EDWIN RAFAEL VILLANUEVA TALAVERA

MÉTODOS BAYESIANOS APLICADOS EM TAXONOMIA MOLECULAR

Dissertação apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Área de concentração: Processamento de Sinais e Instrumentação.

ORIENTADOR:

Prof. Dr. Carlos Dias Maciel

São Carlos

2007

Dedico

A minha esposa por seu amor incondicional,

Aos meus pais por terem educado no caminho do bem.

AGRADEÇO:

Ao meu orientador Prof. Dr. Carlos Dias Maciel pela oportunidade e apoio dado no desenvolvimento do presente trabalho;

Aos Professores, Dra. Vilma Alves de Oliveira e Dr. Estevam Rafael Hruschka, pela participação na banca examinadora e revisão crítica do trabalho;

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro;

Ao pessoal do laboratório LIM pela amizade;

A todos que direta ou indiretamente contribuíram com este trabalho;

“Todo lo que vivamente imaginamos, ardientemente deseamos, sinceramente creamos, y entusiastamente emprendamos… inevitablemente sucederá""

Paul J. Meyer

VILLANUEVA, Edwin R. (2007). **Métodos Bayesianos aplicados em taxonomia molecular**. 107f. Dissertação (Mestrado). Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos.

Neste trabalho são apresentados dois métodos de agrupamento de dados visados para aplicações em taxonomia molecular. Estes métodos estão baseados em modelos probabilísticos, o que permite superar alguns problemas apresentados nos métodos não probabilísticos existentes, como a dificuldade na escolha da métrica de distância e a falta de tratamento e aproveitamento do conhecimento *a priori* disponível. Os métodos apresentados combinam por meio do teorema de Bayes a informação extraída dos dados com o conhecimento *a priori* que se dispõe, razão pela qual são denominados métodos Bayesianos. O primeiro método, *método de agrupamento hierárquico Bayesiano,* está baseado no algoritmo HBC (*Hierarchical Bayesian Clustering*). Este método constrói uma hierarquia de partições (dendrograma) baseado no critério da máxima probabilidade *a posteriori* de cada partição. O segundo método é baseado em um tipo de Modelo Gráfico Probabilístico conhecido como Redes Gaussianas Condicionais, o qual foi adaptado para problemas de agrupamento. Ambos métodos foram avaliados em três bancos de dados donde se conhece a rótulo da classe. Os métodos foram usados também em um problema de aplicação real: a taxonomia de uma coleção brasileira de estirpes de bactérias do gênero *Bradyrhizobium*. (conhecidas por sua capacidade de fixar o N_2 do ar no solo). Este banco de dados é composto por dados genotípicos resultantes da análise do RNA ribossômico. Os resultados mostraram que o método hierárquico Bayesiano gera dendrogramas de boa qualidade, em alguns casos superior que o melhor dos algoritmos hierárquicos analisados. O método baseado em redes gaussianas condicionais também apresentou resultados aceitáveis, mostrando um adequado aproveitamento do conhecimento *a priori* sobre as classes tanto na determinação do número ótimo de grupos, quanto no melhoramento da qualidade dos agrupamentos.

Palavras-Chave: agrupamento, agrupamento hierárquico, modelos probabilísticos, modelos gráficos probabilísticos, taxonomia molecular.

VILLANUEVA, Edwin R. (2007). **Bayesian Methods applied in molecular taxonomy** 107f. M.Sc. (Dissertation) - School of Engineering - University of São Paulo., São Carlos, 2007.

In this work are presented two clustering methods thought to be applied in Molecular Taxonomy. These methods are based in probabilistic models which overcome some problems observed in traditional clustering methods such as the difficulty to know which distance metric must be used or the lack of treatment of available prior information. The proposed methods use the Bayes theorem to combine the information of the data with the available prior information, reason why they are called Bayesian methods. The first method implemented in this work was the Hierarchical Bayesian Clustering, which is an agglomerative hierarchical method that constructs a hierarchy of partitions (dendogram) guided by the criterion of maximum Bayesian posterior probability of the partition. The second method is based in a type of Probabilistic Graphical Model knows as Conditional Gaussian Network, which was adapted for data clustering. Both methods were validated in 3 datasets where the labels are known. The methods were used too in a real problem: the clustering of a Brazilian collection of bacterial strains belonging to the genus *Bradyrhizobium*, known by .their capacity to transform the nitrogen (N_2) of the atmosphere into nitrogen compounds useful for the host plants. This dataset is formed by genetic data resulting of the analysis of the ribosomal RNA. The results shown that the Hierarchical Bayesian Clustering method built dendrograms with good quality, in some cases, better than the other hierarchical methods. In the method based in conditional Gaussian network was observed acceptable results, showing an adequate utilization of the prior information (about the clusters) to determine the optimal number of clusters and to improve the quality of the groups

Keywords: Clustering, hierarchical clustering, probabilistic models, probabilistic graphical models, molecular taxonomy.

n - Número de elementos ou amostras.

N - Número de variáveis.

Δ - Matriz de distância.

d_{uv} - Distância entre os grupos u e v.

