obtidos das imagens.

O primeiro passo na análise de dados de *Microarray* é a análise da imagem obtida por *scanner*. É preciso identificar os *spots* e quantificar o material genético expresso nas lâminas. Para localizar os *spots* é feita a segmentação da imagem. Inicialmente esta segmentação era feita manualmente, num processo trabalhoso e extremamente propício à introdução de erros cujo efeito é difícil de avaliar. Em seguida, a leitura do sinal de cada *spot*, deve ser feita. Nos sistemas existentes há inúmeras opções para esta quantificação, sendo que os resultados numéricos finais podem variar muito, conforme a opção. Essas diferentes opções estão associadas a diferentes suposições de cada modelagem com relação aos fenômenos bioquímicos e fontes de ruído. Por exemplo, não está claro como distinguir *pixels* informativos de não-informativos (ou seja, *pixels* correspondentes à região com *probe* ou não) e sobre como compensar a existência de ruído. Além disso, usualmente, consideramos

importante, apenas razões entre as intensidades nos dois canais, mas não está claro o que isto significa numericamente (razões entre medianas, médias, regressão linear, etc). No entanto, é fato que todas as etapas da análise matemática dos dados de *Microarray* devem ser escolhidas de forma a obter bons estimadores para os níveis de expressão gênica (“Computational analysis of expression data”, Heidelberg 1999).

Uma vez de posse de uma tabela com as quantificações das intensidades de cada gene, o passo seguinte, em boa parte dos trabalhos da literatura, é usualmente realizar a análise exploratória para extrair informações potencialmente úteis dos dados. A análise de agrupamento é uma das ferramentas disponíveis, que consiste em um método de estatística exploratória para agrupar os dados de acordo com uma medida de similaridade. A similaridade pode ser definida segundo uma medida de distância entre as amostras.

3. Agrupamento

A idéia básica de agrupamento começa quando temos vários objetos e queremos agrupá-los pelas suas características que podem ser quantitativas ou qualitativas. Os grupos são formados de modo que as distâncias entre os elementos de um grupo sejam mínimas e as distâncias entre os grupos sejam máximas. Á análise de agrupamento é uma das principais técnicas usadas para verificar as questões específicas de interesse no caso, em particular, dos conjuntos de dados de expressão gênica. Além disso, é utilizada para buscar padrões de comportamento e fazer associações de genes com base em diferentes características ou associações de amostras de pacientes em estados similares de diagnóstico.

Uma classe de métodos amplamente usada envolve agrupamentos aglomerativos hierárquicos, onde dois grupos escolhidos para otimizar algum critério, são unidos a cada estágio do algoritmo. Um critério inclui a soma de quadrados de dentro do grupo (Ward, 1963) e a distância entre grupos, que formam a base do método *single-link*. Outra classe de métodos é baseada na realocação iterativa (também chamado de particionamento iterativo), em que determinados pontos são movidos de um grupo para o outro até que não existam melhorias segundo algum critério. A realocação iterativa com o critério da soma de quadrados é chamada agrupamento de *k*-médias (MacQueen, 1967). Embora haja pesquisa considerável nesta área (por exemplo, análise de dendrograma para agrupamento hierárquico), existem poucas orientações para responder questões básicas encontradas na prática, pertinentes na análise de grupo, tal como: Quantos grupos existem? Que método de agrupamento usar? E como devemos tratar os *outliers*?. Além disso, as propriedades estatísticas desses métodos são geralmente desconhecidas, impossibilitando a aplicação de uma inferência convencional.

Neste processo existem três parâmetros importantes: (i) a função de distância usada para medir a similaridade (proximidade) ou dissimilaridade entre dois pontos (objetos) no espaço *p*-dimensional (*p* é a quantidade de propriedades sendo estudadas); (ii) escolha dos pontos que são usados para medir a distância entre dois grupos (método de união); (iii) quantidade de grupos que serão formados.

Existe toda uma abordagem na qual é discutido o número ótimo de grupos que serão formados a partir de um determinado conjunto de dados.

A análise de grupo pode também ser baseada em modelos de probabilidade (Bock, 1996, 1998a e 1998b). Isto nos permite saber quando um método de agrupamento fornece bons resultados (isto é, quando os dados se ajustam ao modelo) e tem levado ao desenvolvimento de novos métodos. Também tem sido mostrado que alguns dos métodos de agrupamento são métodos de estimação aproximada para determinados modelos de probabilidade. Por exemplo, o agrupamento de *K*-médias padrão e o método de Ward são relativamente próximos ao método de classificação de máxima verossimilhança na normal multivariada quando a matriz de covariância é a mesma para cada componente e proporcional a matriz identidade.

Modelo de mistura finita tem sido proposto e estudado no contexto de agrupamento (Edwards e Cavalli-Sforza, 1965). Mais recentemente tem sido reconhecido que estes modelos podem fornecer princípios estatísticos abordados em questões práticas que surgem na aplicação desses métodos de agrupamento (McLachlan e Basford, 1988; Banfield e Raftery, 1993). Em modelos de mistura finita, cada distribuição de probabilidade da componente corresponde a um grupo.

O problema da determinação do número de grupos e a escolha apropriada do método de agrupamento podem ser transformados em um problema de escolha de modelos estatísticos, e modelos que diferem em número de componentes e/ou em distribuições de componentes que podem ser comparados.

A situação mais comum encontrada é quando os objetos são de amostras diferentes, enquanto que as propriedades desses objetos serão as características de expressão dos genes que estão sendo analisados. Se o objetivo for avaliar a similaridade existente entre os genes, estes são os objetos e suas características as amostras.

3.1 O Problema de Agrupamento

Dado um conjunto com *p* elementos $X = \{X_1, X_2, ..., X_p\}$, o problema de agrupamento consiste na obtenção de um conjunto de *k* grupos, $G = \{G_1, G_2, ..., G_k\}$, tal

que os elementos contidos em um grupo G_i possuam uma maior similaridade entre si do que com os elementos de qualquer um dos demais grupos do conjunto G. O conjunto G é considerado um agrupamento com *k* grupos caso as seguintes condições sejam satisfeitas:

$$\bigcup_{i=1}^{k} G_i = X \qquad (3.1.1)$$

$$G_i \neq \varnothing, \ para\ 1 \leq i \leq k \qquad (3.1.2)$$

$$G_i \cap G_j = \varnothing, \ para\ 1 \leq i, j \leq k \ e \ i \neq j. \qquad (3.1.3)$$

O valor de *k* pode ser conhecido ou não. Caso o valor de *k* seja fornecido como parâmetro para a solução, o problema é conhecido na literatura como "problema de *k* - agrupamento" (Fasulo, 1999). Caso contrário, o *k* seja desconhecido, o problema é conhecido como "problema de agrupamento automático" e a obtenção do valor de *k* faz parte do processo de solução do problema, como em (Doval, 1999).

Em um *k* - agrupamento, o número total de diferentes formas de agrupamento de *p* elementos de um conjunto em *k* grupos, equivale à função N(*p, k),*

$$N(p,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^p. \qquad (3.1.4)$$

Com o intuito de ilustrar o crescimento exponencial do número de soluções possíveis para um problema de *k* - agrupamento, considerando a equação (3.1.4), para combinar 10 elementos em 2 grupos, 100 elementos em 2 grupos e 1000 elementos em 2 grupos, temos N(10, 2) = 511, N(100, 2) = $6{,}33825 \times 10^{29}$ e N(1000, 2) = 5.3575×10^{300} formas diferentes, respectivamente.

Para o problema de "agrupamento automático", o número total de combinações sofre um incremento significativo, sendo definido por,

$$N(p,k) = \sum_{k=1}^{p}\frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^p. \qquad (3.1.5)$$

Dessa forma, para um conjunto com 10 elementos, o agrupamento automático tem que considerar 115.975 diferentes maneiras de combinar os elementos em um número de grupos que pode variar de 1 a 10.

Outro aspecto a ser considerado em relação ao problema de agrupamento é como medir o quanto um elemento é similar a outro e, assim, identificar se ambos devem estar contidos em um mesmo grupo ou não. Para isto deve ser utilizada uma "medida de similaridade", que é específica para cada problema de agrupamento a ser tratado.

Um critério utilizado para identificar a similaridade entre dois elementos é à distância entre eles, que considera as diferenças entre os valores de cada atributo dos elementos. Neste caso, quanto menor for a distância entre um par de elementos, maior será a similaridade entre eles. Como medida de distância, consideramos a distância *Euclidiana Manhattan* e a *Canberra*.

Existem problemas de agrupamento em que a distância não pode ser utilizada, ou não é conveniente que seja utilizada, como medida de similaridade, tendo em vista que os valores dos atributos não são escalares. Como exemplo, ao tratar um problema de agrupamento que envolve atributos como sexo e endereço, são necessárias outras medidas que demonstrem o grau de similaridade entre as instâncias da base de dados. Outro exemplo em que a medida de distância não se aplica diz respeito a alguns problemas de agrupamento de vértices em estruturas de grafos em que não são considerados os pesos das arestas. Nestes problemas, também referenciados como problemas de particionamento de grafos, são necessárias medidas que considerem apenas as conexões entre os seus vértices (esta última classe de problemas não é abordada neste trabalho).

3.2 Métodos para Agrupamentos

No processo de agrupamento, a busca pela melhor solução no espaço de soluções viáveis é um problema difícil. A partir das equações (3.1.4) e (3.1.5), conforme exposto anteriormente, verificamos que a avaliação exaustiva de todas as configurações de agrupamentos possíveis é computacionalmente inviável, para problemas de médio ou grande porte, restringindo com isso o uso de métodos exatos.

Baseado neste fato, métodos aproximados têm sido propostos com freqüência, os quais fornecem soluções sub-ótimas com significativa redução da complexidade na solução do problema. Entretanto, devido à grande heterogeneidade das aplicações de agrupamento, os métodos são normalmente desenvolvidos para determinadas classes de

problemas, ou seja, não existe um método que seja genérico a tal ponto que possa obter bons resultados em todas as aplicações de agrupamento.

Os algoritmos existentes para a solução de problemas de agrupamento podem ser classificados, de forma geral, em métodos hierárquicos e métodos de particionamento (Fasulo, 1999).

3.2.1 Método Hierárquico

O método hierárquico cria uma decomposição da base de dados na forma de árvore, dividindo-a recursivamente em conjuntos de dados menores. Essa divisão pode ser feita de duas formas: *top-down* e *botton-up*. Alguns autores chamam essa divisão de divisivos e aglomerativos, respectivamente.

Na abordagem *top-down*, o processo inicia com todos os objetos no mesmo grupo, o qual vai sendo dividido sucessivamente até que cada grupo contenha um único elemento. Na forma *botton-up*, cada objeto é um grupo e, a cada passo do procedimento, os dois grupos mais próximos (similares) são unidos até que, ao final, exista um único grupo formado por todos os objetos.

Embora os métodos hierárquicos sejam empregados com sucesso em alguns casos, como por exemplo, em aplicações biológicas, não existe uma revisão do agrupamento durante a execução do procedimento, ou seja, no método hierárquico aglomerativo, uma vez realizada a junção de dois objetos dentro de um mesmo grupo, os objetos não podem mais ser separados, permanecendo no mesmo grupo até o final do procedimento. De forma análoga, no hierárquico divisivo, uma vez que dois objetos foram separados, eles nunca mais serão agrupados no mesmo grupo.

AGNES e DIANA são dois algoritmos de agrupamento hierárquico. O primeiro é um algoritmo do tipo *botton-up*, enquanto DIANA é do tipo *top-down.* Se as decisões de agrupamento ou separação tomadas em cada passo da execução não forem bem escolhidas, os grupos gerados podem ser de má qualidade.

Nos algoritmos tradicionais para o agrupamento hierárquico, os grupos vão sendo formados gradativamente através de aglomeração ou divisão de grupos, gerando uma hierarquia de grupos, normalmente representada através de uma estrutura em árvore,

conforme exemplificado na Figura 5. Nesta classe de algoritmos, cada grupo com tamanho maior que 1 pode ser considerado como sendo composto por grupos menores.

Nos algoritmos de aglomeração, que utilizam uma abordagem *botton-up*, a união ocorre de acordo com alguma medida que forneça a informação sobre quais deles estão mais próximos uns dos outros. Nos algoritmos de divisão, com uma abordagem *top-down*, a cada passo, são efetuadas divisões, formando novos grupos de tamanhos menores, conforme critérios pré-estabelecidos.

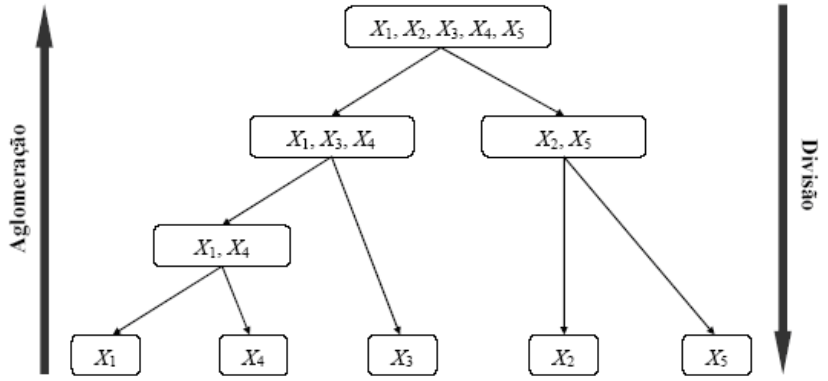


Figura 5: Exemplo de árvore no agrupamento hierárquico.

Berkhin (2002) aponta como vantagens dos algoritmos de agrupamento hierárquicos a facilidade em lidar com qualquer medida de similaridade utilizada e a sua conseqüente aplicabilidade a qualquer tipo de atributo. As desvantagens relacionam-se à imprecisão do critério de parada e ao fato de que a maioria dos algoritmos desta classe não revisitar os grupos formados ao longo de suas execuções. Este último aspecto está relacionado ao fato dos algoritmos para agrupamento hierárquico serem apenas algoritmos construtivos, não permitindo o refinamento de soluções obtidas durante a sua execução. Com relação ao critério de parada destes algoritmos, a formação dos grupos pode ser interrompida quando o número de grupos desejado for obtido, no caso de um *k*-agrupamento.

3.2.2 Método de Particionamento

Os métodos de particionamento buscam encontrar a melhor partição dos n objetos em *k* grupos. Normalmente os *k* grupos encontrados são de melhor qualidade do que os *k* grupos produzidos pelos métodos hierárquicos. Os métodos de particionamento mais utilizados são baseados em um ponto central (média dos atributos dos objetos – *K*-médias) ou em um objeto representativo para o cluster (*k*-medoides).

3.2.2.1 Algoritmo K-Médias

O *k*-médias exige a definição prévia do número de grupos e do posicionamento do centro de cada grupo k_1, k_2, ..., k_p no espaço de atributos. O centro do grupo é chamado centróide, que é o ponto médio do grupo. Esse algoritmo é sensível a ruídos, mas em termos de execução é relativamente eficiente para grandes bases de dados.

Os passos básicos do algoritmo *k*-médias são:

Passo 1: seleção de n objetos para serem centros iniciais dos *k* grupos.
Passo 2: cada objeto é associado a um grupo, esse objeto é alocado no grupo em que a dissimilaridade entre ele e o centro do grupo for menor que a dissimilaridade entre este e os demais centros.
Passo 3: os centros dos grupos são recalculados, redefinindo cada um, em função dos atributos de todos os objetos pertencentes ao grupo.
Passo 4: retorna ao passo 2 até que os centros dos grupos se estabilizem.

3.2.2.2 Algoritmo *K*-Medoides

A diferença básica em relação ao *k*-médias está na utilização de um objeto representativo, chamado medoide, localizado próximo ao centro do grupo ao invés de um centro médio. Por essa razão, este método é menos sensível a outliers que o *k*-médias. Entretanto utiliza um tempo maior de processamento. Os objetos são selecionados

aleatoriamente para serem os centros dos grupos. O algoritmo de agrupamento PAM (Particioning Around Medoids), baseado em *k*-medoides, realiza a cada passo uma busca exaustiva pela troca de um dos *k*-medoides, previamente selecionados por um dos demais *(n-k)* objetos de forma que minimize as dissimilaridades entre os *k*-medoides e os membros dos *k* grupos.

O algoritmo PAM funciona efetivamente com pequenas bases de dados. Para grandes bases de dados, Han (1995) sugere o uso do CLARA (*Clustering LARge Applications*), que utiliza uma amostra e realiza o algoritmo PAM sobre ela. Assim, ao invés de utilizar toda a base de dados, podemos usar uma amostra da base sobre a qual o algoritmo PAM é aplicado para selecionar os medoides. A média de dissimilaridade é calculada sobre toda a base de dados. Dessa forma, várias amostras são coletadas da base de dados e em seguida se aplica o PAM sobre cada amostra. Então, o CLARA pode ser aplicado sobre o melhor agrupamento gerado para cada amostra.

A eficiência do CLARA depende do tamanho das amostras. Enquanto o PAM pesquisa o melhor *k*-medoides numa base de dados, o CLARA procura pelo melhor *K*-medoide entre várias amostras da base de dados.

Nos algoritmos de aglutinação que utilizam algum método de particionamento, caso o agrupamento formado não esteja adequado ao problema em questão, novo agrupamento é obtido através da migração de elementos entre os grupos, e o processo continua de forma iterativa até que critério desejado seja alcançado. Este esquema de migração dos elementos entre os grupos é referenciado na literatura como *otimização iterativa* (Berkhin, 2002), os grupos podem ser melhorados gradativamente, o que não ocorre nos métodos hierárquicos.

4. Técnicas de Agrupamento

No processo de classificação são considerados objetos e atributos. Objeto é tudo aquilo que se quer classificar e atributos são as informações a respeito do objeto que serão consideradas no processo de classificação. Como exemplo, encontramos na Tabela 1 uma

pequena amostra dos dados reais que serão estudados neste trabalho. Construímos um gráfico com as amostras tumorais e normais, o resultado é exposto na Figura 6.

Tabela 1: Amostras tomorais e normais.

Nome	Tumor	Normal
D00265	153.220213	45.580835168
D00596	70.030183	13.821205663
D00749	65.050259	67.789723013
D00760	55.228718	21.248262793
D00761	202.505298	159.08955732
D00762	253.753568	114.95854998

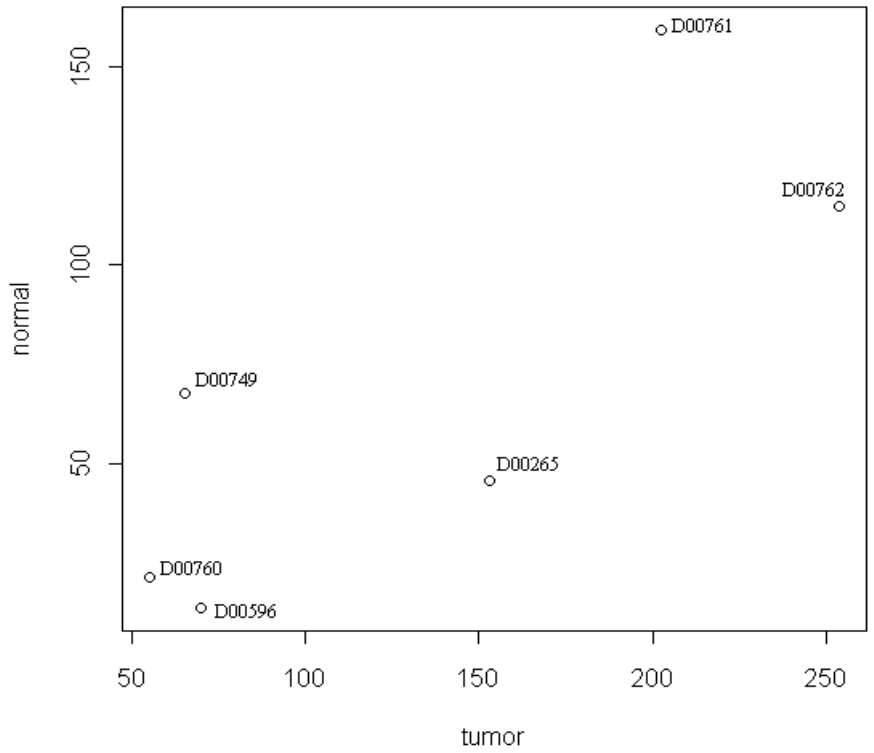


Figura 6: Gráfico relativo aos dados da Tabela 1.

Quando nos referimos às técnicas de agrupamento, o que chamamos de objetos são observações e o que chamamos de atributos são variáveis. Podemos representar essas observações e variáveis através de uma matriz M_{nxp}, onde há n observações (uma em cada linha) e p variáveis (uma em cada coluna),

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Figura 7: Matriz M_{nxp}

As técnicas de agrupamento de dados são ferramentas de análise de dados multivariados. Estas ferramentas têm por objetivo encontrar objetos similares — grupos — dentro de um determinado conjunto de dados, ou seja, dentro de uma matriz M de dados.

O principal objetivo destas técnicas é o de reunir indivíduos semelhantes e separar indivíduos diferentes.

O conjunto de dados pode ser organizado de duas maneiras:

• Matriz com os dados propriamente ditos (matriz M);

• Matriz de distâncias (matriz D) - nela constam valores que irão expressar um grau de distância, ou ainda, o quanto cada objeto é diferente do outro.

Como foi mencionado no capitulo anterior, para um conjunto de dados a ser agrupado podemos utilizar dois métodos de agrupamento, e para cada método há diversos algoritmos (funções), tais como os Métodos Particionais (PAM, Clara, dentre outros) e os Métodos Hierárquicos (Diana, AGNES, dentre outros)

4.1 Cálculo de Distâncias

Se o objetivo é agrupar dados, é importante uma mensuração de distância, ou ainda, de similaridade entre os objetos. Dentre as métricas para cálculo de distância entre objetos iremos nos ater as medidas: Euclidiana, Manhattan e Canberra.

4.1.1. Distância Euclidiana

Para calcular a distância entre cada par de objetos *i* e *j*, objeto *i*, *i*=1,2,..n, caracterizado pelos atributos $x_i = (x_{i1} \ ..., \ x_{ip})$ e objeto *j*, *j*=1,2,...,n, caracterizado pelos atributos $x_j = (x_{j1} \ ..., \ x_{jp})$, sendo *p*, *p*=1,2,...,k, o número de características estudadas (ou o número de colunas), utilizamos

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}, \qquad i \neq j \qquad (4.1.1)$$

Da tabela anterior temos:

i,*j* = 1,2,...,6.

p =1,2.

Assim, $d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}, \qquad i \neq j$

se i=1 e j=2, temos,

$$d(1,2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} = \sqrt{(153.22 - 70.03)^2 + (45.58 - 13.82)^2} = 89.04$$

A equação 4.1.1 corresponde à distância Euclidiana entre os pontos com coordenadas $(x_{i1}, \ldots, x_{ip})$ e $(x_{j1}, \ldots, x_{jp})$.

Na Tabela 2 podemos ver as distâncias, calculadas com a métrica Euclidiana, entre os objetos apresentados na Tabela 1. Podemos verificar numericamente o que já havíamos observado no gráfico. Por exemplo, havíamos verificado que há dois grupos bem definidos. É possível ver também que os genes que estão nos extremos do gráfico (D00760 e D00762) possuem maior distância entre si.

Tabela 2: Matriz de distâncias segundo a métrica Euclidiana.

	D00265	D00596	D00749	D00760	D00761
D00596	89.04637				
D00749	90,92401	54,19779			
D00760	100,96736	165,6033	47,56648		
D00761	123,74672	196,60252	165,01378	201,71915	
D00762	122,14836	209,72135	194,50922	219,53071	67,63084

4.1.2 Distância Manhattan

A Distância Manhattan entre os pontos com coordenadas $(x_{i1},\ldots,x_{ip})$ e $(x_{j1},\ldots,x_{jp})$ é dada por

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|. \qquad (4.1.2)$$

Tabela 3: Matriz de distâncias segundo a métrica Manhattan

	D00265	D00596	D00749	D00760	D00761
D00596	114,94966				
D00749	110,37884	58,94844			
D00760	122,32407	22,22852	56,36300		
D00761	162,79381	277,74347	228,75487	285,11787	
D00762	169,91107	284,86073	235,87214	292,23514	95,37928

4.1.3 Distância Canberra

A Distância Canberra, entre os pontos com coordenadas $(x_{i1},\ldots,x_{ip})$ e $(x_{j1},\ldots,x_{jp})$ é dada por,

$$d(i,j) = \frac{|x_{i1} - x_{j1}|}{|x_{i1} + x_{j1}|} + \frac{|x_{i2} - x_{j2}|}{|x_{i2} + x_{j2}|} + \ldots + \frac{|x_{ip} - x_{jp}|}{|x_{ip} + x_{jp}|} \qquad (4.1.3)$$

Tabela 4: Distâncias entre os objetos da Tabela 1, medidos pela distância Canberra.

	D00265	D00596	D00749	D00760	D00761
D00596	0,9072866				
D00749	0,5998446	0,6981567			
D00760	0,8341998	0,3299483	0,6043710		
D00761	0,6931409	1,3262187	0,9161598	1,3357790	
D00762	0,6791805	1,3527775	0,8500186	1,3305121	0,2733566

4.2 Métodos Particionais

Classificam os dados em *k* grupos, de forma que cada grupo deve conter no mínimo um objeto e cada um deve pertencer a exatamente um grupo, sendo que todos os objetos devem ser alocados em algum grupo.

Note que as condições citadas acima implicam que numa partição existem *k* grupos, tal que $k \leq n$. Assim, se $k = 1$, teremos um único grupo formado por todos os objetos e, se $k = n$, um grupo para cada objeto.

Como *k* é fixo, este deverá ser um dado de entrada do algoritmo definido pelo usuário. Ou seja, o algoritmo construirá tantos grupos quanto forem desejados. Porém, nem todos os valores de *k* levam a grupos "naturais", assim é recomendável que o algoritmo seja executado diversas vezes, objetivando selecionar o valor de *k* mais apropriado.

4.2.1 Método PAM (Partitioning Around Medoids)

O algoritmo começa com a escolha aleatória de *k* objetos da base de dados, que serão os primeiros centros, os chamados medoides, dos *k* grupos. Cada objeto é, então, atribuído ao grupo cujo medoide lhe é mais similar. Temos, portanto, os primeiros *k* grupos formados pelo algoritmo. Assim, repetidamente, são escolhidos *k* novos medoides para cada grupo, gerando *k* novos grupos. Assim como no *k*-médias, o processo termina quando não há mais mudança nos centros dos grupos.

Para escolhermos os novos medoides, todos os objetos são analisados e é escolhido como novo centro de cada grupo aquele objeto que minimiza a função objetivo tanto quanto possível. Essa função é soma de todas as dissimilaridades entre o objeto O_i e os demais objetos O_j do grupo a que ele pertence.

Ao aplicar a função PAM, com $k = 2$, aos dados da Tabela 1 obtemos o resultado a seguir.

Tabela 5: Objetos representativos selecionados dentre os *k* objetos, pelo algoritmo.

Objetos representativos		
Objetos	Tumor	Normal
D00596	70.030.183	13.821.205.663
D00761	202.505.298	15.908.955.732

Tabela 5.1: Resultado do agrupamento.

Vetor de agrupamento:						
Genes	D00265	D00596	D00749	D00760	D00761	D00762
Grupos	1	1	1	1	2	2

Neste caso, foram escolhidos os objetos D00596 e D00761 como objetos representativos e foram criados dois grupos, o primeiro composto pelos genes D00265, D00596, D00749 e D00760, o segundo por D00761 e D00762.

4.3 Métodos Hierárquicos

Neste método de agrupamento, a quantidade de grupos (valor de *k*) a serem formados não é um parâmetro, isto acontece porque o resultado final é só um grupo contendo todos os objetos. A idéia é colocar as uniões numa hierarquia decrescente de similaridade. Primeiro são agrupados os dois objetos que possuem maior similaridade, depois este grupo de dois objetos é unido com o terceiro objeto mais similar e assim sucessivamente até formar um único grupo.

Para permitir a visualização de todos os agrupamentos, construímos um gráfico em forma de árvore chamado dendrograma. Um exemplo de dendrograma pode ser visto na Figura 9. Como mencionado no capitulo anterior, existem dois tipos de métodos hierárquicos: Aglomerativo e Divisivo. Estes dois tipos de algoritmos constroem as suas hierarquias em direções opostas, possivelmente formando diferentes resultados.

4.3.1. Método Aglomerativo

No início deste processo, os agrupamentos são pequenos e os elementos de cada grupo possuem um alto grau de similaridade. Ao final do processo temos poucos agrupamentos, cada um podendo conter muitos elementos e menos similares entre si.

Primeiramente criamos uma matriz de similaridades entre os agrupamentos, lembrando que, no início do algoritmo, cada objeto é um agrupamento. Um problema dos métodos hierárquicos está nessa matriz de similaridade (Viana, 2004).

Depois de criarmos a matriz de similaridade, o passo seguinte é encontrarmos o menor valor na matriz de similaridade. Este 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.

4.3.1.1 Dendrograma

Dendrograma é a representação gráfica, em forma de árvore, da estrutura dos agrupamentos. Nos métodos aglomerativos, o dendrograma representa a ordem em que os dados foram agrupados, como mostra a Figura 8. Com relação aos métodos divisivos, o dendrograma representa a ordem em que os agrupamentos foram divididos.

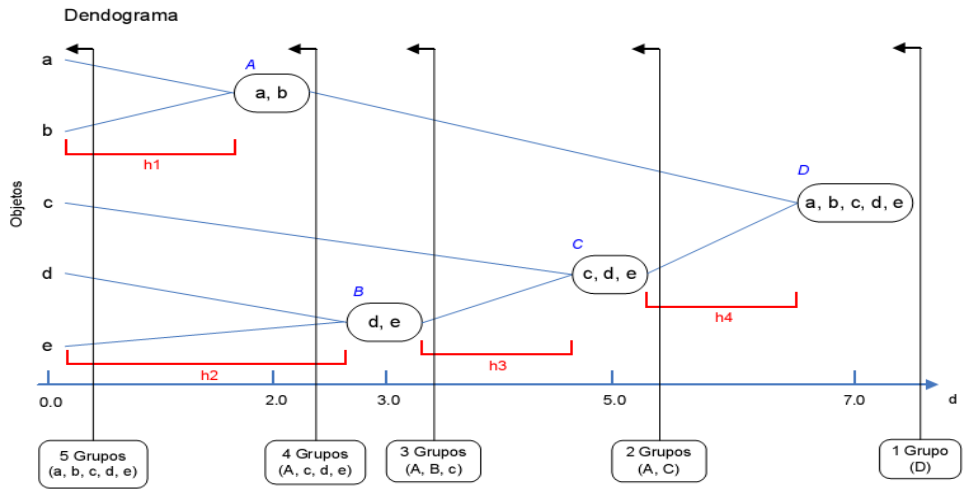


Figura 8: Dendrograma ilustrando a ordem em que os dados foram agrupados.

4.3.1.2 AGNES

O algoritmo AGNES (Agglomerative Nesting) utiliza um método aglomerativo. No início de sua execução todos os objetos formam um agrupamento. Em seguida, os dois objetos mais semelhantes são agrupados e a tabela de distâncias é refeita considerando os objetos agrupados como um só. Na reconstrução da tabela torna-se necessário o cálculo da distância entre cada objeto que não foi agrupado e o grupo recém criado.

Esta técnica adota como padrão *the group average method* introduzido por Sokal e Michener (Kaufman & Rousseeuw, 1990), baseado em argumentos de robustez, monotonicidade e consistência. Sua definição é muito simples. Sejam Ω_R e Ω_Q dois grupos, e $|R|$ e $|Q|$ o número de objetos. Então a dissimilaridade $d(\Omega_R, \Omega_Q)$ entre os grupos Ω_R e Ω_Q é definida como a média de todas as dissimilaridades $d(i, j)$, onde i é qualquer objeto de Ω_R e j é qualquer objeto de Ω_Q. Conforme definição abaixo

$$d(R,Q) = \frac{1}{|R||Q|} \sum_{\substack{i \in \Omega_R \\ j \in \Omega_Q}} d(i, j). \qquad (4.3.1.2)$$

Toma-se como exemplo a Tabela 1 onde utilizaremos a matriz de distância apresentada na Tabela 2. Pode-se observar que a menor distância (47) está entre D00749 e D00760. Agrupando estes e refazendo a matriz de distâncias obtemos a matriz apresentada na Tabela 6.

Tabela 6: Matriz de distâncias obtida depois da primeira etapa do AGNES

	D00760-D00749	D00265	D00596	D00761
D00265	95,94569			
D00596	109,9005	89,04637		
D00761	183,3665	123,74672	196,60252	
D00762	207,0200	122,14836	209,72135	67,63084

Prosseguindo com a execução, são selecionados os objetos que possuem a menor distância, no caso, D00761 e D00762. A matriz de distâncias é refeita e volta para a seleção dos objetos mais semelhantes. Este procedimento se repete até que só restem dois agrupamentos na matriz de distâncias.

Na Figura 9 é apresentado o dendrograma obtido pela execução do AGNES tendo como entrada a matriz de distâncias exposta na Tabela 2.

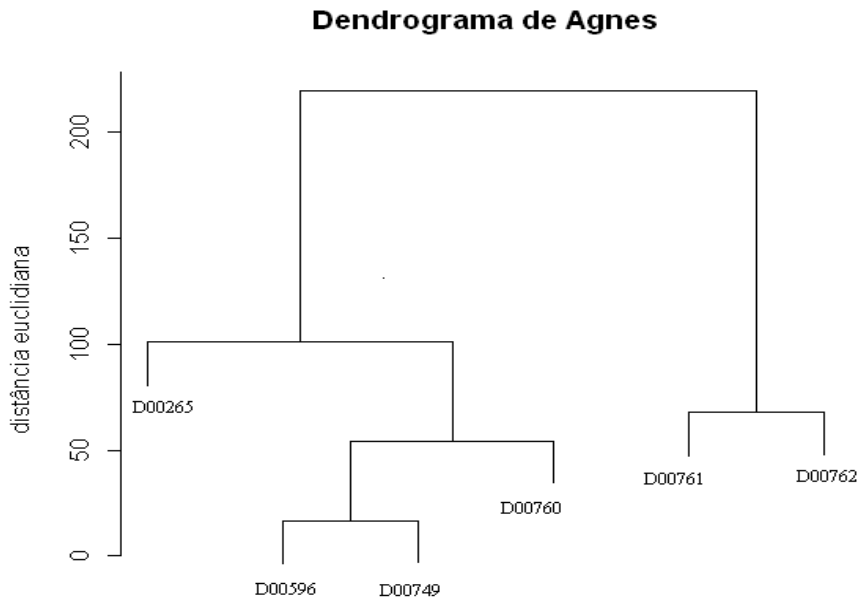


Figura 9: Dendrograma obtido pela execução do AGNES, considerando o método average e a distância Euclidiana.

4.3.2 Método Divisivo

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

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

Os métodos divisivos têm 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.

4.3.2.2 DIANA

Inicialmente todos os objetos são agrupados de forma que componham apenas um grupo. A partir deste momento o algoritmo tem que dividir os dados em dois grupos. É escolhido o objeto que possui a maior distância em relação a um grupo formado pelos

outros, utilizando para isso a Fórmula 4.3.1.2 (a mesma do AGNES). Este irá iniciar um outro grupo, trazendo para o seu grupo os objetos que mais se assemelham a ele. Como exemplo, retorna-se para a situação exposta pela Tabela 1 e a matriz de distâncias apresentada pela Tabela 2. Calculamos a distância de cada objeto em relação ao outro e obtemos os resultados apresentados na Tabela 7. O objeto que possui a maior distância em relação aos outros é D00762, assim este é escolhido para iniciar o grupo chamado dissidente.

Tabela 7: Média da distância de cada objeto em relacão aos outros

Objeto	Média da distância para os outros objetos
D00265	(89,046 + 90,924 + 100,967 + 123,747 + 122,148)/5 = 105,366
D00596	(89,046+ 54,1978 + 16,560 +196,602 + 209,721)/5 = 113,226
D00749	(90,924 + 54,198 + 47,566 + 165,014 + 194,509)/5 = 110,442
D00760	(100,967 + 16,560 + 47,566 + 201,719 + 219,530)/5 = 117,269
D00761	(123,7467 + 196,602 + 165,014 + 201,719 + 67,630)/5 = 150,943
D00762	(122,148 + 209,721 + 194,509 + 219,531 + 67,630)/5 = 162,708

Neste momento temos dois grupos {D00762} e {D00265, D00596, D00749, D00760 e D00761}. Para cada objeto do maior grupo é calculado a média da distância entre o objeto e os outros objetos, e compara esta com a distância dos objetos do grupo dissidente. Na Tabela 8 apresentamos os resultados obtidos. Selecionamos o objeto que apresenta o maior valor positivo para a variável diferença. O valor positivo para a variável diferença significa que existe uma maior afinidade entre o objeto e o grupo dissidente ao objeto com o grupo atual. O objeto selecionado, no caso D00761, vai para o grupo dissidente e a Tabela 8 é refeita agora sem o objeto D00761. A tabela obtida possui valores negativos na variável diferença para todos os objetos e assim é concluído o primeiro passo do algoritmo com os grupos {D00761, D00762} e {D00265, D00596, D00749, D00760}.

Tabela 8: Média da distância de cada objeto em relação aos outros e comparação com média do grupo dissidente.

Objeto	Média da distância para os outros objetos		Aos dissidentes	Diferença
D00265	(89,046 + 90,924 + 100,967 + 123,747)/4 =	101,171	122,148	-20,977
D00596	(89,046+ 54,1978 + 16,560 +196,602)/4 =	89,10145	209,721	-120,619

Cont. Tabela 8: Média da distância de cada objeto em relação aos outros e comparação com média do grupo dissidente.

D00749	(90,924 + 54,198 + 47,566 + 165,0134)/4 =	89,42535	194,509	-105,084
D00760	(100,967 + 16,560 + 47,566 + 201,719)/4 =	91,703	219,531	-127,828
D00761	(123,746 + 196,602 + 165,013 + 201,719)/4=	171,77	67,631	104,139

Para iniciar o próximo passo, precisamos escolher o grupo que será dividido. É escolhido o grupo que possuir o maior diâmetro. Diâmetro é a maior distância entre dois dos objetos do grupo. Em seguida, as distâncias de cada objeto em relação aos outros é calculada, selecionando o objeto que irá "fundar" o grupo dissidente. O procedimento se repete até que só existam grupos formados por um único objeto. Na Figura 11 é apresentado um dendrograma com o resultado da execução do algoritmo.

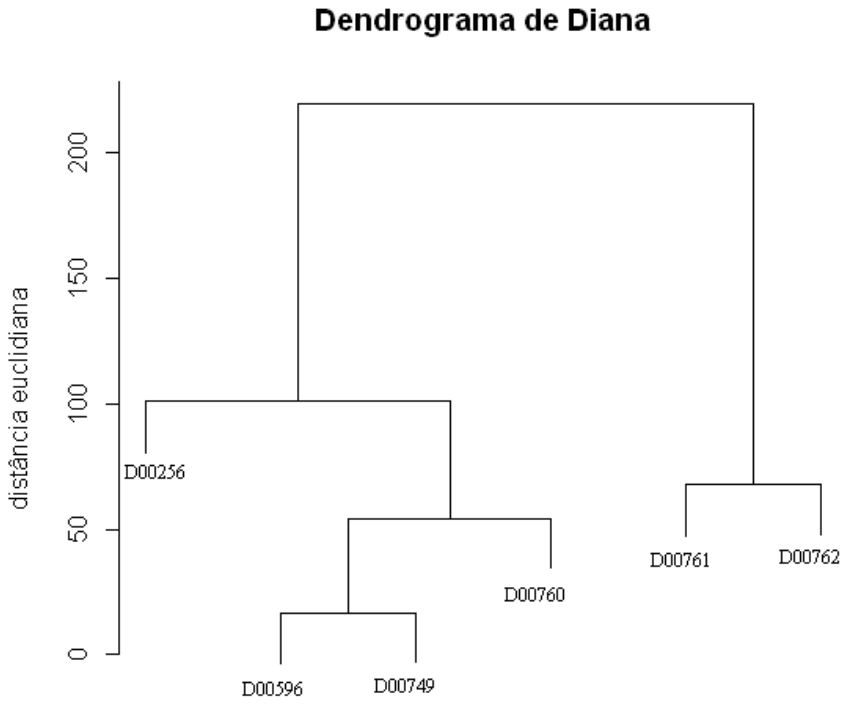


Figura 11: Gráfico obtido pelo agrupamento dos dados com Diana, considerando a distância Manhattan.

5. Redes Neurais

As redes neurais são sistemas inspirados na estrutura de funcionamento do cérebro e dos neurônios biológicos. Esse interesse foi principalmente motivado pela observação da facilidade e eficácia com que o cérebro realiza tarefas difíceis e complexas. As redes neurais resolvem problemas onde é difícil criar modelos adequados à realidade ou, então, situações que mudam muito (problemas não lineares), sem a necessidade de se definir regras ou modelos explícitos.

Devido à similaridade de uma rede neural com a estrutura de um cérebro, elas, também, acabam por exibir características semelhantes tais como:

- ✓ Aprendizado; aprende-se por experiência.
- ✓ Associação; faz associações entre padrões diferentes.
- ✓ Generalização; são capazes de generalizar o conhecimento adquirido a partir de experiências passadas.
- ✓ Abstração; Extrai a essência de um conjunto de informações retirando os ruídos.

Um exemplo da topologia de rede neural encontra-se na Figura 12, que, por convenção, é um formato de rede bastante usado, a Rede *Feedforward*. Nela, podem existir uma ou mais camadas de processamento. Na Figura mencionada, a rede possui três camadas: a camada de entrada, a camada escondida, e a camada de saída. As redes neurais possuem também unidades de processamento da informação. Essas unidades são denominadas neurônios, e são conectadas por pesos sinápticos. Vale mencionar que as redes *feedforward* não possuem realimentação.

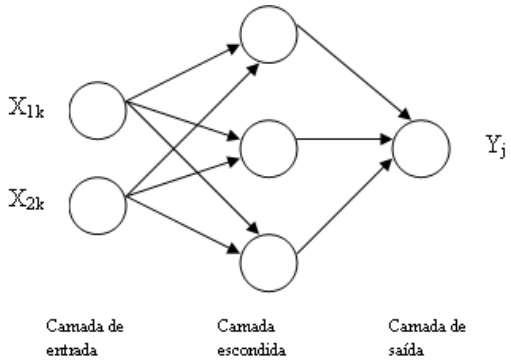


Figura 12: Rede *feedforward.*

Existe também outro tipo de rede, a recorrente. Nesse tipo de topologia existe uma conexão entre os processadores, não só de uma mesma camada como também de camadas diferentes, existindo realimentação na rede.

As redes neurais têm várias fases até que se obtenha o resultado final: a fase de treinamento, onde é retirado o conhecimento do ambiente; a fase de generalização onde o conhecimento adquirido na fase anterior, é testado, para verificar se o que foi aprendido pode ser utilizado para o fim desejado. No entanto existem diversas formas supervisionadas ou não supervisionadas. No primeiro, trabalhamos com conjuntos de pares de entrada e de saída, ambos previamente conhecidos e representantes da realidade. Já no aprendizado não supervisionado, não se trabalha com conjuntos previamente conhecidos, estabelecemos uma medida que representa a qualidade da representação da rede, e os parâmetros são modificados de forma a otimizá-la.

5.1 Método de união

Para avaliar a similaridade entre dois grupos, ou um grupo e um elemento, precisamos obter um ponto que represente o grupo, o qual vai ser usado no cálculo de $d(x_i, y_i)$. A seguir apresentamos o método utilizado.

a) Vizinho mais próximo, *single*

Escolhe de cada grupo um ponto de tal forma que $d(x_i, y_i)$ seja a menor possível.

b) Vizinho mais distante, *complete*

Escolhe de cada grupo um ponto de forma que $d(x_i, y_i)$ seja a maior possível.

c) Centróide, *average*

$d(r,s) = d(\overline{x}_r, \overline{x}_s)$, na qual r e s são os dois grupos analisados, $\overline{x}_r$ e $\overline{x}_s$ os respectivos centróides. Ou seja, para cada grupo, calculamos seu ponto médio, achando

para cada uma das *p* variáveis o valor médio determinado pelo valor que ela apresenta, nos n_r e n_s objetos, respectivamente.