$|u|$ - Cardinalidade do grupo u

D - Conjunto de elementos, amostras, instâncias, casos ou dados.

C - Partição (conjunto de classes).

C_i - i-ésima partição.

$|C_i|$ Número de classes da partição C_i.

$p(C \mid D)$ - Probabilidade da partição C dados os dados D.

$\boldsymbol{X}$ - Variável aleatória, conjunto de variáveis.

X_i - Variável aleatória com posição i, $X_i \in \boldsymbol{X}$.

x_i - Uma instância ou estado da variável aleatória X_i.

s - Grafo ou estrutura de modelo de um MGP.

s^h - Uma estrutura de modelo hipotética.

$\boldsymbol{Pa}(s)_i$ - Conjunto de variáveis aleatórias pais da variável X_i na estrutura de modelo s.

$\boldsymbol{\theta}$ - Conjunto de parâmetros de um MGP.

$\boldsymbol{\theta_i}$ - Conjunto de parâmetros da variável X_i num MGP.

$Cc(s)$ - Conjunto de nós que formam um componente conectado induzido pela estrutura do modelo s.

$CcPa(s)_i$ - Conjunto de nós discretos pais dos nós contidos no componente conectado $Cc(s)_i$.

θ_i^{jk} - Probabilidade condicional de que X_i toma seu k-ésimo valor dado que $\boldsymbol{Pa}(s)_i$ toma seu j-ésimo valor

$s^{(\boldsymbol{Y},X_i)}$ - Sub-grafo induzido no grafo s pelo conjunto de nós correspondentes às variáveis aleatórias $(\boldsymbol{Y}, X_i)$.

ϕ - CG-potencial de crença.

$(g, \boldsymbol{h}, \boldsymbol{K})$ - Características canônicas de um CG-potencial.

$\{p, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ - Características de momentos de um CG-potencial.

V - Clique ou universo de crença.

S - Separador de cliques.

ϕ_V - CG-potencial de clique.

ϕ_S - CG-potencial de separador.

ϕ_U - Sistema de crença conjunta

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Agrupamento de dados, ou *data clustering* em inglês, é um conjunto de técnicas usadas numa grande variedade de campos incluindo reconhecimento de padrões, aprendizado de máquina, mineração de dados, estatística, análise de imagens, bioinformática entre outras. Basicamente estas técnicas têm por finalidade encontrar grupos ou *clusters* dentro de um conjunto dado de elementos, tal que os elementos dentro de cada grupo compartilhem certas características em comum (EVERITT *et. al*., 2001).

É importante entender a diferença entre agrupamento de dados e classificação de dados. Nesta última, tem-se disponível um conjunto de elementos pré-classificados (a etiqueta da classe de cada elemento é conhecida) o qual é usado para aprender as estruturas das classes. Com estas estruturas aprendidas é possível classificar novos elementos com etiquetas desconhecidas (EVERITT *et. al*., 2001). No caso das técnicas de agrupamento, os dados não possuem etiquetas e o problema é justamente encontrar essa informação com base nos atributos dos dados. Em alguns problemas de agrupamento se dispõe, além dos dados, certo conhecimento *a priori* deles, como o número de classes e características gerais das classes, o qual facilita a classificação. Em outros problemas não se tem nenhum conhecimento *a priori* sobre os dados e freqüentemente tem-se que fazer algumas

considerações para poder classificá-los.

A demanda por métodos de agrupamento de dados vem de muitos campos de aplicação. Em biologia, por exemplo, existe um grande interesse em métodos eficientes para avaliar a biodiversidade e posição taxonômica de organismos vivos. Isto é devido à cada vez maior disponibilidade de dados biomoleculares como conseqüência dos avanços nas técnicas de biologia molecular que ocorreram principalmente nas duas últimas décadas. Esta foi a principal motivação do presente trabalho, o qual tem por objetivo principal desenvolver métodos de agrupamento aplicados em taxonomia de organismos com base nas suas características genotípicas.

Existe uma grande quantidade de métodos de agrupamento encontrados na literatura, os quais podem ser classificados de distintas formas. A classificação apresentada em (JAIN; DUBES, 1988) é bastante aceita, na qual se distinguem dois grandes grupos: os métodos hierárquicos e os métodos particionais. Os métodos hierárquicos geram sucessivos agrupamentos baseados em agrupamentos previamente criados formando uma hierarquia de grupos chamada dendrograma, enquanto os métodos particionais determinam todos os grupos em uma vez.

Existem várias limitações e inconvenientes nos algoritmos hierárquicos e particionais tradicionais (não probabilísticos) (EVERITT *et al.*, 2001). Uma delas é a escolha da métrica adequada para medir a similaridade ou dissimilaridade (distância) entre os elementos. Em muitas aplicações, especialmente com dados estruturados como imagens ou seqüências, essa escolha é uma tarefa difícil de cumprir devido ao fato de que freqüentemente distintas métricas geram distintos agrupamentos (HELLER; GHAHRAMANI, 2005). Um outro problema observado é que estes

algoritmos não constroem modelos para descrever os grupos encontrados, sendo difícil comparar distintos agrupamentos e também atribuir rótulos de classe a novos elementos após do agrupamento (HELLER; GHAHRAMANI, 2005). Uma outra limitação nos algoritmos não probabilísticos é que estes não determinam o grau de incerteza com que os dados estão sendo agrupados. Finalmente, estes algoritmos não consideram o conhecimento *a priori* que se dispõe acerca dos dados, o qual pode ter influência na determinação final dos grupos.