5.2 SOM

O algoritmo de aprendizado auto-organizado é chamado de Mapa Auto-Organizado de Kohonen (*Self-Organizing Map*, ou SOM), desenvolvido pelo finlandês Teuvo Kohonen no começo dos anos 80. Um mapa de Kohonen é uma arranjo de neurônios, geralmente restritos a um espaço de dimensão 1 ou 2, que procura estabelecer e preservar noções de vizinhança (preservação topológica). No caso de agrupamento de dados existirá um mapeamento do espaço original (em que os dados se encontram) para o espaço em que está definido o arranjo de neurônios. Como geralmente o arranjo de neurônios ocorre em espaços de dimensão reduzida, será necessária uma redução de dimensionalidade sempre que o espaço original apresentar uma dimensão mais elevada. Toda redução de dimensionalidade pode implicar na perda de informação (por exemplo, violação topológica). Sendo assim, este mapeamento deve ser tal que minimiza a perda de informação. O algoritmo SOM é baseado no chamado aprendizado competitivo.

5.2.1 Aprendizado Competitivo

Uma rede neural típica com aprendizado competitivo é uma rede de uma única camada (uni ou bidimensional) em que todos os neurônios recebem a mesma entrada. Cada neurônio calcula o seu nível de ativação multiplicando o seu vetor de pesos pelo vetor de entrada. A escolha do vetor de pesos pode ser feita atribuindo pequenos valores tomados de um gerador de números aleatórios; desta forma, nenhuma ordem prévia é imposta ao mapa de características. O neurônio que tiver o maior nível de ativação é chamado de vencedor e apenas ele terá atividade diferente de zero na saída da rede, ou seja, o padrão de entrada que estiver sendo apresentado à rede provocará a ativação de apenas um neurônio da rede neural.

Um exemplo de uma rede com aprendizado competitivo em duas dimensões é mostrado na Figura 13. Os padrões de entrada são vetores N dimensionais x_k, k = 1, ..., n. Vamos supor que existam *P* desses padrões. A rede neural consiste de *M* neurônios organizados em uma grade bidimensional. No nosso caso, *M* = 16.

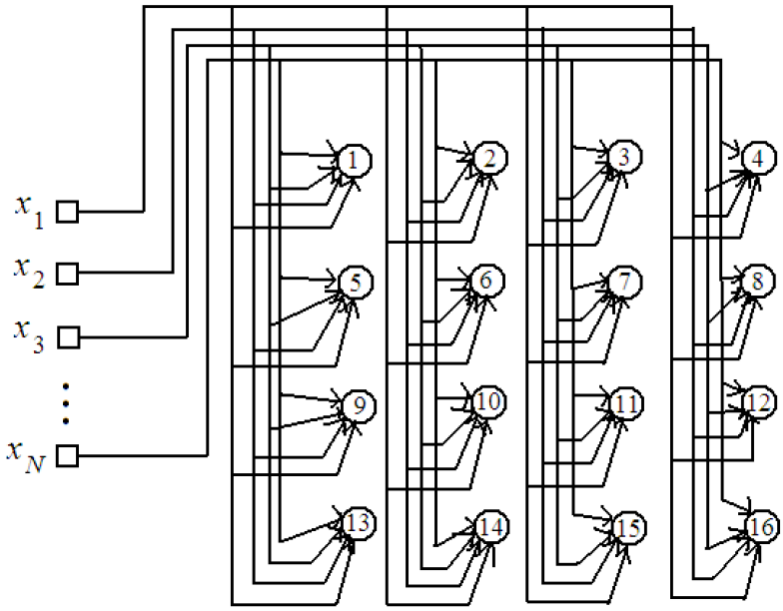


Figura 13: Rede neural bidimensional.

Quando um dos *P* padrões é apresentado na entrada, cada um dos *M* neurônios recebe este padrão e calcula o seu nível de ativação,

$$u_i = \sum_{k=1}^{N} w_{ik} x_k \,, \qquad\qquad i = 1, \ldots, M \,,$$

onde $\boldsymbol{x}$ é o vetor de entrada, *i* é o índice que indica o neurônio e w_i é o vetor de pesos entre o padrão de entrada e o neurônio *i* (os pesos não estão mostrados na figura para não congestioná-la).

O neurônio vencedor é aquele que tiver o maior valor de u_i. Vamos designá-lo por i*. Segundo a regra do aprendizado competitivo, apenas o neurônio vencedor terá saída diferente de zero após a apresentação do padrão *x*. A resposta da rede ao padrão *x* é,

$$y_i(x) = \begin{cases} 1 \ \textit{para} \ i = i^* \\ 0 \ \textit{para} \ i \neq i^* \,. \end{cases}$$

Esse tipo de rede implementa um mapa entre um espaço n-dimensional contínuo de vetores ***x*** e um espaço discreto de *M* neurônios. Note que podemos ter mais de um vetor ***x*** sendo representado pelo mesmo neurônio vencedor. Neste caso, este neurônio é o representante do grupo de padrões ***x*** que o fazem ser vencedor. Desta forma, este tipo de rede neural usa um mecanismo de agrupamento (ou *clustering*) de padrões (padrões com características similares são representados pelo mesmo vetor da rede).

Após a determinação do neurônio vencedor em resposta a um padrão de entrada ***x***, ocorre a alteração nos pesos da rede. Pela regra do aprendizado competitivo, apenas os pesos do neurônio vencedor são modificados. Chamando o vetor de pesos do neurônio vencedor de w_{i*}, a regra do aprendizado competitivo implica em,

$$W_{i^*}(n+1) = W_{i^*}(n) + \eta[x(n) - W_{i^*}(n)] \quad e$$
$$W_{i^*}(n+1) = W_{i^*}(n) \;\; para \; i \neq i^* \;\; ,$$

onde η é a constante da taxa de aprendizagem (um valor entre 0 e 1) que controla a rapidez com que as mudanças nos pesos são feitas.

A normalização dos padrões de entrada provoca alterações neles, o que em alguns casos não é desejável. Para evitar isto, costumamos usar como critério de definição do neurônio vencedor o cálculo da distância euclidiana entre o vetor ***x*** e o vetor de pesos w_i,

$$\| X - W_i \| = \sqrt{\sum_{j=1}^{N} (x_j - w_{ij})^2} \;\; .$$

O neurônio vencedor i* é aquele cujo vetor de pesos tiver a menor distância euclidiana com o padrão de entrada. Este critério de escolha do neurônio vencedor foi utilizado pela primeira vez por Stephen Grossberg em 1969.

Uma vez determinado o neurônio vencedor, o seu vetor de pesos é modificado pela regra do aprendizado competitivo, $\Delta w_{i*} = \eta(\mathbf{x} - w_{i*})$. Ela nos diz que o vetor de pesos do neurônio vencedor deve ser modificado por um fator η na direção de $\mathbf{x} - w_{i*}$. Geometricamente, em duas dimensões,

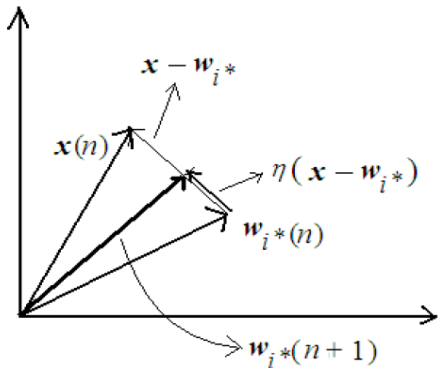


Figura 14: Vetor de pesos do neurônio vencedor.

Esta regra de mudança de pesos faz com que o vetor de pesos do neurônio vencedor, que já era o mais próximo do padrão de entrada, seja arrastado na direção do padrão de entrada, ficando ainda mais próximo dele.

A partir de uma população de padrões de entrada, retiramos, a cada passo, um dos padrões e aplicamos à entrada da rede. O neurônio vencedor é determinado e o seu peso alterado conforme a regra do aprendizado competitivo. Repetindo-se este procedimento várias vezes, os pesos da rede acabam convergindo para uma situação de relativa estabilidade em que eles ficam nos centros de massa de agrupamentos de padrões de entrada.

5.2.2 Agrupamento usando SOM

O método SOM tem sido amplamente utilizado em pesquisas recentes de agrupamento (**"Self-Organizing Maps"**, Kohonen T., 1997), o qual pode ser usado no reconhecimento de padrões, como também para projetar e visualizar objetos, assumindo uma rede neural como uma matriz regular de uma ou duas dimensões.

O algoritmo SOM de Kohonen é baseado no aprendizado competitivo descrito em 5.2.1. Uma rede SOM pode ser vista na Figura 15. A camada de entrada recebe padrões que são vetores *N*-dimensionais, provenientes de alguma população de padrões e a rede SOM é uma camada bi-dimensional (como no caso do desenho) ou uni-dimensional.

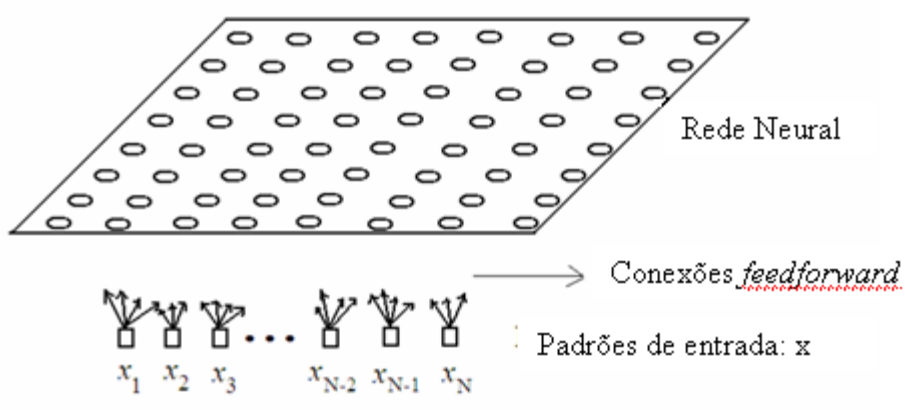


Figura 15: Rede *feedforward.*

Quando um padrão *x* é apresentado na entrada da rede, a unidade vencedora, i*, é aquela cuja distância euclidiana entre o seu vetor de pesos w_{i*} e o padrão *x* for a menor de todas (como no caso do aprendizado competitivo),

$$i^*(x) = \min_j \left\| x - w_j \right\|.$$

A diferença é que agora não é só o neurônio vencedor que tem o seu vetor de pesos atualizado, mas todos os neurônios vizinhos do neurônio vencedor. A regra de mudança de pesos do algoritmo SOM é,

$$w_i(t+1) = w_i(t) + \Lambda_{i,i^*}(t)\eta(t)\left[x(t) - w_i(t)\right],$$

onde Λ_{i,i^*} é a função de vizinhança centrada no neurônio vencedor i^* e $\eta(t)$ é a taxa de aprendizagem. Em geral, tanto Λ_{i,i^*} como $\eta(t)$ variam com o passo de aprendizagem *t*.

Com a introdução da função de vizinhança Λ_{i,i^*} fazemos com que não apenas o vetor de pesos do neurônio vencedor seja modificado na direção do padrão atual, mas também os vetores de pesos de todos os neurônios vizinhos ao neurônio vencedor sejam arrastados na direção do padrão atual, porém por um fator menor vai diminuindo à medida que o neurônio correspondente vai ficando mais distante do vencedor. Quando o vetor de pesos de um neurônio vencedor em um dado passo for modificado, ele irá arrastar consigo os demais vetores de peso, e mais fortemente aqueles dos neurônios mais próximos.

Geralmente usamos uma função gaussiana para implementar a função de vizinhança,

$$\Lambda_{i,i^*}(t) = \exp\left(-\frac{d^2{}_{i,i^*}}{2\sigma^2(t)}\right),$$

onde $d_{i,\,i^*}$ é a distância entre um neurônio *i* e o neurônio vencedor *i**. Se a rede neural for uni-dimensional, essa distância é simplesmente o módulo da diferença entre os índices de *i* e *i**, se a rede for bi-dimensional essa distância é dada pela distância euclidiana entre os seus vetores de posição,

$$d^2{}_{i,i^*} = \left\| r_i - r_{i^*} \right\|^2,$$

onde r_i é o vetor posição do neurônio i e r_{i^*} é o vetor posição do neurônio vencedor, os dois sendo medidos nos espaço discreto definido pelos nodos da rede neural.

O desvio padrão da função de vizinhança, $\sigma_{i,i^*}(t)$, diminui com o número de passos t. Uma maneira comum de implementar essa diminuição é por um decaimento exponencial,

$$\sigma(t) = \sigma_0 \exp(-\frac{t}{\tau_1}),$$

onde τ_1 é uma constante temporal (determinada empiricamente).

Usualmente, fazem a taxa de aprendizagem $\eta(t)$ diminuir com o passo de iteração de uma maneira exponencial,

$$\eta(t) = \eta_0 \exp(-\frac{t}{\tau_2}),$$

onde τ_2 é outra constante temporal (também determinada empiricamente).

A função de vizinhança, aplicada ao algoritmo SOM, faz com que a representação dos padrões do espaço de entrada pela rede neural em grupos preserve a topografia do espaço de entrada. Isto implica que padrões vizinhos no espaço de entrada são representados por neurônios vizinhos na rede neural. Além disso, regiões do espaço de entrada cujos padrões *x* tenham maior probabilidade de ocorrer são representadas por um número maior de neurônios da rede neural, isto é, elas são representadas com uma resolução maior do que as regiões cujos padrões ocorrem menos freqüentemente. Deste

modo, um mapa entre o espaço contínuo de entrada e o espaço discreto da rede neural, executado pelo algoritmo SOM, tende a preservar tanto a métrica como a distribuição do espaço de entrada. É válido ressaltar que isto ocorre apenas a partir da informação contida nos padrões de entrada; o mapa criado pelo SOM não é supervisionado.

Durante o treinamento segundo o algoritmo SOM ocorrem, em geral, duas fases distintas:

1. Fase de auto-organização ou de ordenamento: É nesta fase que ocorre o ordenamento topográfico dos vetores de pesos. Inicialmente, os vetores de pesos têm valores aleatórios e não possuem qualquer tipo de ordenação. À medida que a rede vai treinando vetores de neurônios vizinhos entre si, no espaço da rede neural, começam a se aproximar uns dos outros, de maneira que neurônios de uma mesma área da rede neural acabam representando padrões vindos de uma mesma região do espaço de entrada. Algumas dicas úteis para que ela ocorra da melhor maneira possível, podem ser vistas em (Haykin, 2001).

2. Fase de Convergência: Nesta fase ocorre o refinamento do mapa, levando a uma representação mais acurada do espaço de entrada por parte da rede neural. A taxa de aprendizagem η deve permanecer pequena, da ordem de 0,01 e a função de vizinhança de englobar apenas o próprio neurônio vencedor e, no máximo, os seus primeiros vizinhos.

A seguir uma síntese do algoritmo SOM de Kohonen:

1. Inicialização: Escolha valores aleatórios para as componentes iniciais dos vetores de pesos $w_i(0)$, i = 1,..., M. A única restrição é que os M vetores de pesos sejam diferentes uns dos outros e é conveniente que os seus módulos sejam pequenos;

2. Escolha do padrão de entrada: De acordo com alguma distribuição de probabilidades *p(x)*, escolha um padrão *x* da população para ser colocado na camada de entrada da rede;

3. Determinação do neurônio vencedor: Use o critério de similaridade baseado na distância euclidiana entre o vetor de entrada e os vetores de peso para determinar o neurônio vencedor *i** para o passo atual,

$$i^*(x) = \min_j \left\| x - w_j \right\|;$$

4. Atualização dos pesos: Modifique os vetores de pesos dos neurônios da rede de acordo com a regra,

$$w_i(t+1) = w_i(t) + \Lambda_{i,i^*}(t)\eta(t)\left[x(t) - w_i(t)\right];$$

5. Continuação: Volte para o passo 2 e continue até que não sejam observadas mudanças significativas no mapa formado.

Existem diferentes topologias para estruturação de um Mapa Auto-Organizável, sendo que a estrutura mais comum é a de duas dimensões. Podemos visualizar na Figura 16, uma rede SOM unidimensional (a), uma rede bidimensional com organização hexagonal dos neurônios (b) e uma rede bidimensional com disposição retangular dos neurônios (c).

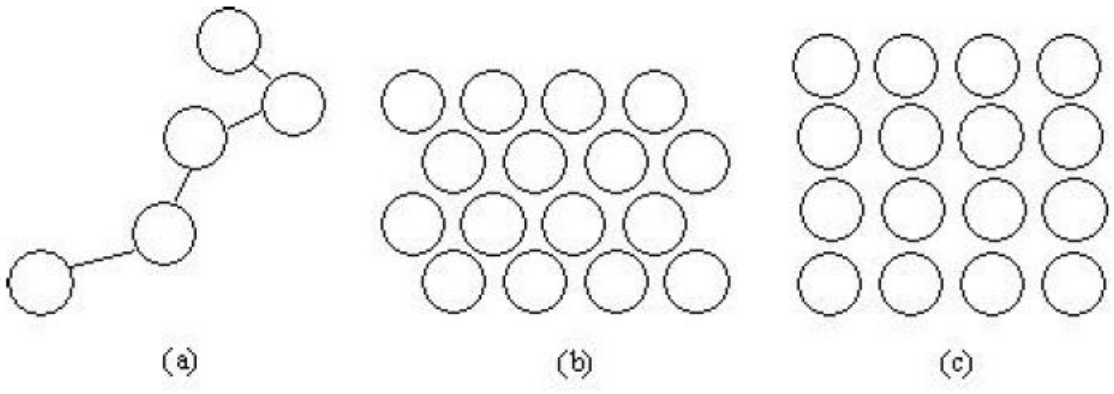


Figura 16: Diferentes topologias para estruturação dos mapas auto-organizáveis:
a) unidimensional; b) bidimensional hexagonal; c) bidimensional retangular.

A escolha do tipo de topologia pode influenciar no número de elementos em cada grupo. Se considerarmos a forma hexagonal, esta permite a cada neurônio interagir com até seis vizinhos. A outra forma de rede normalmente utilizada, retangular, só permite no máximo quatro vizinhos, o que diminui a cooperação entre os neurônios. Podemos verificar tais diferenças no exemplo a seguir.

Exemplo 1.

Também aplicamos o método SOM ao conjunto de dados Íris. Esta é uma base de dados que apresenta as medidas em centímetro da largura e comprimento das pétalas e sépalas de três espécies de flor íris (Setosa, Virginica e Versicolor) (Fisher, 1936). Esse conjunto de dados contem 150 amostras e 5 variáveis (Sépala - largura e comprimento, Pétala – largura e comprimento e Espécie). Utilizamos estes dados por conhecermos a classificação correta por espécies, podendo assim avaliar a eficiência do método. Podemos visualizar na Tabela 9.

Tabela 9: Conjunto de dados Íris.

Observação	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Espécie
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
. . .	. . .	. . .	. . .	. . .	. . .
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

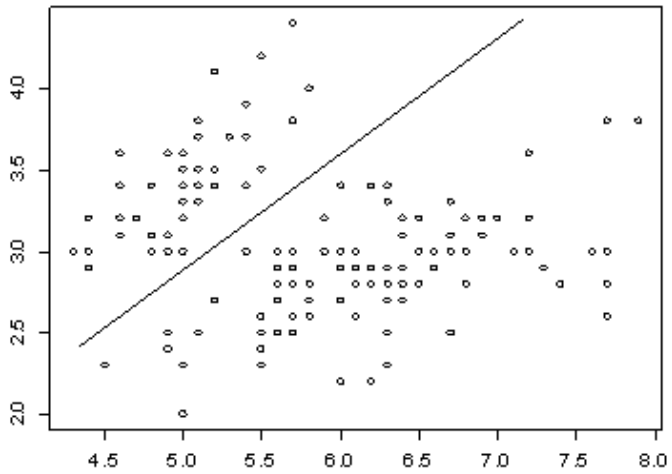


Figura 17: Gráfico com os dados Íris.

Caso não soubéssemos a verdadeira classificação dos dados, poderíamos supor que os mesmos se dividiram em 2 grupos, pois de acordo com a Figura 17 é possível ver um agrupamento natural dos dados em 2 grupos.

A seguir mostramos os mapas com as duas topologias: Retangular e Hexagonal.

Aqui podemos verificar a mudança no número de genes nos grupos devido à topologia aplicada, como havíamos mencionado anteriormente.

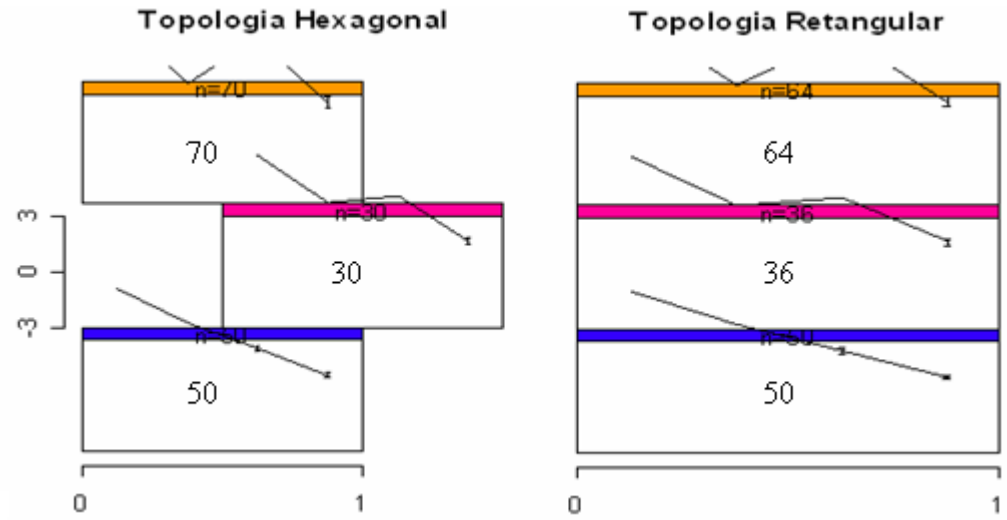


Figura 18: Gráficos com as duas topologias, hexagonal e retangular, do método SOM, considerando a função de vizinhança gaussiana e dimensão do mapa (1 x 3).

Tabela 10. Quantidade de observações em cada grupo, pelo método SOM.

Grupos	Topologia hexagonal	Topologia retangular
1	70	64
2	30	36
3	50	50

De acordo com os gráficos podemos observar que o algoritmo está fazendo uma confusão nos grupos 1 e 2, tanto na topologia retangular quanto na hexagonal. Uma explicação para tal, é que o agrupamento natural para estes dados é de apenas dois grupos, assim quando impomos um mapa 1x3 (ou seja, 3 grupos) o algoritmo não consegue formar os três grupos corretamente.

6. Agrupamento baseado em Modelos de Mistura

A análise de agrupamento pode também ser baseada em modelos de probabilidade (ver Bock, 1996, 1998a, 1998b). Isto nos permite saber quando um método de agrupamento fornece bons resultados (isto é, quando os dados se ajustam ao modelo) e tem levado ao desenvolvimento de novos métodos. Também tem sido mostrado que alguns dos métodos heurísticos de agrupamentos, mais populares, são métodos de estimação aproximada para determinados modelos de probabilidade. Por exemplo, o agrupamento de *k*-médias padrão e o método de *Ward* são relativamente próximos ao método de classificação de máxima verossimilhança na normal multivariada quando a matriz de covariância é a mesma para cada componente e proporcional a matriz identidade.

Modelos de mistura finita têm sido muitas vezes propostos e estudados no contexto de agrupamento (Wolf, 1963, 1965, 1967, 1970; Edwards e Cavalli-Sforza, 1965). Nesses modelos, cada distribuição de probabilidade da componente corresponde a um grupo. O problema da determinação do número de grupos e a escolha apropriada do método de agrupamento podem ser transformados em um problema de escolha de modelos estatísticos, e modelos que diferem em número de componentes e/ou em distribuições de componentes que podem ser comparados. Os *outliers* são tratados adicionando uma ou mais componentes representando uma distribuição diferente para estes dados.

Dentre outras, apresentamos a abordagem proposta por Fraley e Raftery (2002), que aborda dois aspectos relacionados aos parâmetros de interesse do modelo de mistura.

O primeiro aspecto diz respeito à quantidade de componentes que particionam a população. Neste caso, o número de componentes M é desconhecido, então utilizamos uma técnica de seleção de modelos para estimar o valor de M. O segundo aspecto é referente à questão da estimação dos parâmetros desconhecidos. Uma possível abordagem é a utilização de um processo via máxima verossimilhança para obter as estimativas dos parâmetros. Dependendo da forma da função de verossimilhança o problema de maximização pode se tornar bastante complexo exigindo técnicas mais elaboradas. Uma destas técnicas é o algoritmo EM.

6.1 Modelo com mistura

6.1.1 Fundamentos

Os modelos estatísticos formulados partindo de misturas de distribuições são aplicados quando as observações são provenientes de uma população particionada em vários grupos ou componentes, sendo que não se sabe à qual destas componentes pertence cada observação em particular. Desta forma, os modelos com mistura apresentam características que expressam a diversidade do número de componentes, que particionam a população, bem como as diferentes distribuições existentes.

Para definirmos um modelo com mistura de distribuições, considere um vetor aleatório y = (y_1,..., y_v) assumindo valores no conjunto dos números reais R^v. Se a distribuição de *Y* pode ser representada por uma função densidade de probabilidade, no caso de *Y* assumir valores contínuos, ou função de probabilidade no caso do vetor assumir valores discretos, da forma,

$$f(y) = \sum_{k=1}^{G} \tau_k \, f_k(y_i) \qquad (6.1)$$

onde $\sum_{k=1}^{G} \tau_k = 1$, $\tau_k \geq 0$, $f_k(.) \geq 0$ e $\int f_k(y)\,dy = 1$ com $k = 1,2,...,G$ e i = 1, ..., v.

Assim, podemos afirmar que y = (y_1,..., y_v) possui uma distribuição com mistura sendo que $f(.)$, definida em (1), é uma função densidade com mistura finita.

Temos que em (6.1) o parâmetro τ_k é a probabilidade de uma observação pertencer à *k-ésima* componente e $f_1(.),..., f_G(.)$ são chamadas de densidades componentes da mistura e representam qualquer distribuição.

Considerando $f_k(y) = f_k(y \mid \theta_k)$ para $k = 1,2,...,G$, ou seja, as densidades da mistura $f_k(y)$ pertencendo a uma mesma família, têm função de densidade com mistura finita é expressa como,

$$f(y \mid \theta, \tau) = \sum_{k=1}^{G} \tau_k \, f(y \mid \theta_k). \qquad (6.2)$$

Assim, τ_k é a probabilidade associada a cada componente da mistura produzir as observações y = (y_1, ..., y_v).

6.2 Função de verossimilhança

Dado o vetor $y_i = x_{i1}, x_{i2} ..., x_{iv}$ para $i = 1,..., n$, a função de verossimilhança construída a partir do modelo (6.2), é dada por

$$L_{mix}(\theta_1,...,\theta_G;\tau_1,...,\tau_G \mid y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \, f_k(y_i \mid \theta_k) \qquad (6.3)$$

e pode ser pensada como uma função dos parâmetros quando os dados estão fixos.

Em alguns casos, f_k (.) é a densidade ϕ_k de uma distribuição normal multivariada parametrizada pelo vetor de médias μ_k e a matriz de covariâncias Σ_k ,

$$\phi_k(y_i \mid \mu_k, \Sigma_k) \equiv \frac{\exp\{-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\}}{\sqrt{\det(2\pi\Sigma_k)}} \, . \qquad (6.4)$$

Os dados gerados pela densidade normal multivariada são caracterizados por grupos centrados na média μ_k. As correspondentes superfícies de densidades constantes são elipsoidais. Características geométricas (forma, volume, direção) de um grupo são determinadas pelas covariâncias Σ_k, que podem também ser parametrizadas para impor restrições a grupos desfavoráveis. Um caso comum é quando $\Sigma_k = \lambda I$, onde todos os grupos são esféricos e do mesmo tamanho, $\Sigma_k = \Sigma$, através de grupos constantes, onde todos possuem a mesma geometria, mas não necessariamente esférica (Friedman e Rubin,

1967); e Σ_k não restrito, onde cada grupo pode ter uma geometria diferente (Scott e Symans, 1971). Para $\Sigma_k = \lambda I$, somente um parâmetro é necessário para caracterizar a estrutura de covariância de uma mistura. Considerando $d(d+1)/2$ e $G(d(d+1)/2)$ parâmetros são necessários para Σ_k constante e Σ_k não restrita, respectivamente, se os dados são *d*-dimensionais.

Banfield e Raftery (1993) propuseram uma estrutura geral para a geometria de restrições de grupos desfavoráveis em mistura de normais multivariadas pela parametrização das matrizes de covariância, através da decomposição em autovalores na forma,

$$\Sigma_k = \lambda_k D_k A_k D_k^T , \qquad (6.5)$$

sendo D_k a matriz ortogonal de autovetores, A_k é a matriz diagonal cujos elementos são proporcionais aos autovalores. λ_k está associada a uma constante de proporcionalidade. A idéia é ajustar $\lambda_k, A_k \, e \, D_k$ como conjuntos independentes com as mesmas restrições dentro do grupo podendo haver variações entre grupos. Quando os parâmetros são fixos, os grupos partilham certas propriedades geométricas; D_k dirige a orientação da *k*-ésima componente da mistura, A_k é a forma e λ_k é o volume que é proporcional a $\lambda_k^d \det(A_k)$. Por exemplo, se o maior autovalor de Σ_k é muito maior que os outros autovalores, então o *k*-ésimo grupo estará aproximadamente concentrado em uma linha no espaço *d*-dimensional, e será a primeira componente principal da distribuição do *k*-ésimo grupo. Similarmente, se os dois maiores autovalores são da mesma magnitude e dominam os outros autovalores, então o *k*-ésimo grupo estará concentrado próximo a um plano no espaço *d*-dimensional. O *k*-ésimo grupo será aproximadamente esférico se o maior e o menor autovalor de Σ_k são da mesma magnitude.

Esta abordagem generaliza o trabalho de Murtagh e Raftery (1984), que utilizou o modelo de forma ou volume ($\Sigma_k = \lambda D_k A D_k^T$) igual para agrupamento em reconhecimento de características e outras situações envolvendo pequenos grupos

amplamente lineares, e possivelmente sobrepondo estes com orientações diferentes. Isto também classifica três modelos mais comuns: λI , variância igual e variância não restrita, mencionados anteriormente, bem como outros modelos úteis, tais como $\Sigma_k = \lambda_k I$, onde todos os grupos são esféricos, mas possuem volumes diferentes, e $\Sigma_k = \lambda_k A_k$, onde todas as covariâncias são diagonais diferindo em suas formas, tamanhos e permitindo orientações variadas. Para uma extensa enumeração de possíveis modelos resultantes de (6.5), ver Celeux e Govaert, 1995.

6.3 Variáveis Latentes

Variáveis latentes são usadas, num sentido estatístico, para descreverem efeitos genéticos ou ambientais compartilhados pelos indivíduos, ou ainda, covariáveis não consideradas no estudo.

Segundo Gelman *et al.* (1995), o procedimento básico a ser feito quando se trabalha com misturas de distribuições é utilizar um algoritmo baseado em uma amostra ampliada que tem por objetivo classificar as observações em relação as suas componentes. Para ampliar a amostra, agregamos ao vetor de observações $y_i = (y_{i1}, y_{i2}..., y_{ik})$ para $i = 1,..., n$, vetores $z_i = (z_{i1}, z_{i2}..., z_{ik})$ de variáveis aleatórias indicadoras não observáveis, chamadas de variáveis latentes. Estas variáveis se comportam de forma que $z_{ik} = 1$ se a *i*-ésima observação é proveniente da *k*-ésima componente da mistura e $z_{ik} = 0$ caso contrário, observando a restrição $\sum_{k=1}^{G} z_{ik} = 1$, que faz com que o vetor z_i indique que a observação y_i seja atribuída a somente uma das componentes da mistura.
O vetor z_i segue uma distribuição multinomial,

$$z_i \mid y, \theta, \tau \sim multinomial(1; q_i = (q_{i1},...,q_{iG})) \qquad (6.6)$$

sendo que segundo Gilks *et al.* (1996), cada componente é dada por

$$q_{ik} = \frac{\tau_k f(y_i \mid \theta_k)}{\sum_{k=1}^{G} \tau_k f(y_i \mid \theta_k)} \qquad (6.7)$$

para $i = 1,\ldots,n$ e $k = 1,\ldots,G$.

Segundo Diebolt *et al.* (1994), a classificação feita das observações em relação às componentes com a inclusão das variáveis latentes atua como uma estrutura oculta do modelo, que pode ser visto como dados perdidos (*missing*). A inclusão das variáveis latentes z_i, para $i = 1,\ldots,n$, além de classificar as observações, contribui para a simplificação da função de verossimilhança, pois após a inclusão de z_i, a função de verossimilhança (6.3) passa a ser descrita como,

$$L_{mix}(\theta_1,\ldots,\theta_G;\tau_1,\ldots,\tau_G \mid y) = \prod_{k=1}^{G}\prod_{i=1}^{n}[\tau_k f_k(y_i \mid \theta_k)]^{z_{ik}}. \qquad (6.8)$$

Vamos abordar dois aspectos relacionados aos parâmetros de interesse no modelo (6.8).

O primeiro aspecto diz respeito à quantidade de componentes que particionam a população. Quando existe a informação a respeito da quantidade de componentes, então não existe problema em relação à dimensionalidade do espaço paramétrico, mas quando o número M de componentes é desconhecido, devemos utilizar procedimentos como, por exemplo, técnicas de seleção de modelos para estimar o valor M. Neste trabalho consideramos o número de componentes desconhecido.

O segundo aspecto é referente à questão da estimação dos parâmetros desconhecidos $\theta e \tau$. Uma possível abordagem é a utilização de um processo via máxima verossimilhança para obter as estimativas dos parâmetros. Neste tipo de abordagem, o objetivo é encontrar valores das estimativas dos parâmetros que maximizam o logaritmo da verossimilhança, ou seja, tentamos maximizar o $\log[L(y,z \mid \theta,\tau)]$. Dependendo da forma de $L(y,z \mid \theta,\tau)$ o problema de maximização pode se tornar bastante complexo exigindo técnicas mais elaboradas. Uma destas técnicas é o algoritmo EM (*Expectation-Maximization*). O algoritmo EM é um método para encontrar estimadores de máxima verossimilhança dos parâmetros de modelos que possuem uma estrutura com variáveis latentes ou dados perdidos.

6.4 Algoritmo EM para Modelos de Mistura

O algoritmo EM (Dempster, Laird e Rubin, 1977; McLachlan e Krishnan, 1977) é uma abordagem geral para estimação de máxima verossimilhança em problemas onde os dados podem ser vistos como consistindo de n observações multivariadas x_i, recuperáveis a partir de (y_i, z_i), em que y_i é observado e z_i é não observado. Se os x_i são independentes e identicamente distribuídos (iid), segundo uma distribuição de probabilidade *f(.)* com parâmetros θ, então a verossimilhança completa dos dados é

$$L_c(x_i \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Além disso, se a probabilidade que uma variável seja não observada depende somente dos dados observados em y e não em z, então a verossimilhança observada, $L_o(y \mid \theta)$, pode ser obtida integrando em z a verossimilhança completa dos dados:

$$L_o(y \mid \theta) = \int L_c(x \mid \theta) dz.$$

A estimativa de máxima verossimilhança (MLE) para θ baseado nos dados observados, maximiza $L_o(y \mid \theta)$.

O algoritmo EM alterna entre dois passos, no "passo E" calculamos a esperança condicional da log-verossimilhança completa dos dados, dado os dados observados e a estimativa do parâmetro, e no "Passo M" determinamos os parâmetros que maximizam a log-verossimilhança encontrada no "Passo E".

No algoritmo EM para modelos de mistura, os "dados completos" são vistos como $x_i = (y_i, z_i)$, onde $z_i = (z_{i1},...z_{iG})$ é a parte não observada dos dados, e cada z_i é iid segundo uma distribuição multinomial com *G* categorias com probabilidades $\tau_1,...,\tau_G$, como discutida na seção 5.3, e que a densidade de uma observação y_i dado z_i é obtida por $\prod_{k=1}^{G} f_k(y_i \mid \theta_k)^{z_{ik}}$, a log-verossimilhança completa resultante dos dados é

$$l(\theta_k, \tau_k, z_{ik} \mid x) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \log[\tau_k f_k(y_i \mid \theta_k)]. \qquad (6.9)$$

O passo E do algoritmo EM para modelos de mistura é dado por

$$\hat{z}_{ik} = \frac{\hat{\tau}_k f_k(y_i / \hat{\theta}_k)}{\sum_{j=1}^{G} \hat{\tau}_j f_j(y_i / \hat{\theta}_j)}, \qquad (6.10)$$

enquanto no passo M devemos maximizar (7.2) em termos de τ_k e θ_k com z_{ik} fixo nos valores calculados no passo E, $\hat{z}_{ik}$. O valor z_{ik}^* de $\hat{z}_{ik}$ no máximo de (6.3) é a probabilidade condicional estimada de que esta observação pertença ao grupo *K*. A classificação de máxima verossimilhança da observação i é $\{j / z_{ij}^* = \max_k z_{ik}^*\}$, de modo que $\{1 - \max_k z_{ik}^*\}$é uma medida de incerteza na classificação (Bensmail, Ceulex, Raftery e Robert, 1997).

Para misturas de normais multivariadas, o passo E é dado por (7.2) com f_k substituído por ϕ_k como definido em (6.4), independente da parametrização. Para o passo M, estimativas das médias e probabilidades possuem expressões simples de forma fechada envolvendo os dados e $\hat{z}_{ik}$ do passo E,

$$\tau_k = \frac{n_k}{n}\ ; \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} y_i}{n_k}\ ; \qquad n_k = \sum_{i=1}^{n} \hat{z}_{ik}\ .$$

A estimativa de covariância $\hat{\Sigma}_k$ depende de sua parametrização. Detalhes do passo M para Σ_k parametrizado pela decomposição em autovalores (6.5) é dado por Celeux e Govaert (1995).

A estimação EM para modelos de mistura tem um número de limitações. Primeiro, a taxa de convergência pode ser lenta, entretanto EM fornece bons resultados se os dados se ajustam razoavelmente ao modelo e a iteração é iniciada com valores razoáveis. Segundo, o algoritmo EM para modelos de mistura de normais multivariadas falha quando a covariância associada com uma ou mais componentes é singular ou aproximadamente singular. Pode falhar ou dá resultados incorretos se um ou mais grupos contiver poucas observações, ou se as observações que contem estiverem concentradas perto de um subespaço linear de dimensão inferior à dos dados.

7. Seleção de Modelos

Dois problemas básicos surgem na aplicação de análise de agrupamento: a seleção de métodos de agrupamento e a determinação do número de grupos. Na abordagem de modelagem de mistura, estes problemas podem ser reduzidos a um único interesse: a seleção de modelos. Reconhecendo que cada combinação do número de grupos e um modelo de agrupamento correspondem a diferentes modelos estatísticos para os dados, isto reduz o problema de comparação entre os membros de um conjunto de possíveis modelos.

Nossa abordagem para o problema de seleção de modelos é baseada na seleção bayesiana de modelos via fator de Bayes e probabilidade de modelos *a posteriori* (e.g., Kass e Raftery, 1995). A idéia básica é que todos os modelos, $M_1,...,M_k$, sejam considerados com probabilidades *a prior*i $P(M_k)$, $k=1,...,K$ (podem ser iguais), então, pelo teorema de Bayes, a probabilidade *a posteriori* do modelo M_k, dado os dados D, é proporcional à probabilidade dos dados do modelo M_k, vezes a probabilidade *a priori* do modelo,

$$P(M_k \mid D) \propto P(D \mid M_k)P(M_k).$$

Quando existem parâmetros desconhecidos, pela lei da probabilidade total, $P(D \mid M_k)$ é obtida pela integração sobre os parâmetros,

$$P(D \mid M_k) = \int P(D \mid \theta_k)P(\theta_k \mid M_k)d\theta_k$$

onde $P(\theta_k \mid M_k)$ é a distribuição *a priori* de θ_k, o vetor de parâmetros para o modelo M_k. A quantidade $P(D \mid M_k)$ é conhecida como verossimilhança integrada do modelo M_k.

Uma abordagem bayesiana natural para seleção de modelo é escolher o modelo com maior probabilidade *a posteriori*. Se as probabilidades *a priori* dos modelos, $p(M_k)$, são as mesmas, então isto significa escolher o modelo com maior verossimilhança integrada. Para comparar dois modelos $M_1 \, e \, M_2$, o fator de Bayes é definido como a razão de duas verossimilhanças integradas,

$$B_{12} = P(D \mid M_1) / P(D \mid M_2),$$

com a comparação favorecendo M_1 *se* $B_{12} > 1$ e convencionalmente será visto como fornecendo fortes evidências para M_1 *se* $B_{12} > 100$ (Jeffreys, 1961).

Esta abordagem é apropriada no presente contexto, porque isto se aplica quando existe mais do que dois modelos e pode ser usada para comparação de modelos não hierárquicos. Além disso, existe uma solução bayesiana para o problema que possui algumas propriedades freqüentemente desejáveis. Por exemplo, se temos exatamente dois modelos e eles estão hierarquizados, então o modelo escolhido baseado no fator de Bayes minimiza a razão do erro total, que é a soma das razões do erro do tipo I e do tipo II (Jeffreys, 1961).

A principal dificuldade em usar o fator de Bayes é a resolução da integral que define a verossimilhança integrada. A verossimilhança integrada poder ser aproximada simplesmente pelo BIC,

$$2\log p(D \mid M_k) \approx 2\log p(D \mid \hat{\theta}, M_k) - v_k \log(n) = BIC_k ,$$

sendo v_k o número de parâmetros independentes a serem estimados no modelo M_k (Schwarz, 1978; Houghston, 1988).

8. Análise de Agrupamento

A proposta da análise de agrupamento é classificar dados de estrutura previamente desconhecidas dentro de grupos significativos. Nesta seção nós descrevemos uma estratégia para análise de agrupamento baseada em modelos de mistura. Nós usamos a parametrização (6.5) como a base para uma classe de modelos que é flexível para acomodar dados com características amplamente variadas. A estratégia inclui três elementos centrais: inicialização via modelos baseados em agrupamentos hierárquicos aglomerativos, estimação de máxima verossimilhança via algoritmo EM, e seleção de modelos e número de grupos usando o fator de Bayes com aproximação BIC.

8.1 Modelos baseados em agrupamentos hierárquicos

O modelo baseado em agrupamento aglomerativo hierárquico é uma abordagem para calcular uma aproximação para a máxima verossimilhança de classificação,

$$L_{CL}(\theta_1,\ldots,\theta_G;l_1,\ldots,l_n \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta_{l_i}) \qquad (8.1)$$

sendo l_i rótulos que indicam a classificação de cada observação, $l_i = k$ se y_i pertence a *k*-ésima componente. Na mistura de verossimilhança (6.3), cada componente é ponderada pela probabilidade de uma observação pertencer a uma determinada componente. A presença de classes rotuladas na verossimilhança de classificação (8.1) introduz um aspecto combinatorial em que uma maximização exata é pouco prática.

Murtagh e Raftery (1984) aplicaram, com sucesso, modelos baseados em agrupamentos aglomerativos hierárquicos em problemas de reconhecimento de caracteres usando um modelo de normal multivariada parametrizada, como em (6.5), mantendo volume e forma (λ_k e A_k) constantes nos grupos. Esta abordagem foi generalizada por Banfield e Raftery para outros modelos e aplicações, incluindo segmentação de tecidos em imagens médicas.

O procedimento para modelos baseados em agrupamentos aglomerativos hierárquicos é dado por sucessivas uniões de pares de grupos, correspondendo a um aumento máximo na verossimilhança de classificação (8.1) entre todos os possíveis pares. Na falta de alguma informação sobre o agrupamento, o procedimento inicia tratando cada observação como um grupo. Quando o modelo de probabilidade em (8.1) é uma normal multivariada com uma covariância de volume esférico igual a λI, o critério de seleção é conhecido como critério da soma de quadrados.