Os métodos de agrupamento baseados em modelos probabilísticos foram propostos para superar as limitações mencionadas acima (BERKHIN, 2002). No caso dos algoritmos particionais probabilísticos, os dados são modelados como sendo gerados por uma mistura de distribuições de probabilidades (EVERITT *et al*., 2001). No caso dos algoritmos hierárquicos probabilísticos, a hierarquia de grupos é construída com base em algum critério probabilístico (IWAYAMA; TOKUNAGA, 1995; FRALEY; RAFTERY, 1998; HELLER; GHAHRAMANI, 2005). Em ambos casos é possível avaliar as classificações resultantes mediante a comparação dos modelos aprendidos com algum critério estatístico.

No presente trabalho foram implementados dois métodos de agrupamento baseados em modelos probabilísticos voltados para aplicações em taxonomia molecular. Estes métodos podem considerar (se disponível) o conhecimento a priori dos dados. A forma como combinam esse conhecimento com o conhecimento extraído dos dados é por meio do teorema de Bayes, por tal razão são denominados métodos Bayesianos. O primeiro método é o método de agrupamento hierárquico Bayesiano - HBC (do inglês, *Hierarchical Bayesian Clustering*). Este método pertence ao grupo dos métodos hierárquicos aglomerativos. Basicamente este método ordena os dados numa hierarquia de grupos procurando a maximização da

probabilidade *a posteriori* de cada agrupamento. O segundo método implementado corresponde a um método particional probabilístico baseado em um tipo de Modelo Gráfico Probabilístico (PEARL, 1988) conhecido como Redes Gaussianas Condicionais (PEÑA, 2001) o qual foi adaptado para realizar agrupamento de dados. Este tipo de modelo foi escolhido porque além de agrupar pode dar uma idéia do relacionamento entre os atributos dos dados e também por sua capacidade para lidar com conhecimento *a priori* disponível e com dados incompletos como é o caso de problemas de agrupamento.

O presente trabalho está organizado em 5 capítulos. No capítulo 2, seguindo esta introdução, descrevem-se os fundamentos teóricos implicados nos métodos implementados. No Capítulo 3 são descritos os materiais e métodos utilizados no presente trabalho. No Capítulo 4 são apresentados e discutidos os resultados obtidos. Finalmente, no Capítulo 5 são apresentadas as conclusões e tópicos para trabalhos futuros.

Capítulo 2

FUNDAMENTOS TEÓRICOS

Este capítulo apresenta os fundamentos teóricos necessários para o desenvolvimento do trabalho. O capítulo é dividido em duas partes. Na primeira parte é apresentada a teoria subjacente aos métodos de agrupamento hierárquico tanto probabilísticos e não probabilísticos. Na segunda parte é introduzida a teoria dos Modelos Gráficos Probabilísticos (MGP) com principal ênfase nas redes Gaussianas Condicionais como tipos de MGP usados em problemas de agrupamento de dados.

2.1. AGRUPAMENTO HIERÁRQUICO

No agrupamento hierárquico, os dados são organizados em uma série de agrupamentos ou partições, os quais podem variar desde agrupamentos com todos os elementos em um único grupo até agrupamentos com n grupos unitários. Os métodos de agrupamento hierárquico podem ser divididos em aglomerativos e divisivos. Os métodos aglomerativos começam formando grupos unitários e vão fundindo-os sucessivamente formando grupos maiores. Entretanto, os métodos divisivos começam com um único grupo formado por todos os elementos e vão dividindo-os sucessivamente em grupos menores.

A saída produzida por ambos os métodos é uma hierarquia de grupos freqüentemente representada por uma árvore binária chamada *dendrograma*, a qual mostra as fusões ou divisões realizadas em cada passo do processo. Um exemplo de dendrograma é mostrado na Figura 2.1, na qual os números acima e abaixo indicam um determinado passo no processo de aglomeração ou divisão respectivamente. Este tipo de representação é bastante usado em aplicações de biologia para representar filogenias e taxonomias. Outras áreas onde o agrupamento hierárquico é apropriado são em museologia, sistemas sociais, biblioteconomia, e outros (EVERITT *et al.*, 2001). O agrupamento hierárquico também tem sido aplicado em áreas onde não existe necessariamente uma estrutura hierárquica natural, mas tem mostrado sua utilidade para organizar os dados e prover um ponto de partida para outros métodos de agrupamento (EVERITT *et al.*, 2001).

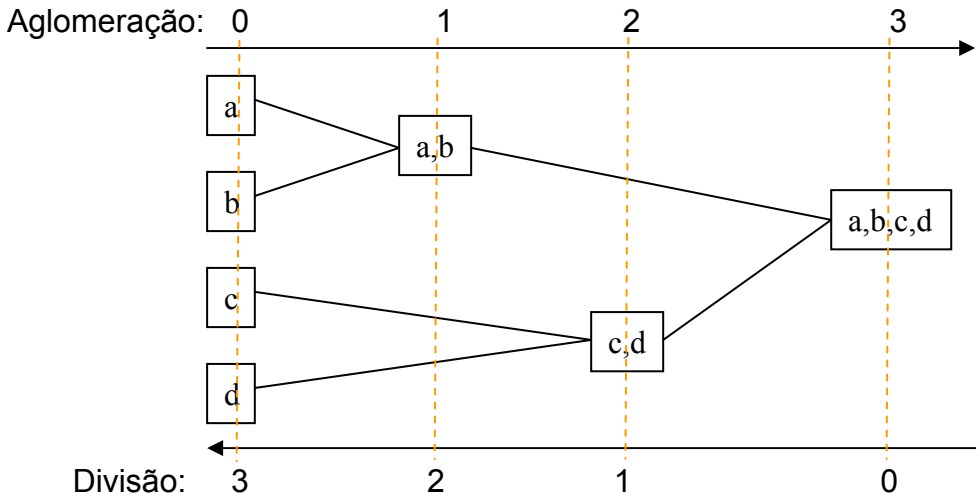


Figura 2.1: Exemplo de representação em dendrograma, na qual cada retângulo representa um grupo resultante da aglomeração ou divisão do método hierárquico respectivo.

Os métodos aglomerativos atualmente são mais amplamente usados que os

divisivos devido a que estes últimos são computacionalmente exigentes pela enorme quantidade de possíveis divisões inicias, razão pela qual não são estudados no presente trabalho. A seguir são descritos os métodos aglomerativos não probabilísticos comumente usados, os quais são chamados no resto do texto como métodos aglomerativos tradicionais. Logo são apresentados os métodos probabilísticos para agrupamento hierárquico aglomerativo.

2.1.1. MÉTODOS AGLOMERATIVOS TRADICIONAIS

Os métodos aglomerativos tradicionais produzem uma série de partições. No começo são iniciados n grupos (n = número de elementos) com cada elemento em um grupo distinto. À cada passo do processo de aglomeração um par de grupos é fundido de acordo com algum critério de similaridade. O processo finaliza quando todos os elementos são juntados em um único grupo.

Existe uma variedade de métodos aglomerativos, que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos. Entretanto, a maioria dos métodos parecem ser formulações alternativas de três grandes métodos de agrupamento aglomerativo (ANDERBERG, 1973):

1) Métodos de ligação (*single linkage, complete linkage, average linkage*);
2) Métodos de centróide;
3) Métodos de soma de erros quadráticos ou variância (método de Ward).

Entrada: Banco de dados $D = \{d_1, d_2, ..., d_n\}$ com n elementos.

Função de distância entre elementos $f : (d_i, d_j) \rightarrow R^+$

Função de distância entre grupos $g : (v, w) \rightarrow R^+$

Saída: Uma hierarquia de grupos.

1.Iniciar n grupos, contendo um elemento em cada grupo

2.Calcular a matriz de distância $\Delta = (\delta_{ij})$, onde $\delta_{ij} = f(d_i, d_j)$, $i, j \in [1, n]$

3.Repetir;

4. Localizar os grupos u e v com menor distância δ_{uv} em Δ ($u \neq v$);

5. Construir um novo grupo (uv) pela fusão desse par de grupos de distância mínima

6 Atualizar a matriz Δ, suprimindo as linhas e as colunas correspondentes aos grupos fusionados u e v;

7. Atualizar a matriz Δ, adicionando uma linha e uma coluna correspondente às distâncias $\delta_{(uv)w}$ entre o novo grupo (uv) e cada grupo existente w, onde $\delta_{(uv)w} = g(uv, w)$;

8. Até $n-1$, quando todos elementos estarão em um único grupo.

Figura 2.2: Algoritmo padrão dos métodos de agrupamento hierárquicos aglomerativos tradicionais.

Os métodos aglomerativos seguem, de modo geral, o algoritmo padrão descrito na Figura 2.2. A matriz de distância inicial $\Delta = (\delta_{ij})$ no passo 2 do algoritmo é formada aplicando a função de distância $f(.)$ a cada par de elementos d_i, d_j do banco de dados D, isto é, $\delta_{ij} = f(d_i, d_j)$. A escolha da função de distância entre elementos é depende do tipo de dado e do problema abordado. Algumas funções de distâncias conhecidas são (EVERITT, 2001): distância Euclidiana, distância C*ity Block,* distância Minkowski, correlação de Pearson, distância co-seno. A diferença entre os métodos ocorre no passo 7, onde a função de distância entre grupos $g(.)$ é usada para calcular as distâncias $\delta_{(uv)w}$ entre o novo grupo (uv) e cada grupo

existente w , isto é, $\delta_{(uv)w} = g((uv), w)$ (JOHNSON, 1992). Estas funções são mostradas a seguir para cada método junto com algumas propriedades importantes deles.