8.2 Combinando Aglomerados Hierárquicos, EM e Fator de Bayes

Em aglomerados hierárquicos, cada estágio de união corresponde a um número de grupos e uma única partição dos dados. Uma determinada partição pode ser transformada em variáveis indicadoras (seção 6.3), que podem então ser usadas como probabilidades condicionais no passo M da estimação de parâmetro por EM, inicializando o algoritmo EM. Isto, combinado com o fator de Bayes, aproximado pelo BIC, para seleção de modelos, garante uma estratégia de agrupamento:

- Determinar o número máximo de grupos M e o conjunto de modelos de mistura a considerar.
- Executar aglomerações hierárquicas para maximizar aproximadamente a verossimilhança de classificação para cada modelo, e obter as classificações correspondentes para até M grupos.
- Aplicar o algoritmo EM para cada modelo e cada número de grupos 2, ..., M, iniciando com classificações a partir de aglomerações hierárquicas.
- Calcular o BIC para o caso de um grupo para cada modelo e para o modelo de mistura com parâmetros ótimos a partir do algoritmo EM, para 2, ..., M grupos.

Uma forte evidência na escolha de um modelo e o número de grupos associados é o valor dado pelo BIC.

Mistura de normais multivariadas parametrizadas como em (6.5) representa um bom conjunto de modelos para agrupamento em muitas situações encontradas na prática. Com estes modelos, os cálculos feitos por aglomerações hierárquicas podem ser aproveitados para somente um dos modelos (covariância não restrita) usando as partições resultantes como valores iniciais para o EM com qualquer outra parametrização.

8.3 Estimação via EM

Para estimação de parâmetros via algoritmo EM, nos modelos com misturas de distribuições normais e com uma variedade de estruturas de covariância, podem ser vistas no quadro abaixo.

Quadro 1: Parametrizações da matriz de covariância Σ_k para agrupamento hierárquico (HC) e/ou dados multidimensionais para EM. (‘ • ’ indica a utilidade):

Simbolo	Modelo	HC	EM	Distribuição	Volume	Forma	Orientação
E		•	•	Univariada	Igual		
V		•	•	Univariada	Variável		
EII	λI	•	•	Esférico	Igual	Igual	NA
VII	$\lambda_k I$	•	•	Esférico	Variável	Igual	NA
EEI	λA		•	Diagonal	Igual	Igual	Eixos coordenados
VEI	$\lambda_k A$		•	Diagonal	Variável	Igual	Eixos coordenados
EVI	λA_k		•	Diagonal	Igual	Variável	Eixos coordenados
VVI	$\lambda_k A_k$		•	Diagonal	Variável	Variável	Eixos coordenados
EEE	λDAD^T	•	•	Elipsoidal	Igual	Igual	Igual
EEV	$\lambda D_k A D_k^T$		•	Elipsoidal	Igual	Igual	Variável
VEV	$\lambda_k D_k A D_k^T$		•	Elipsoidal	Variável	Igual	Variável
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Elipsoidal	Variável	Variável	Variável

Como ilustração, considere o conjunto de dados Íris. Obtemos como melhor modelo o VEV, que é um modelo com volume e orientação variáveis e forma igual, formando dois grupos ou duas componentes. O agrupamento resultante é mostrado na Figura 19.

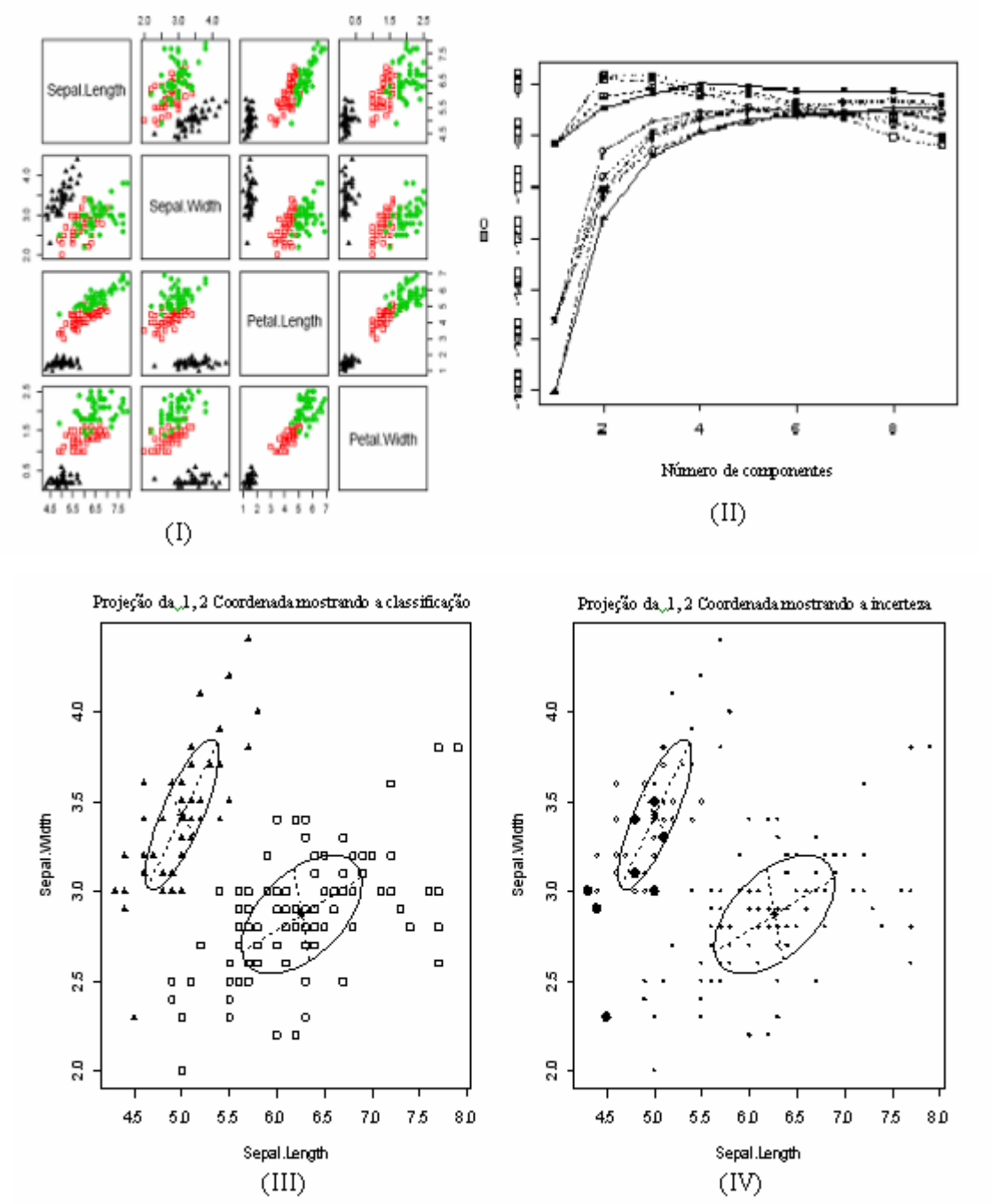


Figura 19: (I) - pares de plot; (II)- BIC; (III)- classificação; (IV) - Incerteza.

Quadro 2: Símbolos usados para representar diferentes modelos e outros significados:

Símbolo	Esférico/Univariado:		Diagonal:		Elipsoidal:
▲	EII/E Variância igual	●	EEI Variância igual	■	EEE Variância igual
△	VII/V não restrita	⊕	EVI Volume igual	⊞	EEV Volume e forma igual
		⊠	VEI Forma igual	⊠	VEV Forma igual
		○	VVI não restrita	▭	VVV não restrita

Também é possível observar os resultados para o mesmo conjunto de dados, mas com diferentes tipos de modelos e/ou diferentes números de componentes. Uma abordagem alternativa é dividir a análise em diferentes partes usando a função *mclustBIC* (Fraley e Raftery, MCLUST, 1999).

Obtivemos como resultado os seguintes modelos e os respectivos números de grupos.

Valores de BIC

Grupos	EII	VII	EEI	VEI	EVI	VVI	EEE	EEV	VEV	VVV
1	-1.804.085	-1.804.085	-1.522.120	-1.522.120	-1.522.120	-1.522.120	-829.978	-829.978	-829.978	-829.978
2	-1.123.412	-1.012.235	-1.042.968	-956.282	-1.007.308	-857.552	-688.097	-644.599	-561.728	-574.018
3	-878.765	-853.814	-813.051	-779.156	-797.836	-744.636	-632.966	-610.085	-562.551	-580.839
4	-784.310	-783.827	-735.482	-716.525	-732.458	-705.069	-591.409	-646.001	-603.927	-628.965
5	-734.386	-746.993	-694.392	-703.052	-695.674	-700.910	-604.929	-621.691	-635.209	-683.821
6	-715.715	-705.781	-693.801	-675.583	-722.152	-696.902	-621.818	-669.719	-681.306	-711.573
7	-712.101	-708.721	-671.676	-666.867	-704.165	-703.992	-617.621	-711.315	-715.210	-728.551
8	-686.097	-707.261	-661.085	-657.245	-703.660	-702.114	-622.422	-750.189	-724.175	-801.729
9	-694.524	-700.022	-678.599	-671.825	-737.311	-727.635	-638.208	-799.641	-810.132	-835.909

Melhores modelos:

VEV,2 (volume e orientação variáveis e forma igual, indicando dois agrupamentos, BIC = -561.7285)

VEV,3 (volume e orientação variáveis e forma igual, indicando três agrupamentos, BIC = -562.5514)

VVV,2 (volume, orientação e forma variáveis, indicando dois grupos, BIC = -574.0178)

O modelo VEV,2 apontado como melhor, não está classificando corretamente pois como conhecemos a verdadeira classificação sabemos que o método deveria indicar 3 grupos.

8.3.1 Adicionando uma *priori* ao modelo

Podemos especificar opcionalmente uma distribuição *a priori* conjugada. Em alguns casos, quando não utilizamos uma informação *a priori*, o *BICplot* mostra um número de picos irregulares, e muitos valores do BIC são *missing* para alguns modelos devido ao insucesso no cálculo do EM, causado pela singularidade e/ou componentes restritas. Com a inclusão de uma distribuição *a priori*, os BIC´s tendem a apresentar curvas mais suaves e existem poucos insucessos no EM. Entretanto, neste caso, especificamente com os dados de Íris, não houve grandes diferenças, ver Figura 20.

Para dados univariados, utilizamos *a priori* normal a seguir

$$\mu | \sigma^2 \sim N(\mu_p, \sigma^2 / k_p)$$

e para variância, uma distribuição *a priori* gama inversa

$$\sigma^2 \sim GI(\nu_p / 2\ , \varsigma_p^2 / 2).$$

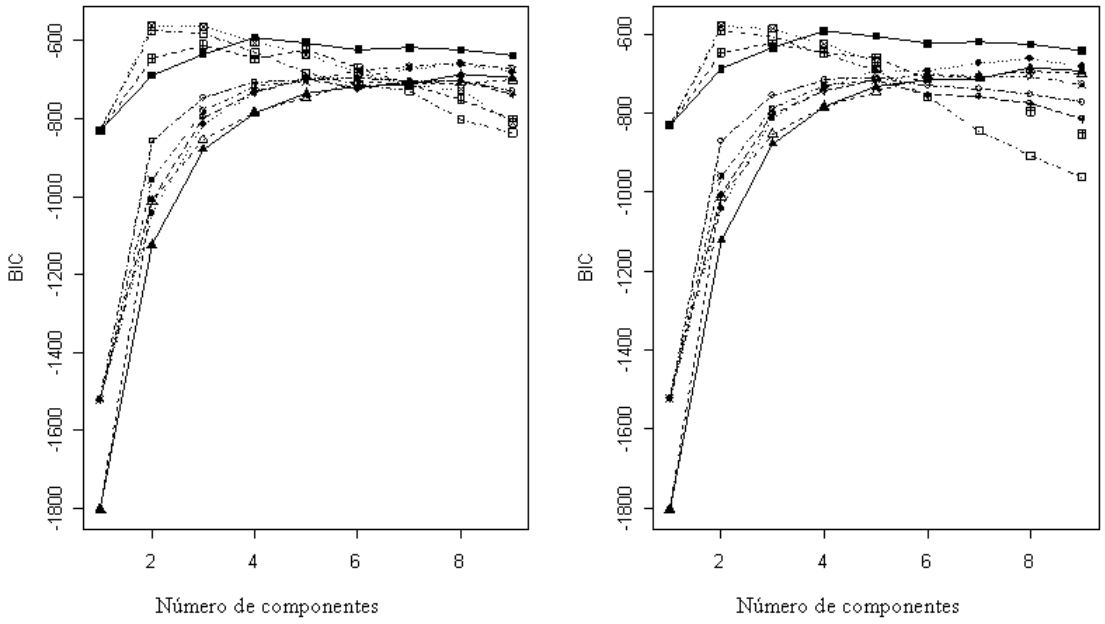
Para dados multivariados, usamos uma distribuição *a priori* normal dada por

$$\mu | \Sigma \sim N(\mu_p, \Sigma / k_p),$$

e uma distribuição *a priori* Wishart inversa para a matriz de covariância

$$\Sigma \sim WI(\nu_p, \Lambda_p).$$

Os hiperparâmetros μ_p, k_p e ν_p são a média, *shrinkage* (recuo) e os graus de liberdade, respectivamente. Os parâmetros ς_p^2 (um escalar) e Λ_p (uma matriz) são os parâmetros de dispersão da distribuição *a priori* nos casos univariados e multivariados, respectivamente. Estas distribuições são chamadas de conjugadas de uma distribuição normal porque a distribuição *a posteriori* pode ser expressa como um produto de uma distribuição normal e uma gama inversa ou distribuição Wishart.



(I) (II)

Figura. 20: (I) - Sem o acréscimo de uma distribuição *a priori*; (II) – Acrescentando a de uma distribuição *a priori*.

Como conhecemos a verdadeira classificação, a incerteza relativa das observações classificadas incorretamente pode ser mostrada, como feito abaixo para o exemplo íris (ver Figura 21).

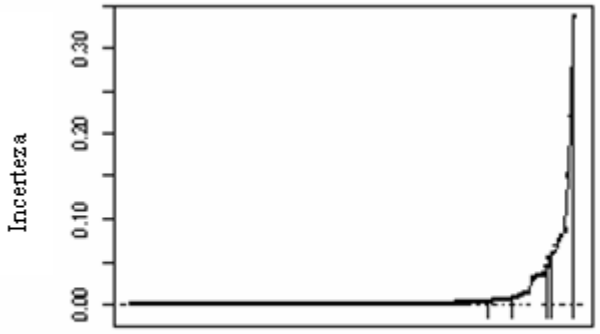


Figura 21: Número de observações em ordem de incerteza.

Figura 21: Gráfico de incerteza para 3-grupos do modelo de mistura ajustado ao conjunto de dados Íris, via EM, baseado em uma mistura gaussiana não restrita. As linhas verticais indicam as observações classificadas incorretamente. O gráfico foi criado com a função *uncerplot*, e mostra a incerteza relativa das observações classificadas incorretamente.

- Modelo baseado em agrupamento Hierárquico

Usando o agrupamento Hierárquico para o mesmo conjunto de dados, encontramos as seguintes observações classificadas incorretamente:

Tabela 11: Mostra as observações classificadas incorretamente.

Observações classificadas incorretamente												
102	107	114	115	120	122	124	128	134	139	143	147	150

Considerando que temos 150 observações e que 14 foram classificadas incorretamente, temos uma taxa de classificação incorreta de aproximadamente 9,3%.

9. Vantagens e desvantagens de cada método

Os métodos de agrupamento atendem a diferentes tipos de requisitos, como encontrar ou não um número adequado de grupos, ser capaz ou não de desprezar ruídos, identificar grupos de tamanhos variados, fornecer resultados interpretáveis e utilizáveis e ainda apresentar o resultado em um tempo satisfatório.

Como não há um método que atenda a todos esses requisitos é importante que ao utilizarmos um desses métodos, analisemos quais dos aspectos citados anteriormente ele satisfaz e assim escolher o que mais se adapta as nossas necessidades.

Os métodos hierárquicos, apesar de serem amplamente empregados com sucesso em aplicações biológicas, como na taxonomia de animais e plantas, existem deficiências inerentes a estes métodos. Neles não existe uma revisão dos agrupamentos durante a execução do procedimento, ou seja, no método hierárquico aglomerativo, uma vez realizado a fusão de dois objetos dentro de um mesmo grupo, os objetos não são mais separados, permanecendo até o final do procedimento, sempre juntos em um mesmo grupo. De forma análoga, no hierárquico divisivo, uma vez separados dois objetos, eles não são mais agrupados em um mesmo grupo (Ng e Han, 1994). É dispendioso computacionalmente e raramente utilizado quando existe um grande número de objetos a serem agrupados, por exemplo, em sensoriamento remoto, esse método é praticamente

descartado pelo grande número de objetos. Este é o caso em processamento de imagens onde o número de pixels gerado é muito elevado.

Os métodos de particionamento, discutido no Capitulo 4, buscam encontrar, iterativamente, a melhor partição dos *n* objetos em *k* grupos. Freqüentemente os *k* grupos encontrados pelos métodos de particionamento são de melhor qualidade (grupos internamente mais homogêneos) do que os *k* grupos produzidos pelos métodos hierárquicos. Devido a este melhor desempenho, os algoritmos de particionamento têm sido mais investigados e utilizados. Os métodos de particionamento mais utilizados são baseados em um ponto central (média dos atributos dos objetos - *k*-médias) (Zhang *et al*., 2001) ou em um objeto representativo para o grupo (*k*-medoides), (Kaufman e Rousseeuw, 1990). Apesar do método de *k*-médias ser bastante utilizado é válido ressaltar que ele é aleatório e, por isso, a qualidade de seus resultados depende muito dos *k* centros escolhidos inicialmente. Uma solução seria executar o algoritmo várias vezes em busca de uma melhor solução. Além disso, o *k*-médias é muito sensível a ruídos, visto que um objetivo bem distante dos demais em certo grupo pode afetar o cálculo do centro de gravidade do grupo a que ele pertence.

Ng e Han (1994) afirmam que os métodos baseados em *k*-medoides apresentam duas vantagens em relação aos métodos baseados em centros médios. A primeira vantagem está relacionada a sua robustez quanto à presença de *outliers* (objetos fora do padrão dos demais objetos). Outra vantagem é que são independentes da ordem na qual os objetos são examinados.

O método de agrupamento PAM (*Particioning Around Medoids*), baseado em *k*-medoides, verifica todas as trocas possíveis entre um medoide e um objeto não selecionado, escolhendo um novo conjunto de *k*-medoides. Devido à procura exaustiva, pelo conjunto *k*-medoides que minimiza as dissimilaridades, o PAM não é eficiente, principalmente quando aplicado a grandes volumes de dados. Para cada medoide, são investigadas (*n*-*k*) possibilidades de troca; considerando todos os medoides, temos um total de *k*(*n*-*k*) trocas. Como exemplo, considere 1000 objetos e 10 grupos, neste caso seriam avaliadas 9.900 trocas em cada iteração do algoritmo. Assim como o método k-médias, seu resultado ainda é aleatório, pois também dependem dos k centros escolhidos inicialmente. Tendo a necessidade de ser executado várias vezes para aumentarmos a chance de conseguirmos obter grupos de boa qualidade, isso se agrava ainda mais no

PAM, pois, por sempre percorrermos todos os objetos de cada grupo em busca daquele que provocaria maior redução do custo total caso fosse eleito como seu medoide, computacionalmente pode se tornar muito custoso.

Com relação ao método SOM, sabemos que, por se tratar de um tipo de rede neural, ele possui uma importante característica, a sua capacidade de aprender a partir de estímulos fornecido pelo meio ambiente. A versatilidade e robustez do SOM também trazem alguns problemas posteriormente. O SOM contém muito mais parâmetros do que, por exemplo, agrupamentos hierárquicos. Além disso, o SOM é um algoritmo matematicamente complicado e algumas das propriedades teóricas ainda permanecem sem prova (Kohonen, 2001). Portanto o SOM é mais difícil de ser aplicado em dados de *microarray* que os agrupamentos hierárquicos. Além do que, amplamente usado na visualização para amostragem de vetores de referência, faz a identificação de correlações entre amostras difíceis. Essas duas questões podem explicar porque o SOM não tem sido largamente aceito como uma ferramenta de agrupamento padrão para análises de dados de *microarray*. Também deve ser notado que se o conjunto de dados consiste em centenas de amostras, a representação plana pode ser difícil de ser interpretada.

A seguir mostramos uma breve comparação entre os Modelos de Mistura e o k-médias. O agrupamento baseado em Modelos de Mistura pode ser visto como uma generalização do método *k*-médias.

K-médias	Modelos de Mistura
Atribuição determinística dos objetos aos grupos	Atribuição probabilística dos objetos aos grupos
Variáveis quantitativas	Variáveis de qualquer tipo
Grupos de forma esférica	Independência condicional
Minimização (maximização) da variação dentro (entre) grupos	Maximização de máxima verossimilhança
Critérios para determinação do número de grupos não são objetivos	Há vários diagnósticos que ajudam a decidir sobre o número de grupos

A vantagem dos Modelos de Mistura sobre os métodos tradicionais é ser mais flexível e ser baseado num modelo estatístico.

10. Dados Artificiais

Neste capitulo analisaremos dados gerados, que tentam se aproximar das características encontradas, na prática, em dados de *microarray*. Assim podermos julgar os métodos de acordo com o poder de classificação de cada um. Foram gerados 3 matrizes (M_1, M_2, e M_3), cada matriz é formada por três grupos. Estes grupos foram gerados por distribuições normais (função *rnorm* do software R), diferenciando média e variância como é mostrado à abaixo:

Tabela 10.1: Média e desvio padrão (dp) dos dados gerados.

	n	**média**	**dp**	**n**	**média**	**dp**	**n**	**média**	**dp**
Dados 1	20	0,13	0,53	20	0,16	0,80	20	0,32	0,51
Dados 2	200	0,13	0,53	200	0,16	0,80	200	0,32	0,51
Dados 3	1000	0,13	0,53	1000	0,16	0,80	1000	0,32	0,51
Dados 4	60	0,13	0,53	120	0,16	0,80	420	0,32	0,51
Dados 5	100	0,13	0,53	200	0,16	0,80	200	0,32	0,51

1. Aplicação dos métodos para os Dados 1:

- Aplicando o método SOM.

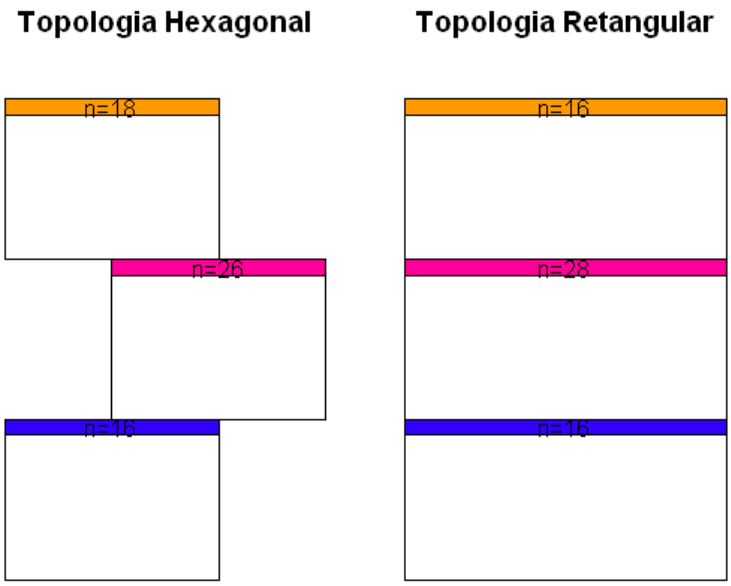


Figura 10.1 Resultado do agrupamento da Matriz de Dados 1 obtidos pelo método SOM.

Tabela 10.1. Quantidade de observações em cada grupo, pelo método SOM.

Grupos	Topologia hexagonal	Topologia retangular
1	18	16
2	26	28
3	16	16

- Aplicando o método Agnes.

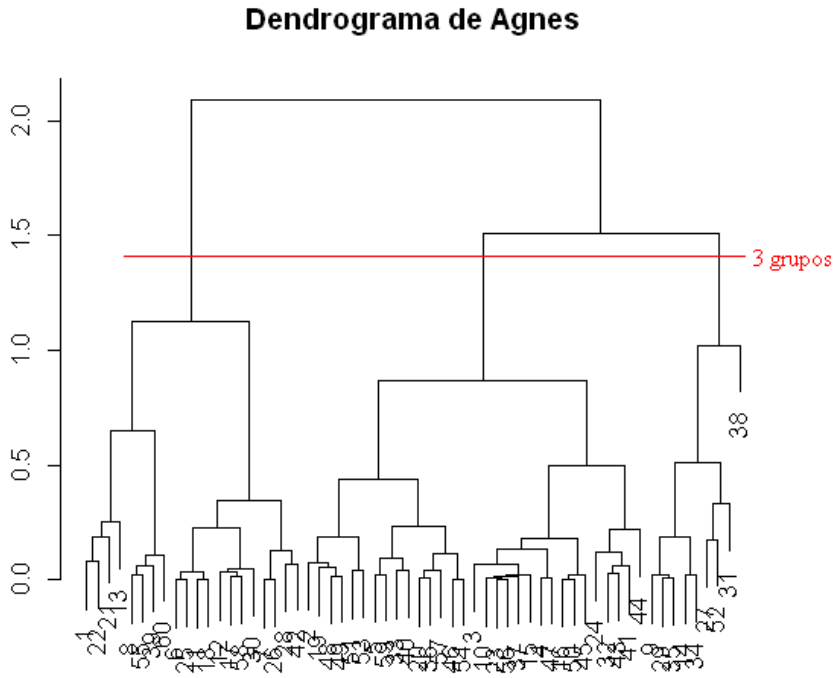


Gráfico 10.2. Dendrograma obtido pelo método Agnes.

Tabela 10.2. Quantidade de observações em cada grupo, pelo método Agnes.

Grupos	Qtde de Obs
1	20
2	31
3	9
Total	60

- Aplicando o método DIANA

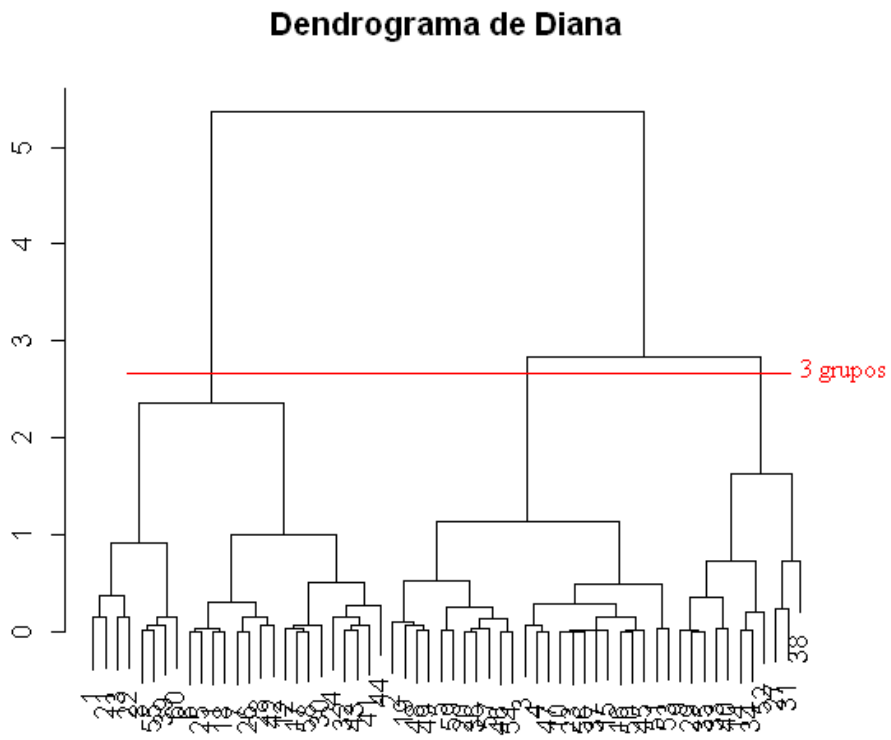


Gráfico 10.3. Dendrograma obtido pelo método DIANA.

Tabela 10.3. Quantidade de observações em cada grupo, pelo método Diana.

Grupos	Qtde de Obs
1	25
2	24
3	11
Total	60

- Usando o método K- médias.

Tabela 10.4. Quantidade de observações em cada grupo, pelo método K-médias.

Grupos	Qtde de Obs
1	22
2	18
3	20
Total	60

Tabela 10.5. Centro de Grupos.

Grupos	Centro de Grupos
1	0,49
2	1,17
3	-0,26

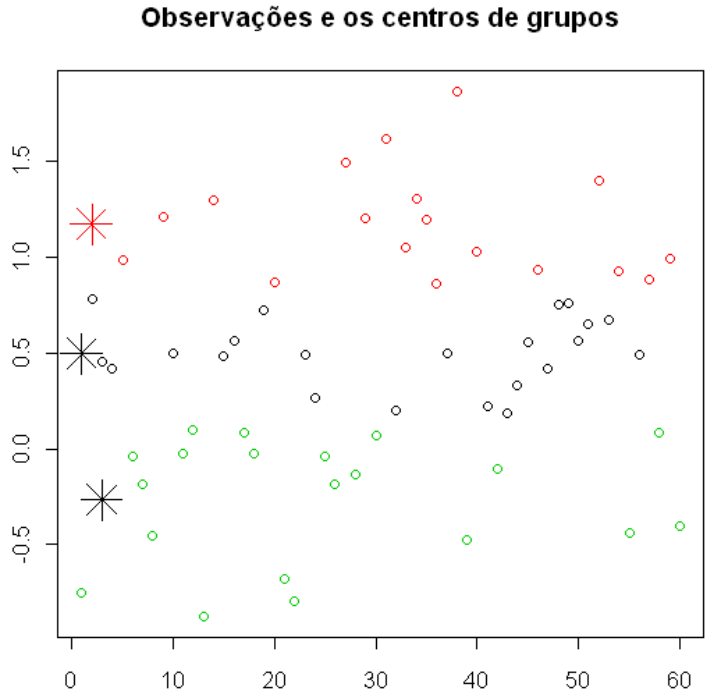


Gráfico 10.4. Gráfico com as observações e os centros de grupos. Cada cor representa um agrupamento

- Usando modelos de mistura

Modelo escolhido: VII com 3 componentes. Neste caso o melhor modelo é um modelo com forma igual, volume variáveis com 3 componentes ou grupos. Parametrização da matriz de covariância: $\lambda_k I$.

Tabela 10.6. Quantidade de observações em cada grupo, pelo modelo de misturas.

Grupos	Qtde de Obs
1	25
2	24
3	11
Total	60

- Conclusão Dados 1:

Para os Dados 1 o método que melhor classificou as observações foi o k-médias, em seguida melhor classificação foi obtida pelo método SOM. Para este tamanho, n=20, o fator tempo não foi significativo para escolha do método, ou seja, todos executaram em tempos semelhantes.

2. Aplicação dos métodos para os Dados 2:

- Aplicando o método SOM.

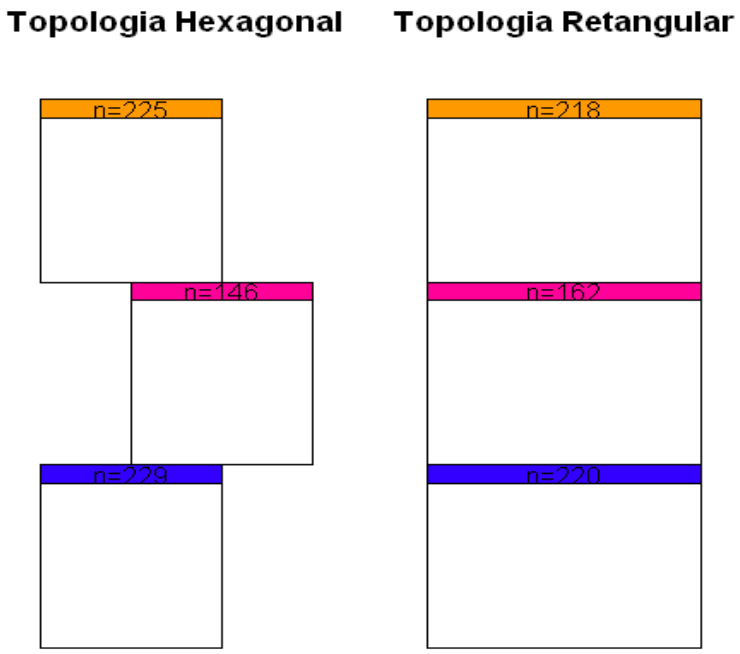


Gráfico 10.5: Resultado do agrupamento da Matriz de Dados 2 obtidos pelo método SOM.

Tabela 10.7. Quantidade de observações em cada grupo, pelo método SOM.

Grupos	Topologia hexagonal	Topologia retangular
1	225	218
2	146	162
3	229	220

- Utilizando o método Agnes

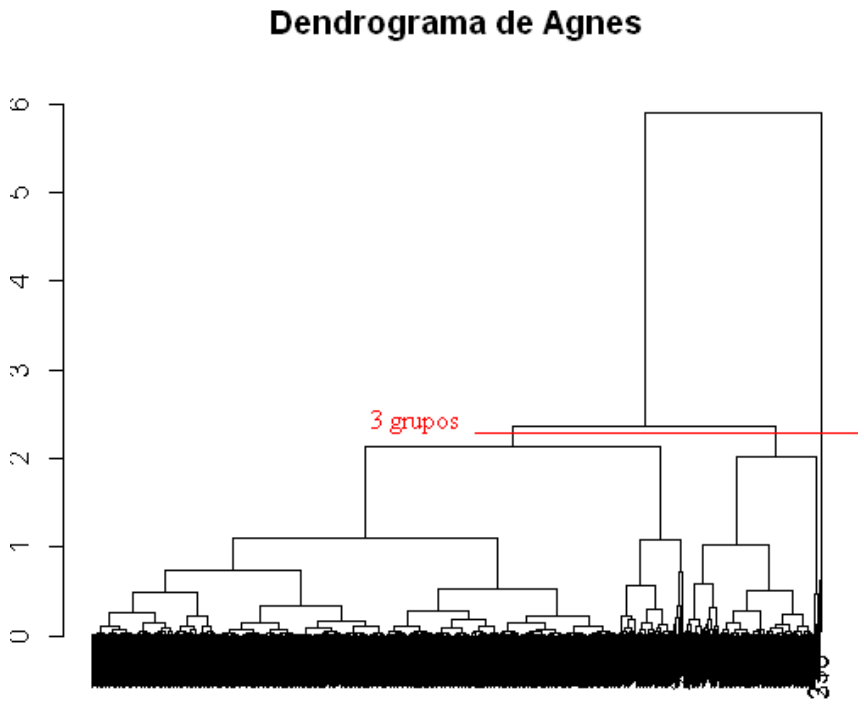


Gráfico 10.6. Dendrograma obtido pelo método Agnes.

Tabela 10.8. Quantidade de observações em cada grupo, pelo método Agnes.

Grupos	Qtde de Obs
1	486
2	112
3	2
Total	600

- Aplicando o método DIANA.

Tabela 10.9. Quantidade de observações em cada grupo, pelo método Diana.

Grupos	Qtde de Obs
1	245
2	78
3	277
Total	600

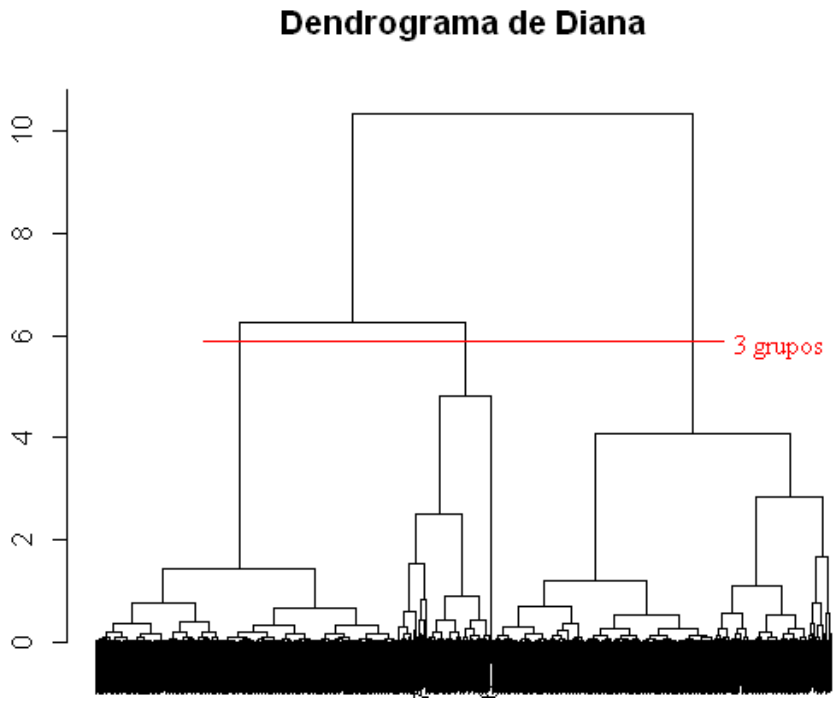


Gráfico 10.8: Dendrograma obtido pelo método Diana.

- Aplicando o k-médias.

Tabela 10.10: Quantidade de observações em cada grupo, pelo método k-médias.

Grupos	Qtde de Obs
1	123
2	124
3	353
Total	600

Tabela 10.11: Centro de grupo.

Grupos	Centro de Grupos
1	-0,62
2	0,23
3	1,13

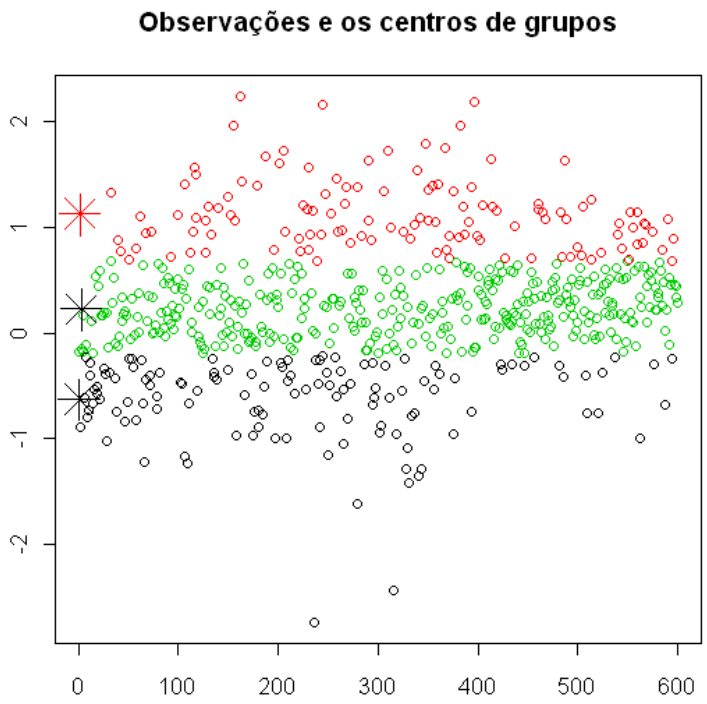


Gráfico 10.9: Gráfico com as observações e os centros de grupos. Cada cor representa um agrupamento

- Usando modelos de mistura.

Tabela 10.12: Quantidade de observações em cada grupo, usando modelos de Mistura.

Grupos	Qtde de Obs
1	295
2	139
3	166
Total	600

- Conclusão Dados 2:

Para os Dados 2 o método que melhor classificou as observações foi o SOM, em seguida melhor classificação foi obtida pelo Modelo de Misturas. Para este tamanho, n=200, o fator tempo também não foi significativo para escolha do método, ou seja, todos executaram em tempos semelhantes.

1. Usando o agrupamento *K*-médias.

Usando o k-médias com três grupos não obtivemos uma classificação satisfatória, pois deveríamos obter como resultado aproximadamente 20 observações em cada grupo.

Tabela 10.1 Centro de grupos.

Grupo		
1	2	3
0,6927	26,096	15,6453

Tabela 10.2. Quantidade de observações em cada grupo.

K-médias		
Gupo 1	Grupo 2	Grupo 3
38	13	9

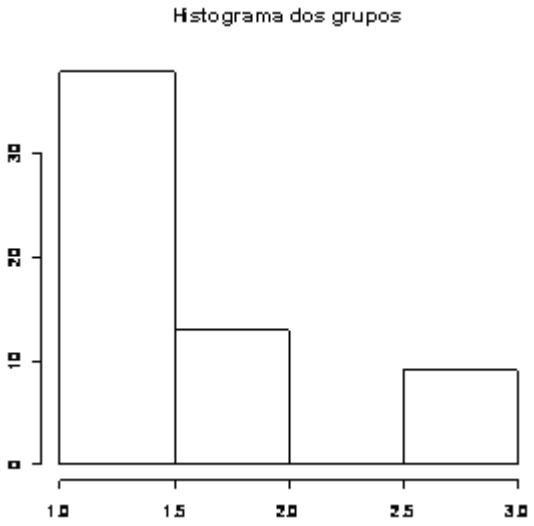


Figura 10.2: Histograma das observações agrupadas por K - médias.

Tabela 10.3 Observações classificadas incorretamente.

Grupo 1	Grupo 2	Grupo 3
18	5	2

Considerando que temos 60 observações e que 25 foram classificadas incorretamente, temos uma taxa de classificação incorreta de aproximadamente 42%.

2. Aplicando o SOM

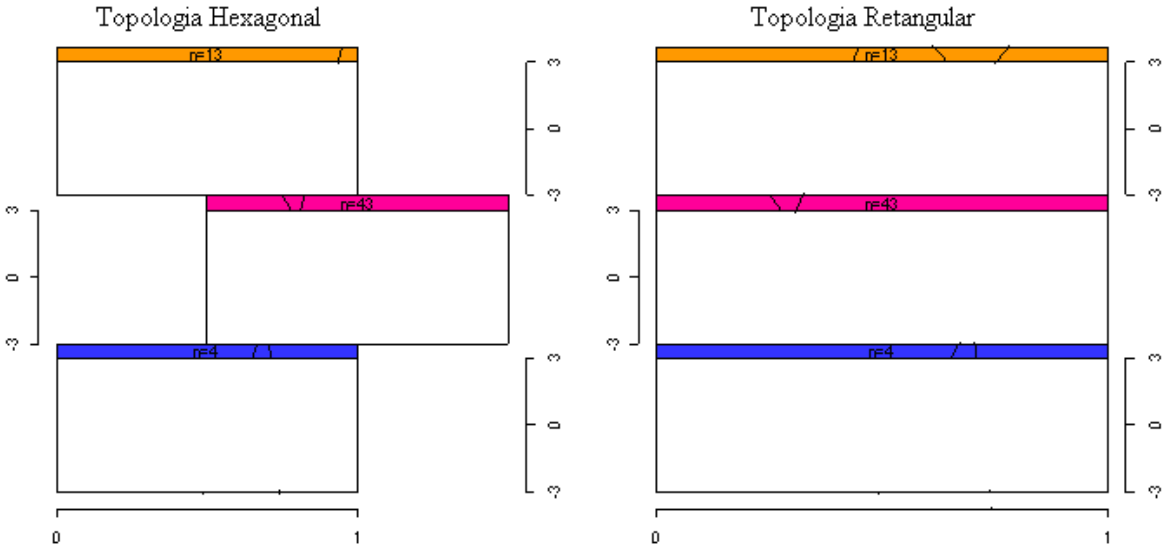


Figura 10.3: Agrupamento SOM com topologia hexagonal e retangular.

Tabela 10.4. Quantidade de observações em cada grupo, pelo método SOM.

Grupos	Topologia hexagonal	Topologia retangular
1	4	4
2	43	43
3	13	13

Na Tabela 10.4, é apresentado um resumo da Figura 10.3. O agrupamento não muda com a topologia, continua com uma má distribuição das observações. Temos o grupo 2 concentrando quase 72% das observações.

3. Aplicando o PAM

Na Tabela 10.5 temos a distribuição das observações nos grupos após a execução do SOM.

Tabela 10.5. Quantidade de observações em cada grupo, pelo método PAM.