Método de ligação por vizinho mais próximo (*Single Linkage*): Neste método, a distância $\delta_{(uv)w}$ entre o grupo fundido (uv) e qualquer outro grupo existente w é calculada com a menor distância entre os elementos integrantes de ambos grupos:

$$\delta_{(uv)w} = \min(\delta_{uw}, \delta_{vw}) . \tag{2.1}$$

Algumas características importantes deste método são (ANDERBERG, 1973):

- Permite detectar grupos de formas não-elípticas;
- Grupos muito próximos podem não ser identificados;
- Pouca tolerância a ruído;
- Tendência a formar longas cadeias (encadeamento).

Método de ligação por vizinho mais distante (*Complete Linkage*): Neste método, a distância $\delta_{(uv)w}$ entre o grupo fundido (uv) e qualquer outro grupo w é calculada com a maior distância entre os elementos integrantes de ambos grupos:

$$\delta_{(uv)w} = \max(\delta_{uw}, \delta_{vw}) . \tag{2.2}$$

Algumas características importantes são (KAUFMANN, 1990):

- Tendência a formar grupos compactos;
- Os ruídos demoram a serem incorporados ao grupo;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras.

Método de ligação média (*Average Linkage*): Neste método, a distância entre o grupo fundido (uv) e qualquer outro grupo w é calculada com a distância média de u e v com respeito a w:

$$\delta_{(uv)w} = \frac{|u| \cdot \delta_{uw} + |v| \cdot \delta_{vw}}{|u| + |v|} \tag{2.3}$$

em que, $|u|$ e $|v|$ é a cardinalidade dos conjuntos u e v respectivamente.

Algumas características importantes são (KAUFMANN, 1990):

- Menor sensibilidade a ruídos que os outros métodos de ligação;

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

- Tendência a formar grupos com número de elementos similares.

Método de ligação por centróide (*Centroid Linkage*) : Neste método, a distância entre o grupo fundido (uv) e qualquer outro grupo existente w é atualiza como:

$$\delta_{(uv)w} = \frac{|u| \cdot \delta_{uw} + |v| \cdot \delta_{vw}}{|u| + |v|} - \frac{|u| \cdot |v| \cdot \delta_{uv}}{(|u| + |v|)^2}. \tag{2.4}$$

Algumas características importantes são:

- Tolerância à presença de ruído;

- Fenômeno da reversão, isto é, novos grupos podem-se formar a um nível inferior dos grupos existentes, tornando o dendrograma confuso.

Método de ligação de Ward: Neste método, a distância entre o grupo fundido (uv) e qualquer outro grupo existente w é atualiza como:

$$\delta_{(uv)w} = \frac{(|w| + |u|) \cdot d_{uw} + (|w| + |v|) \cdot \delta_{vw} - |w| \cdot \delta_{uv}}{|u| + |v| + |w|}. \tag{2.5}$$

Algumas características importantes são:

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

- Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é significativamente diferente;

- Tendência a combinar grupos com poucos elementos;

- Sensível à presença de *outliers*[1].

2.1.2. MÉTODOS AGLOMERATIVOS PROBABILÍSTICOS

A principal limitação dos métodos aglomerativos tradicionais é que estes não proporcionam nenhum guia para selecionar o agrupamento ótimo ou "natural" dos dados, isto é, o nível do dendrograma onde deve ser cortado. Outra limitação dos métodos tradicionais é a dificuldade na escolha da métrica de distância adequada, especialmente para dados estruturados como imagens ou seqüências (HELLER; GHAHRAMANI, 2005).

Métodos hierárquicos aglomerativos probabilísticos são o híbrido entre métodos hierárquicos aglomerativos e modelos probabilísticos. A característica principal destes métodos é que modelam os grupos com distribuições de probabilidades e realizam o processo de aglomeração guiados por algum critério probabilístico. Em (HELLER;GHAHRAMANI, 2005) é apresentada uma revisão do estado da arte de tais algoritmos. As diferenças encontradas nestes algoritmos estão principalmente:

- No modelo probabilístico usado para descrever os dados, entre eles se encontram: mistura de gaussianas, mistura de exponenciais, entre outras;

- No critério usado para guiar o processo de aglomeração, podendo ser: máxima verossimilhança, verossimilhança marginal, máxima probabilidade *a posteriori*.

Entre estes algoritmos encontra-se o *Hierarchical Bayesian Clustering* - HBC exposto em (IWAYAMA; TOKUNAGA, 1995), o qual é descrito a seguir pois constitui a base do método de agrupamento hierárquico implementado neste trabalho.

[1] Observações que são numericamente distantes do resto do banco de dados

Algoritmo de agrupamento HBC

Este algoritmo foi inicialmente proposto para classificação de documentos de texto, onde apresentou bons resultados (IWAYAMA; TOKUNAGA, 1995). Basicamente, o algoritmo constrói uma hierarquia de grupos (dendrograma) fundindo em cada interação o par de grupos que produzem o agrupamento com a maior probabilidade *a posteriori*. Os agrupamentos produzidos em cada interação são chamados partições.

No inicio é formada uma partição com n grupos unitários. A cada interação o algoritmo calcula para cada par de grupos a probabilidade da partição resultante de juntá-los, selecionando assim o par de grupos que ao serem fundidos apresentem a maior probabilidade da partição resultante. O par de grupos encontrados são logo substituídos por um novo grupo, o qual contêm os elementos de ambos grupos. O algoritmo toma $n-1$ iterações até juntar os n grupos iniciais em um único grupo.

Considere-se $D = \{d_1, d_2, ..., d_n\}$ o conjunto de dados a agrupar. No inicio é formada a partição inicial $C_0 = \{c_1, c_2, ..., c_n\}$ na qual cada grupo é formado por um único elemento, $c_i = \{d_i\}$. Após k interações tem-se a partição C_k formada pelo conjunto de grupos $\{c_1, c_2, ..., c_{n-k}\}$, onde $1 \le k \le n-1$. Seja $p(C_k \mid D)$ a probabilidade *a posteriori* de agrupar os dados D no conjunto de grupos C_k. O par de grupos escolhido para ser fundido na iteração k é aquele que apresenta o máximo valor da probabilidade $p(C_k \mid D)$ sobre o espaço de todos os pares de grupos possíveis a fundir. Considerando independência condicional entre os dados dentro dos grupos, a probabilidade de agrupar os dados na partição C_k pode ser calculada como:

$$p(C_k \mid D) = \prod_{c_i \in C_k} \prod_{d_j \in c_i} p(c_i \mid d_j) \tag{2.6}$$

aplicando o teorema de Bayes e desenvolvendo a equação anterior, tem-se:

$$p(C_k \mid D) = \prod_{c_i \in C_k} \prod_{d_j \in c_i} \frac{p(d_j \mid c_i) p(c_i)}{p(d_j)}$$

$$= \frac{\prod_{c_i \in C_k} p(c_i)^{|c_i|}}{p(D)} \prod_{c_i \in C_k} \prod_{d_j \in c_i} p(d_j \mid c_i)$$

$$= \frac{p(C_k)}{p(D)} \prod_{c_i \in C_k} SC(c_i) \tag{2.7}$$

onde (IWAYAMA; TOKUNAGA, 1995):

- $p(D)$ é a probabilidade marginal dos dados, a qual é constante para qualquer partição.
- $p(C_k)$ é a probabilidade *a priori* de agrupar os n dados na partição C_k:

$$p(C_k) = \prod_{c_i \in C_k} p(c_i)^{|c_i|} \tag{2.8}$$

  em que $p(c_i)$ é a probabilidade *a priori* do grupo c_i e $|c_i|$ sua cardinalidade.

- $SC(c_i)$ é a verossimilhança do grupo c_i, isto é, a probabilidade do grupo c_i de agrupar todos seus elementos d_j, calculada como:

$$SC(c_i) = \prod_{d_j \in c_i} p(d_j \mid c_i) \tag{2.9}$$

- $\prod_{c_i \in C_k} SC(c_i)$ é definido como a verossimilhança da partição C_k, a qual mede quão prováveis são os dados se a partição fosse certa. É calculado como a produtória das verossimilhanças de cada grupo integrante.

Para determinar os dois grupos que devem ser fundidos na interação k

considera-se o par de grupos candidatos c_x e c_y. O grupo resultante da fusão é $c_z = c_x \cup c_y$ e a partição é atualizada da seguinte forma:

$$C_k = C_{k-1} + c_z - c_x - c_y \tag{2.10}$$

Usando (2.7) para expressar a probabilidade posterior da partição C_{k-1}, tem-se:

$$p(C_{k-1} \mid D) = \frac{p(C_{k-1})}{p(D)} \prod_{c_i \in C_{k-1}} SC(c_i) \tag{2.11}$$

Dividindo (2.11) e (2.7) pode-se isolar a probabilidade *a posteriori* da partição C_k em termos da probabilidade da partição anterior C_{k-1} :

$$p(C_k \mid D) = \frac{p(C_k)}{p(C_{k-1})} \frac{SC(c_z)}{SC(c_x)SC(c_y)} p(C_{k-1} \mid D). \tag{2.12}$$

Note-se que esta atualização é local e pode ser feita eficientemente desde que tudo o que se precisa calcular é a verossimilhança do novo grupo $SC(c_z)$. O termo $\frac{p(C_k)}{p(C_{k-1})}$ é o fator no qual a probabilidade *a priori* da partição muda a cada passo. Usando o principio *Minimum Description Length* - MDL (RISSANEN, 1989): A probabilidade *a priori* de um modelo (neste caso, um grupo) é uma função decrescente do seu tamanho. Pode-se definir essa função como $p(c) \propto (A)^{-|c|}$ (IWAYAMA; TOKUNAGA, 1995), onde $A > 1$. De acordo com este principio, (2.8) pode ser reescrita como:

$$p(C_k) = \prod_{c_i \in C_k} p(c)^{|c_i|} \propto \prod_{c_i \in C_k} A = A^{|C_k|} . \tag{2.13}$$