PAM		
Gupo 1	Grupo 2	Grupo 3
31	19	10

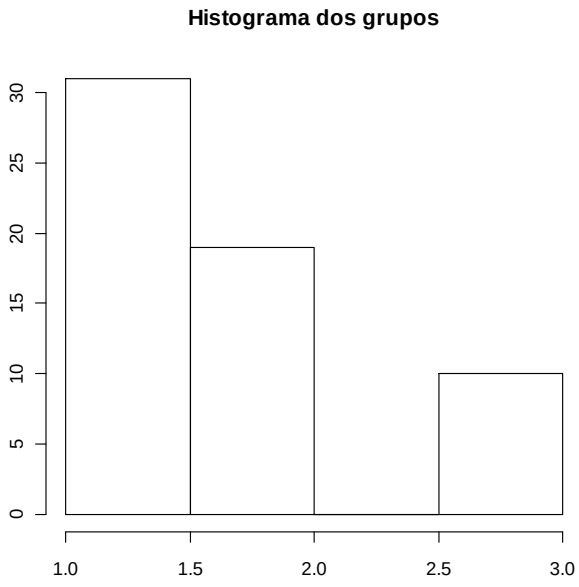


Figura 10.4: Histograma das observações agrupadas por PAM.

Nesse caso também não há uma boa classificação, tendo uma maior concentração no grupo 1.

4. Aplicando o Agnes

Na Tabela 10.6 temos a distribuição das observações nos grupos após a execução do Agnes.

Tabela 10.6. Quantidade de observações em cada grupo, pelo método AGNES.

Grupos	Qtde de Obs
1	20
2	25
3	15
Total	60

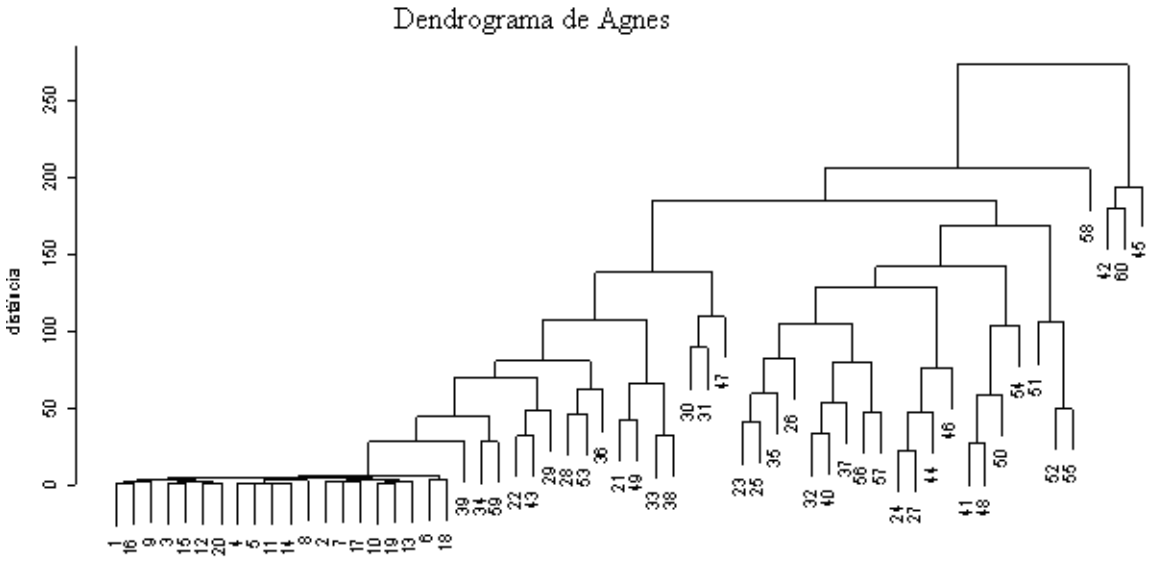


Figura 10.5: Dendrograma das observações agrupadas por Agnes.

5. Usando Modelos de Mistura.

Encontramos como melhor modelo o VII. É um modelo com volume variável e forma igual, formando dois grupos ou duas componentes. O agrupamento resultante é mostrado na Tabela 10.7.

Tabela 10.7. Quantidade de observações em cada grupo, pelo Modelo de Misturas

Grupos	Qtde de Obs
1	20
2	40
Total	60

- Usando o valor do BIC para selecionar os melhores modelos.

Tabela 10.8. Valores do BIC para os melhores modelos.

VII,2	VEI,2	VVI,2
-2399,175	-2404,095	-2420,045

No primeiro grupo todas as observações foram classificadas corretamente, entretanto as observações que deveriam formar dois grupos formaram apenas 1.

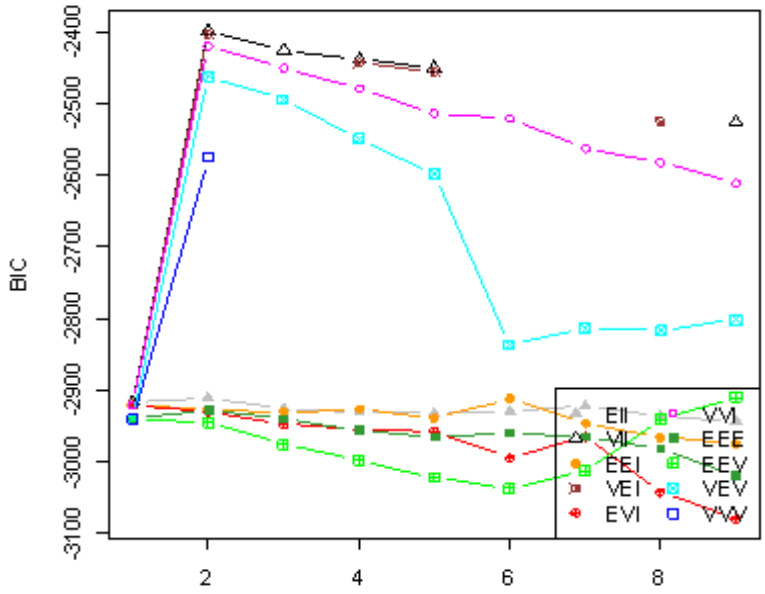


Tabela 10.6. Valores do BIC para os melhores modelos.

A seguir, apresentaremos os resultados obtidos pelos mesmos métodos, porém com uma amostra de tamanho maior. A matriz M_2.

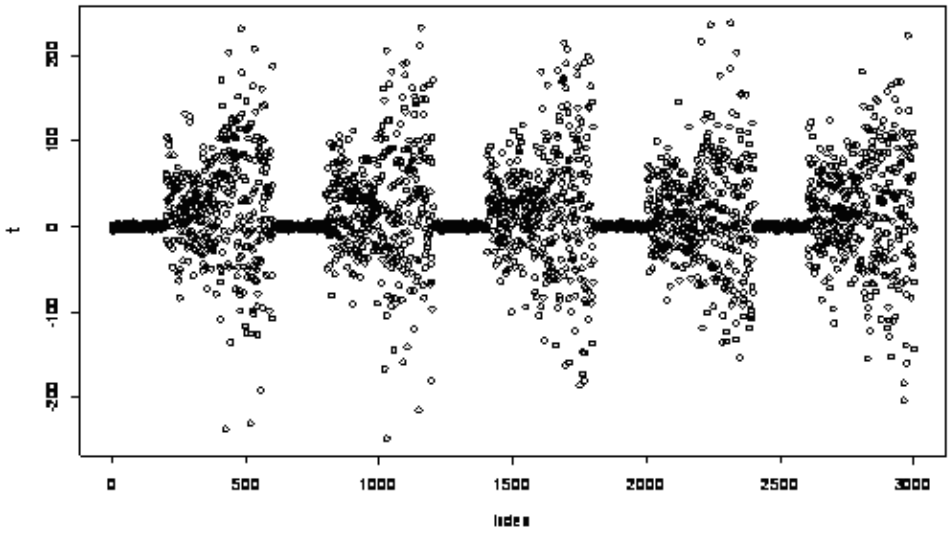


Figura 10.7: Gráfico das 600 observações pertencentes aos três grupos. Cada grupo com 200 observações. Este gráfico também não sugere nenhuma forma de agrupamento.

1. Usando o agrupamento *K*-médias.

Usando o *k*-médias com três grupos não obtivemos uma classificação satisfatória, pois deveríamos obter como resultado aproximadamente 200 observações em cada grupo.

Tabela 10.9 Classificação obtida por k-médias

Grupo		
1	2	3
121	368	111

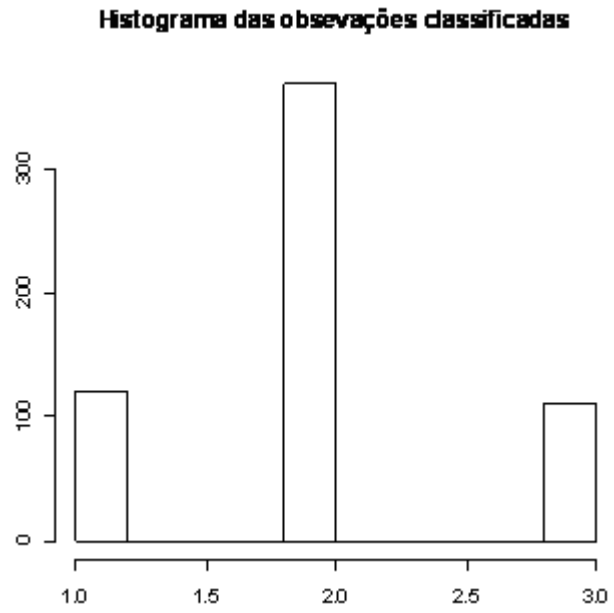


Figura 10.8: Histograma das observações agrupadas por K - médias.

2. Usando o método SOM

Tabela 10.10 Classificação obtida pelo método SOM

Grupos	Topologia hexagonal	Topologia retangular
1	203	191
2	69	87
3	328	322

Na Tabela 10.10, é apresentado um resumo da Figura 10.9. O agrupamento muda muito pouco com a mudança de topologia, continua com uma má distribuição das observações. O grupo 3 foi o que mais concentrou observações.

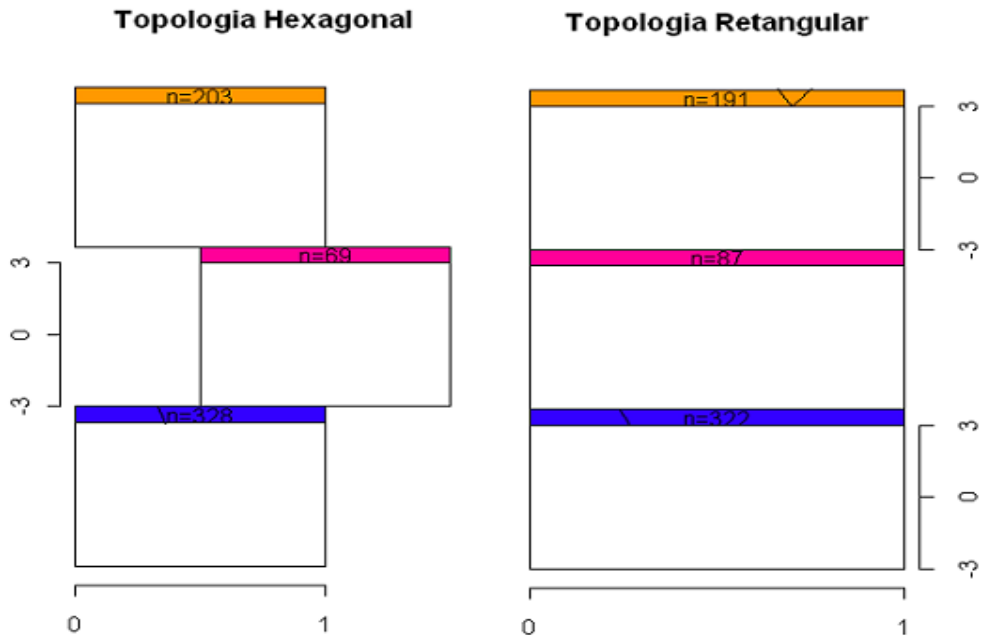


Figura 10.9: Mapa obtido pelo agrupamento SOM

3. Aplicando o PAM

Tabela 10.11 Classificacão obtida pelo método PAM

PAM		
Gupo 1	Grupo 2	Grupo 3
230	210	160

Os métodos Agnes e Diana tem como resultado uma classificação parecida com o método PAM. Se compararmos os resultados de M2 com os apresentados para M1 obtém-se melhores resultados, mas ainda não são satisfatórios.

Tabela 10.12 Classificacão obtida

Gupo 1	Grupo 2	Grupo 3
201	176	223

- Usando os valores de BIC para selecionar os melhores modelos

Tabela 10.13. Valores do BIC para os melhores modelos.

VII,3	VEI,3	VII,4
-27248,31	-27270,19	-27289,07

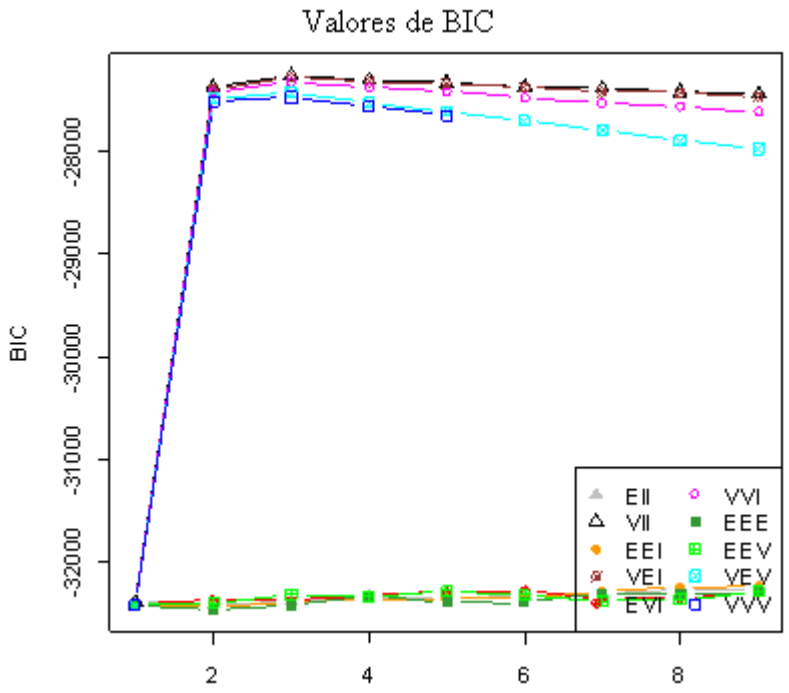


Figura 10.10: Gráfico com os valores de BIC para os modelos propostos.

Ao adicionar a priori, não observamos nenhuma diferença considerável. O modelo eleito continua o mesmo, inclusive as mesmas quantidades de observações em cada agrupamento

- Aplicação dos métodos para a matriz M_3.

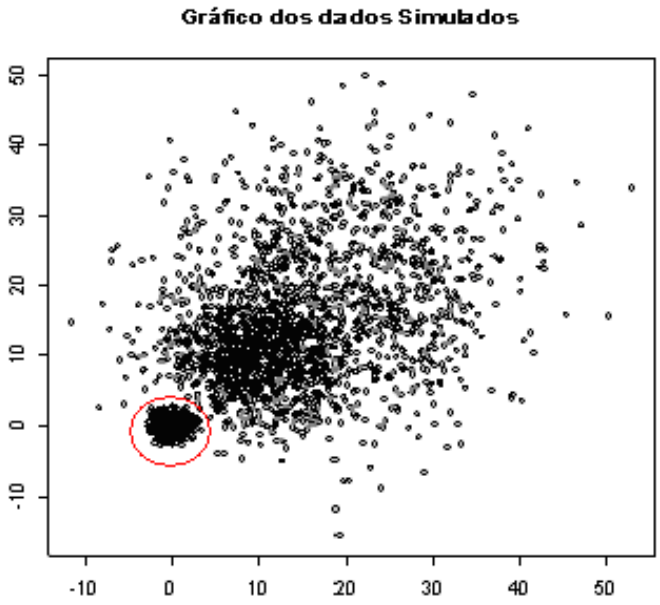


Figura 10.11: Gráfico das 3000 observações pertencentes aos três grupos.

Note que na Figura 10.11 podemos visualizar uma aglomeração de observações, dando a idéia de um primeiro agrupamento, as demais observações aparecem embaralhadas no gráfico.

- Usando o agrupamento *K*-médias.

Utilizamos o agrupamento K-médias com 3 grupos e obtivemos uma má classificação das observações. Houve uma concentração no grupo 3, como podemos observar na Tabela 10.15.

Tabela 10.14 Centro de grupos.

Grupo		
1	2	3
0,48575	23,31622	10,72009

Tabela 10.15. Quantidade de observações em cada grupo.

Grupos	Qtde de Obs
1	1075
2	716
3	1209
Total	3000

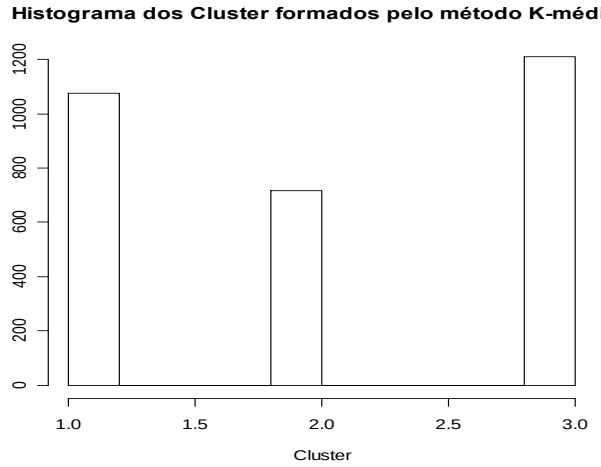


Figura 10.12: Histograma das observações agrupadas por K - médias.

Note que o grupo três absorveu o maior número de observações. Deveríamos obter como resultado aproximadamente 1.000 observações, em cada grupo. No agrupamento formado pelo algoritmo obtivemos 363 observações classificadas incorretamente, ver Tabela 10.16.

Tabela 10.16. Observações classificadas incorretamente.

Grupo 1	Grupo 2	Grupo 3
75	6	282

Considerando que temos 3000 observações e que 363 foram classificadas incorretamente, temos uma taxa de classificação incorreta de aproximadamente 12,1%.

- Aplicando o algoritmo SOM para agrupar os mesmos dados gerados acima.

Na Tabela 10.17, é apresentado um resumo da Figura 10.13. Quando consideramos a topologia do Mapa observamos uma sensível melhora no agrupamento para a topologia retangular, neste mapa as observações estão melhores distribuídas.

Tabela 10.17. Quantidade de observações em cada grupo, pelo método SOM.

Grupos	Topologia hexagonal	Topologia retangular
1	1025	991
2	650	730
3	1325	1279

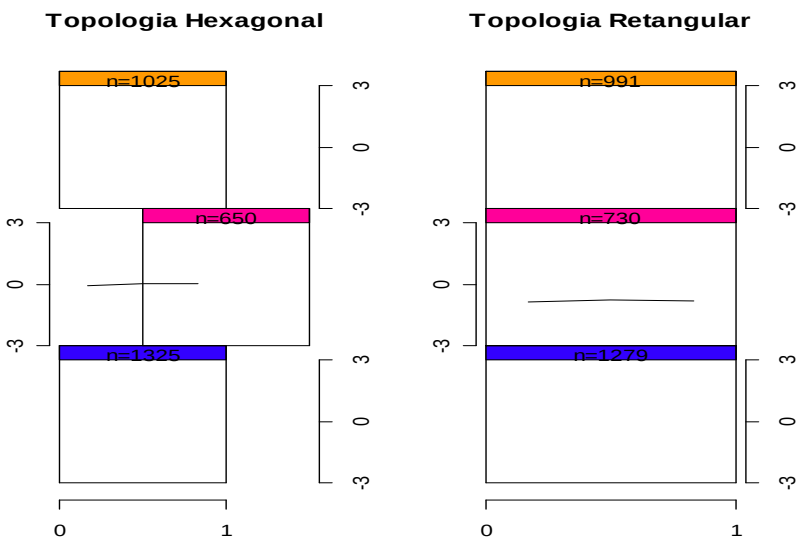


Figura 10.13: Agrupamento SOM com topologia hexagonal e retangular.

- Aplicando o algoritmo PAM.

Tabela 10.18. Quantidade de observações em cada grupo, pelo método PAM.

Grupos	Qtde de Obs
1	1075
2	1220
3	705
Total	3000

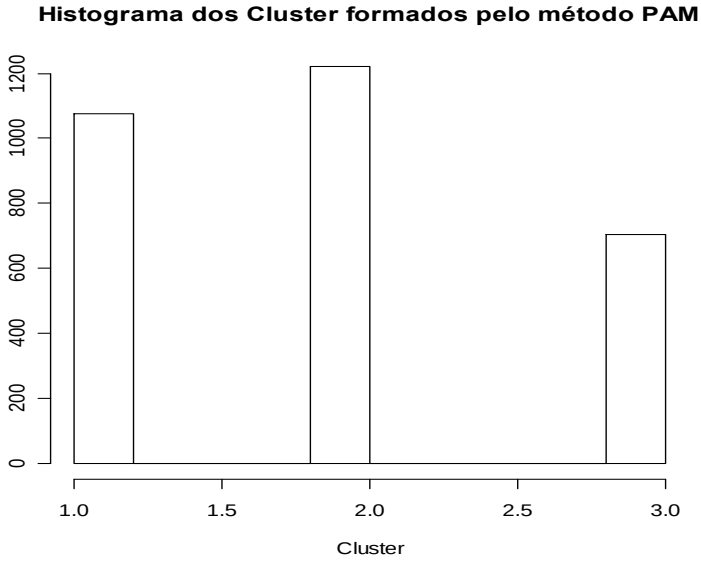


Figura 10.14: Histograma das observações agrupadas por PAM.

- Aplicando o algoritmo DIANA.