Cada partição $|C_k|$ decresce de uma unidade a cada passo, conseqüentemente $\frac{PC(C_k)}{PC(C_{k-1})}$ se reduz à constante A^{-1}. A maximização é então feita sobre o quociente

dos termos SC em (2.12), o qual é denotado como $U(c_x, c_y)$:

$$U(c_x, c_y) = \frac{SC(c_x \cup c_y)}{SC(c_x)SC(c_y)} \tag{2.14}$$

portanto, os grupos a serem fundidos serão os que maximizem (2.14), isto é:

$$(c_x, c_y) = \arg\max_{c_x, c_y}(U(c_x, c_y)) . \tag{2.15}$$

A probabilidade elementar $p(d_j \mid c_i)$ necessária para calcular $SC(c_i)$ em (2.9) é a probabilidade do grupo c_i de agrupar o seu dado membro d_j. Esta probabilidade é determinada de acordo com o modelo probabilístico usado para modelar os grupos, o qual é dependente da aplicação em questão. Na Figura 2.3 é apresentado um sumário do algoritmo descrito na presente seção.

Entrada :

$D = \{d_1, d_2, ..., d_n\}$, conjunto de dados

Saída :

Uma hierarquia de partições $\{C_0, ..., C_k, ...C_{n-1}\}$

Iniciar:

$c_i = \{d_i\}$, para $i = 1...n$

$C_0 = \{c_1, c_2, ...c_n\}$, Partição inicial

Repetir:

$U = \{\frac{SC(c_i \cup c_j)}{SC(c_i)SC(c_j)}\}$, $c_i, c_j \in C_{k-1}$, $c_i \neq c_j$

$(c_x, c_y) \leftarrow \arg\max_{c_i, c_j}(U)$

$c_z \leftarrow c_x \cup c_y$

$C_k \leftarrow C_{k-1} + c_z - c_x - c_y$

Até $n-1$ vezes, quando todos elementos estarão em um único grupo.

Figura 2.3: Pseudocódigo do algoritmo de agrupamento hierárquico Bayesiano (HBC).

2.2. AGRUPAMENTO COM MODELOS GRÁFICOS PROBABILÍSTICOS

Nesta seção é apresentada uma breve introdução aos modelos gráficos probabilísticos (MGP) em geral e redes gaussianas condicionais - CGN (do inglês, *Conditional Gaussian Networks*) em particular, como um tipo de MGP que pode lidar com dados tanto discretos como contínuos.

Esta seção inicia com uma breve introdução aos MGP para logo apresentar as CGNs e sua forma como são adaptadas para serem usadas em problemas de agrupamento; são apresentados também os algoritmos de inferência e aprendizado neste tipo de modelos.

2.2.1. MODELOS GRÁFICOS PROBABILÍSTICOS

Os modelos gráficos probabilísticos são ferramentas poderosas para modelar e fazer inferência em domínios complexos e com incerteza. Estes têm apresentado um interesse crescente nos últimos anos, devido basicamente a sua flexibilidade e fácil interpretação. Especificamente estas ferramentas provêem uma forma modular de codificar a distribuição de probabilidade conjunta de um problema com uma quantidade estritamente necessária de parâmetros, denominados parâmetros do modelo. Os MGP também codificam os relacionamentos entre as variáveis do problema mediante uma estrutura de grafo denominada estrutura do modelo. Estas vantagens são oferecidas pelos MGP por explorarem as asserções de independência condicional existentes entre as variáveis do problema (PEÑA, 2001).

Seja $\boldsymbol{X} = (X_1,...,X_N)$ uma variável aleatória (VA) N-dimensional. Um modelo gráfico probabilístico (MGP) para $\boldsymbol{X}$ é uma fatoração gráfica da distribuição de probabilidade conjunta para $\boldsymbol{X}$, $p(x)$ (PEARL, 1988). Um MGP para $\boldsymbol{X}$ consiste de dois componentes: um grafo s (estrutura do modelo) que determina as

independências condicionais entre as variáveis aleatórias de $\boldsymbol{X}$, e um conjunto de distribuições de probabilidades locais para a estrutura do modelo. Cada nó da estrutura do modelo representa uma variável aleatória unidimensional X_i, para todo i. Assim, cada nó no grafo s será chamado indistintamente variável aleatória ou nó na continuação do texto. O presente trabalho está focalizado em MGPs com parte estrutural limitada a grafos acíclicos dirigidos (DAGs).