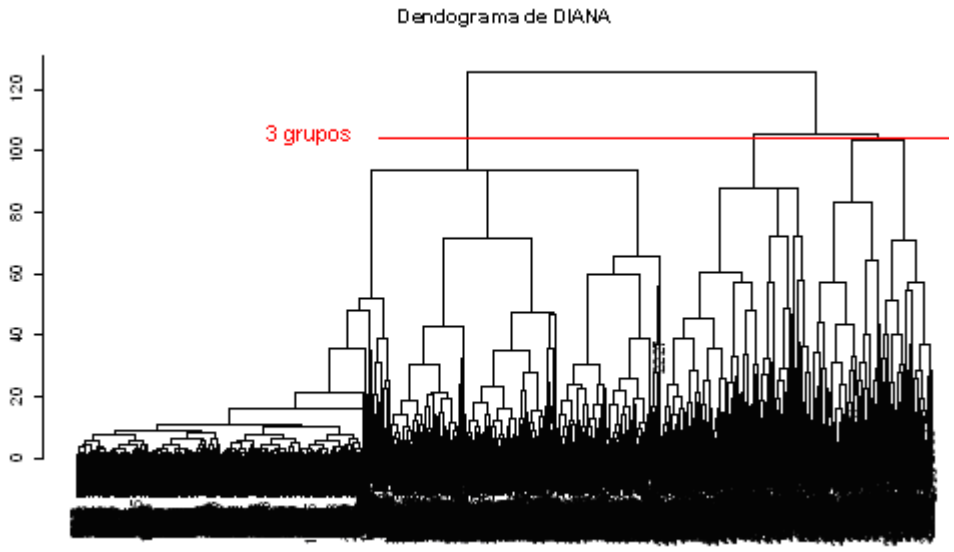


Figura 10.15: Dendrograma formado pelo método DIANA.

Fazendo um corte no dendrograma, formando 3 grupos, podemos avaliar a classificação feita pelo método na Tabela 10.19.

Tabela 10.19. Quantidade de observações em cada grupo, pelo método DIANA.

Grupos	Qtde de Obs
1	2048
2	529
3	423
Total	3000

- Aplicando o algoritmo Agnes.

Da mesma forma que foi feita no método DIANA fizemos para o Agnes. Podemos avaliar a classificação feita pelo método na Tabela 10.20.

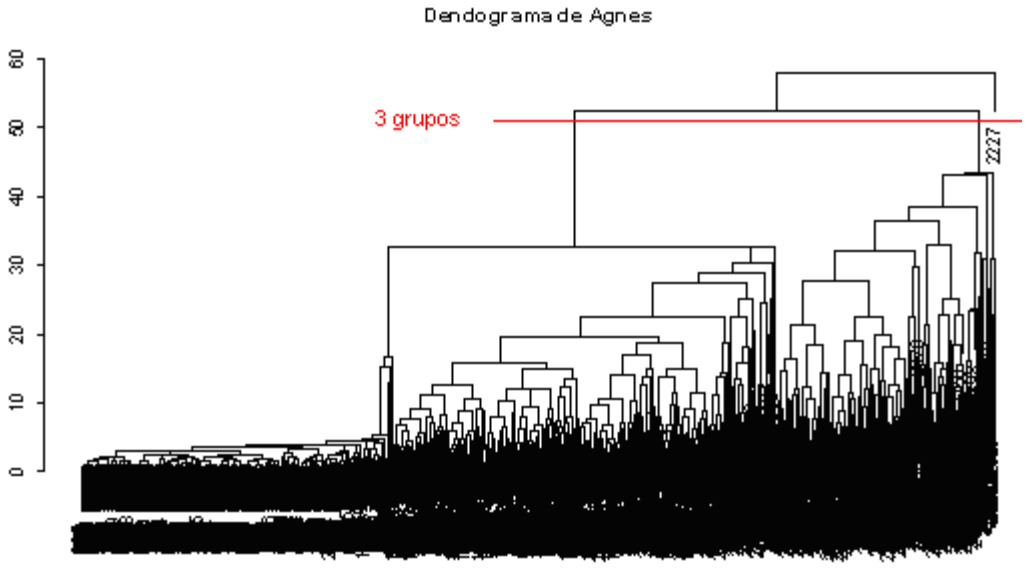


Figura 10.16: Dendrograma formado pelo método Agnes.

Tabela 10.20. Quantidade de observações em cada grupo, pelo método Agnes (average).

Grupos	Qtde de Obs
1	2275
2	724
3	1
Total	3000

Tabela 10.21. Quantidade de observações em cada grupo, pelo método Agnes (single).

Grupos	Qtde de Obs
1	2998
2	1
3	1
Total	3000

- Aplicando Modelos de Mistura para agrupar os dados.

Obtemos como melhor modelo o VII, é um modelo com volume variável e forma igual, formando três grupos ou três componentes. O agrupamento resultante é mostrado na Figura 10.22.

Tabela 10.22. Quantidade em cada grupo, usando o modelo de misturas de normais.

Grupos	Qtde de Obs
1	1003
2	948
3	1049
Total	3000

Usando o valor do BIC para selecionar os melhores modelos.

Tabela 10.23. Valores do BIC para os melhores modelos.

VII,3	VEI,3	VII,4
-54829,53	-54843,78	-54863,4

Tabela 10.24. Quantidade em cada grupo, usando o modelo de misturas de normais.

Grupos	Qtde de Obs
1	1003
2	938
3	1059
Total	3000

Tabela 10.25. Quantidade de observações classificadas incorretamente em cada grupo.

Grupo 1	Grupo 2	Grupo 3
3	60	112

Considerando que temos 3000 observações e que 175 foram classificadas incorretamente, temos uma taxa de classificação incorreta de aproximadamente 5,8%.

Depois de aplicarmos os mesmos modelos nas matrizes simuladas chegamos a conclusão que os Modelos que usam misturas de distribuições, neste caso misturas de normais, para agrupamento fornece um melhor resultado na maioria das vezes. Portanto utilizaremos este para agrupamento dos dados reais analisados neste trabalho.

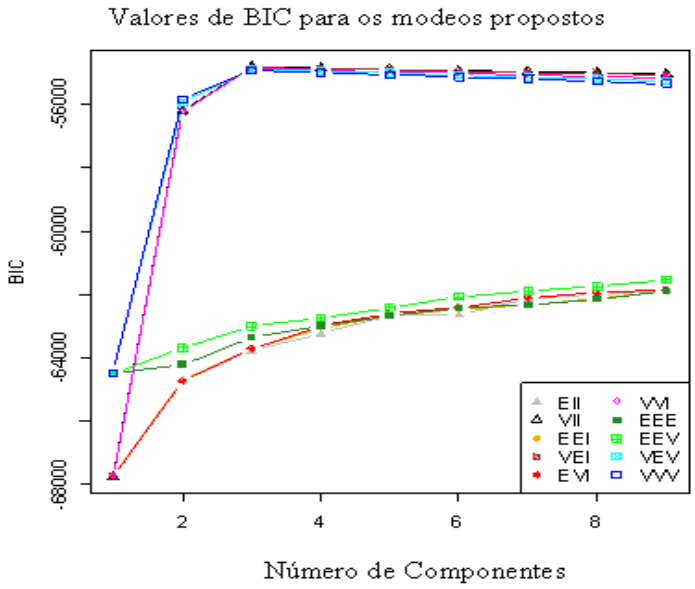


Figura 10.17: Gráfico com os valores de BIC.

Adicionando a priori, não observamos nenhuma diferença considerável, ver Tabela 10.26. O modelo eleito continua o mesmo, inclusive as mesmas quantidades de observações em cada agrupamento. Podemos ver os dois gráficos na Figura 10.18.

Tabela 10.26. Valores do BIC para os melhores modelos.

VII,3	VEI,3	VII,4
-54830,06	-54847,45	-54863,05

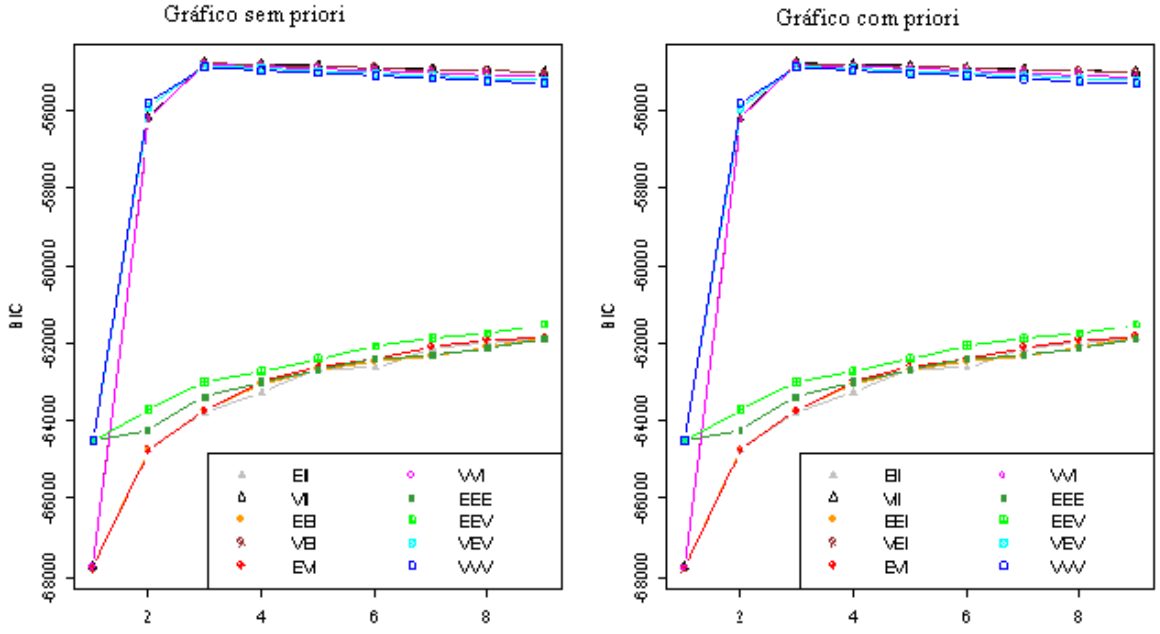


Figura 10.18: Gráfico com os valores de BIC.

11. Aplicação dos Modelos de Mistura para um conjunto de dados Expressão Gênica

Esta base é composta por 62 amostras provenientes de *microarrays* de oligonucleotídeos que medem os níveis de expressão de 6500 genes humanos. São 22 amostras normais e 40 com câncer de cólon. Alon e *cia*. (1999) disponibilizaram essa mesma base de dados reduzindo o número de genes 1908, após a eliminação daqueles com baixa intensidade do sinal. Os dados podem ser obtidos em,
http://microarray.princeton.edu/oncology/.

O resultado da aplicação do método resultou em um modelo VEV com 4 grupos. É um modelo com volume e orientação, vaiáveis e forma igual. Na Tabela 11.1, podemos ver como foram distribuído às observações nos grupos.

Tabela 11.1. Quantidade em cada grupo, usando o modelo de misturas de normais.

Grupos	Qtde de Obs
1	752
2	352
3	670
4	134
Total	1908

Tabela 11.2. Valores do BIC para os melhores modelos.

VEV,4	VEV,5	VVV,3
-705493,9	-706741,5	-707733,9

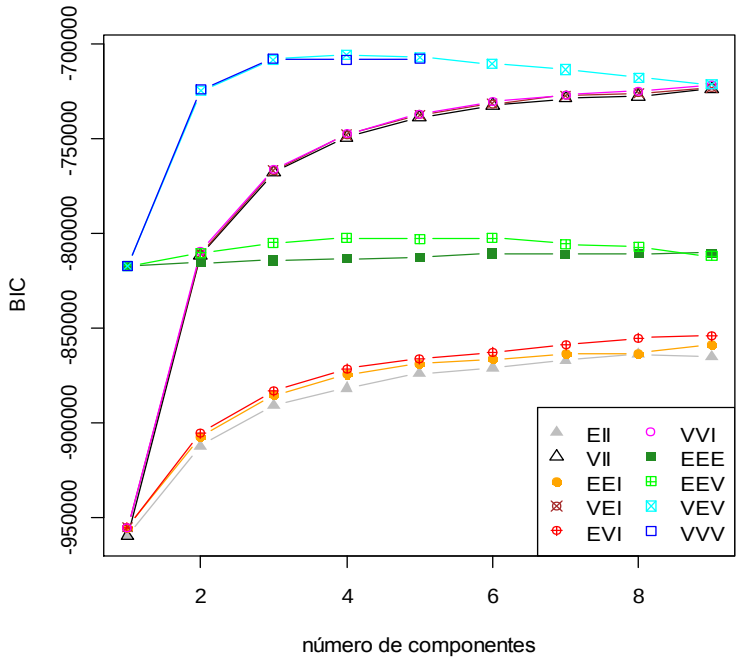


Figura 11.1: Gráfico com os valores de BIC.

Foi adicionada uma priori informativa, como mencionado anteriormente uma priori normal para o vetor de médias e uma Wishart para matriz de covariâncias, mas não houve nenhuma mudança no agrupamento, ou seja, as quantidades foram as mesmas em cada grupo. Portanto, como podemos ver na Figura 11.2 não houve mudanças significativas nos valores de BIC.

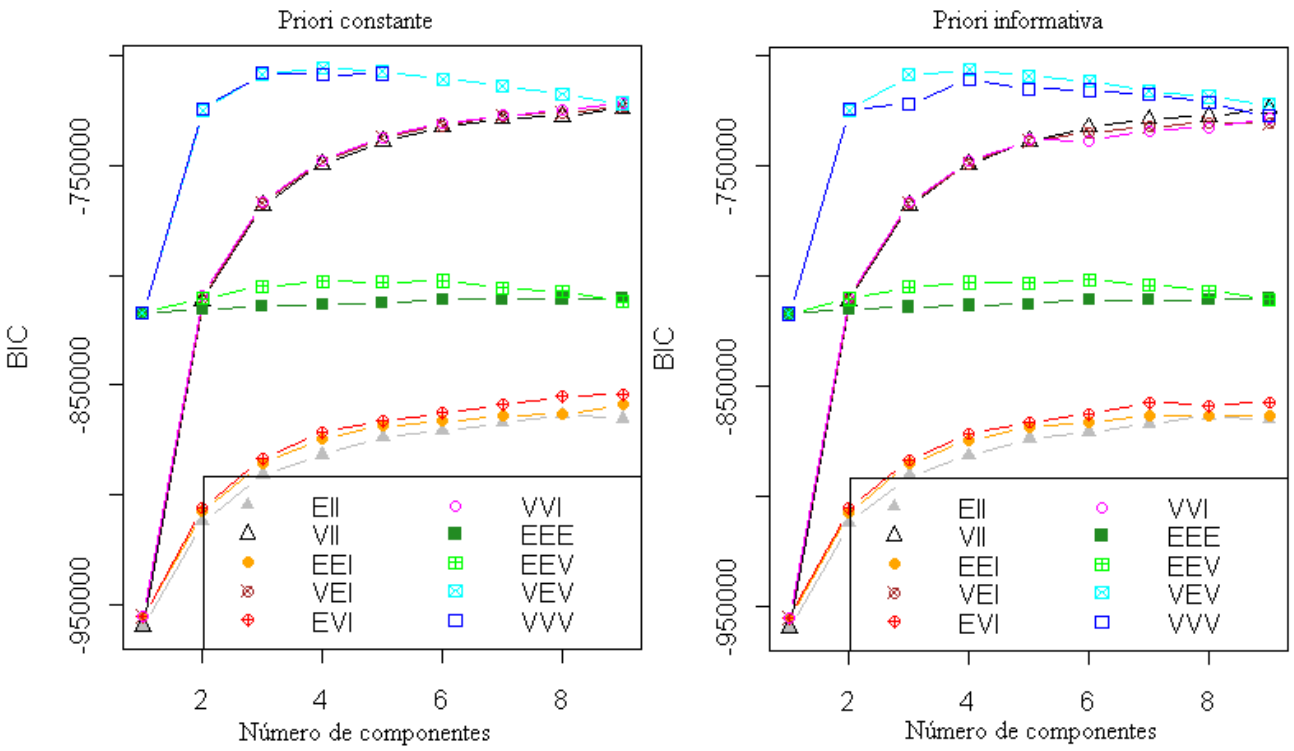


Figura 11.2: Gráfico com os valores de BIC com priori constante e priori informativa.

- Conclusão

Como não sabemos a classificação verdadeira desses genes, não podemos avaliar o erro na classificação. No entanto, temos acompanhado um bom desempenho deste método ao longo das aplicações neste trabalho, como por exemplo, nas simulações, e em outro conjunto de dados também analisado neste trabalho, os dados Íris. Em todos esses casos o método que apresentou a menor quantidade de observações classificadas incorretamente foi o Modelo de Mistura de Normais. Outra coisa a ser considerada é que os tempos de execução de cada método não diferiram de forma significativa, não sendo um ponto para julgar a eficiência dos métodos.

12. Referências Bibliográficas

Kohonen, T. 1997. **"**Self-Organizing Maps".

Edwards, A. W. F. e Cavalli-Sforza, L. L (1965). "A Method for Cluster Analysis".

Flaley, C. e Raftery, A. E. 2002. "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97.

Banfield, J. D. e Raftery, A. E. 1993. "Model-Based Gaussian and Non-Gaussian Clustering"**.**

Murtagh, F.. e Raftery, A. 1984. "Fitting Straight Lines to Point" Patterns, Pattern Recognition.

Fraley, C., e Raftery, A.. 1998. **"**How Many Clusters? Which Clustering Method? - Answers via Model-Based Cluster Analysis**"**, *Computer Journal*.

Bensmail, H., Ceulex, G., Raftery, A. E. and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis", *Estatistics and Computing*. 7, 1-10.

Friedman, H. P., and Rubin, J. (1967), "On Some Invariant Criteria for Grouping Data*." Journal of the American Statistical Association*, 62, 1159 – 1178.

McLachlan, G. J., and Basford, K. E. (1988). *Mixture Models: Inference and Aplications to Clustering*. New York: Marcel Dekker.

Celeux, G. e Govaert, G. 1995. Gaussian parsimonious clustering models**.**

Scott, A. J., and Symons, M. J. (1971). "Clustering Methods Based on Likelihood Ratio Criteria." *Biometrics*, 27, 387 – 244.

Nature Genetics Supplement Vol 21 no. 1, January 1999.

MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations," *in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Vol 1, eds. L. M. L. Cam and J. Neyman, Berkeley, CA: University of California Press, pp. 281- 297.

Kass, R. e Raftery, A. 1995. "Bayes Factors"**.** *Journal of the American Statistical* Association.

Dempster, Laird e Rubin, D. B. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm"**.** *Journal of the Royal Statistical Society*. Series B (Methodological).

Mclachlan, G. e Krishnan, T.. "The EM algorithm and extensions"**.**

Diebolt, J. e Robert, C. 1994. "Estimation of Finite Mixture Distributions through Bayesian Sampling"**.** *Journal of the Royal Statistical Society*. Series B (Methodological), 56.

Schwarz, G.. 1978. "Estimating the Dimension of a Model"**.** *The Annals of Statistics,* 6.

Jeffreys, H. "Theory of Probability"**. (**3rd, ed.), Oxford, UK: Clarendon Press.

Houghton, D. M. A. 1988. "On the Chose of a Model to Fit Data From an Exponential Family*". The Annals of Statistics,* 16.

Bock, H (1996). "Probabilistic Models in Cluster Analysis". *Computation Statistics and Data Analysis*, 23, 5-28.

(1998a). Probabilistic Approaches in Cluster Analysis". *Bulletin of the International Statistical Institute,* 57, 603-606.

(1998b). "Probabilistic Aspects in Classification" in Data Science. *Classification and Related Methods,* eds Hayashi, C. *et al*. New York: Springer Verlag, pp, 3-21.

Raftery. A. E. (1995). "Bayesian Model Selection in Research (with discussion)". *Sociological Methodology*, 25, 111-193.

(1995). "Bayes Factors and BIC: Comment on A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods and Research*, 27, 411-427.

C. Fraley and A. E. Raftery. MCLUST: Software for model-based cluster analysis. Journal of Classification, 16:297–306, 1999.

Kaufman, L. e Rousseeuw, P. J. "*Finding Groups in Data*"**.** New York: Wiley, 1990.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor e normal colon tissues probed by oligonucleotide arrays. *In Proc Natl Acad Sci USA*, volume 96, pages 6745–6750. National Academy of Sciences.

"R". http://www.r-project.org/

FASULO, D. "An Analysis of Recent Work on Clustering Algorithms", Technical Report 01-03-02, Department of Computer Science & Engineering, University of Washington, 1999.

COSTA, J.A.F. "Classificação Automática e Análise de Dados por Redes Neurais Auto-Organizáveis", Tese de Doutorado, Faculdade de Engenharia Elétrica e de Computação (FEEC/Unicamp), 1999.

Everitt, B., Landau, S. & Leese, M. (2001), *Cluster Analysis*, 4 edn, Arnould, New York.

Meireles S, Carvalho A, Hirata R, Montagnini A, Martins W, Runza F, Stolf B, Termini L, Neto C, Silva R, Soares F, Neves E and Reis L. **"Differentially expressed genes in gastric tumors identified by cDNA array"**. Cancer Letters Vol 190 pg 199-211, 2003.

Duggan D, Bittner M, Chen Y, Meltzer P and Trent J. **"Expression profiling using cDNA microarrays"**. Nature Genetics Vol 21 pg 10-14, 1999.

Ward, J. H.. Hierarchical groupings to optimize an objective function. Journal of the American Statistical Association, 58:234–244, 1963.

ZHANG, B.; HSU, M.; DAYAL, U. K-harmonic means - A spatial clustering algorithm with bossting. In: RODDICK, J.F.; HORNSBY, K., Eds., *Temporal, Spatial and Spatio-Temporal Data Mining, Lecture Notes in Artificial Intelligence*. Berlin: Springer, 2001.

Berkhin, P. Survey of Clustering Data Mining Techniques. Accrue Software, 2002.

Doval, D., Mancoridis, S. e Mitchell, B. S. Automatic Clustering of Software Systems using a Genetic Algorithm. In *1999 International Conference on Software Tools and Engineering Practice (STEP '99)*, 1999.

Ng, R. T. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In International Conference on Very Large Data Bases (VLDB) Santiago, Chile.

HAN J.; FU Y. Discovery of multiple-level association rules from large databases. **Proceedings...** Int. Conference Very Large Data Bases, Zurich, Switzerland, 1995.

KOHONEN, T. Self-Organizing Maps. Berlim: Springer, 2001.

HAYKIN, Simon. *Redes Neurais: Princípios e Prática*. Porto Alegre: Bookman, 2001 2ed. 2001

Heidelberg. "Computational Analysis of expression data", 1999.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, Part II, 179–188.