Aplicando a regra da cadeia, tem-se que a distribuição da probabilidade conjunta codificada por um MGP para $\boldsymbol{X}$ pode ser fatorada como:

$$p(x) = p(x_1,...,x_N) = \prod_{i=1}^{N} p(x_i \mid x_1,...x_{i-1}) . \tag{2.16}$$

Pode-se assumir sem perda de generalidade (PEÑA, 2001), que as variáveis de $\boldsymbol{X}$ seguem uma ordem ancestral, isto é, a variável X_i não tem descendentes (representados no grafo s) no conjunto de variáveis $(X_1,...,X_{i-1})$. Conseqüentemente, a estrutura do modelo s codifica um conjunto de independências condicionais da forma $CI(X_i,(X_1,...X_{i-1}) \setminus \boldsymbol{Pa}(s)_i \mid \boldsymbol{Pa}(s)_i)$, onde $\boldsymbol{Pa}(s)_i$ são os nós pais de X_i no grafo s. Logo, (2.16) pode ser reescrita como:

$$p(x) = p(x_1,...,x_N) = \prod_{i=1}^{N} p(x_i \mid \boldsymbol{pa}(s)_i) . \tag{2.17}$$

As distribuições de probabilidades locais do MGP são aquelas induzidas pelos produtos que aparecem em (2.17). É assumido que estas são especificadas com um conjunto de parâmetros $\boldsymbol{\theta_s} = (\boldsymbol{\theta_1},...,\boldsymbol{\theta_N})$. Denotando s^h como a hipótese de que a verdadeira distribuição da probabilidade conjunta para $\boldsymbol{X}$ é fatorada de acordo com o conjunto de independências condicionais codificadas no grafo s, então (2.17) pode ser escrita como:

$$p(x \mid \boldsymbol{\theta}_s, s^h) = p(x_1,...,x_N \mid \boldsymbol{\theta}_s, s^h) = \prod_{i=1}^{N} p(x_i \mid \boldsymbol{pa}(s)_i, \boldsymbol{\theta}_i, s^h). \quad (2.18)$$

Estes resultados foram apresentados no trabalho de (PEARL,1988) e posteriormente por outros pesquisadores (BOUCKAERT,1995; CASTILLO *et al.*,1997; PEÑA, 2001).

Resumindo, para definir um MGP, é necessário especificar: um grafo acíclico dirigido (DAG), o qual codifica o conjunto de independências condicionais entre as variáveis aleatórias do problema em questão e um conjunto de distribuições de probabilidades condicionais, uma para cada nó. Estas distribuições correspondem aos termos do produtório em (2.18).

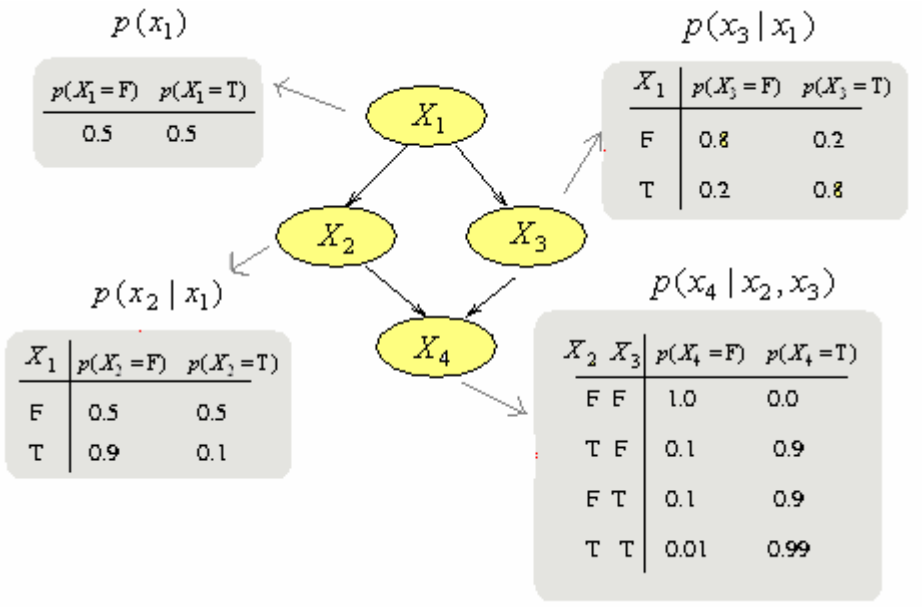


Figura 2.4: Exemplo de estrutura de modelo e distribuições de probabilidades locais (parâmetros) para um MGP de 4 variáveis aleatórias discretas binárias.

Na Figura 2.4 é mostrado um exemplo de MGP para 4 variáveis aleatórias discretas binárias $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$, na qual pode ser visto a estrutura do modelo e as distribuições de probabilidades locais em forma de tabelas de probabilidades. A

fatoração gráfica da probabilidade conjunta para $\boldsymbol{X}$ é obtida segundo (2.18): $p(x_1, x_2, x_3, x_4) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1) \cdot p(x_4 \mid x_{2,} x_3)$. As independências condicionais induzidas pela estrutura de modelo deste exemplo são: $CI(X_4, X_1 \mid X_2, X3)$ e $CI(X_3, X_2 \mid X_1)$.

2.2.2. REDES GAUSSIANAS CONDICIONAIS

As redes gaussianas condicionais - CGNs são uma classe de MGP que pode lidar com dados de natureza mista (discretos e contínuos).

Seja $\boldsymbol{X} = (X_1,...,X_N)$ uma variável aleatória composta por 2 subconjuntos de variáveis: $\boldsymbol{Y} = (Y_1,...,Y_R)$ e $\boldsymbol{Z} = (Z_1,...,Z_{N-R})$, tal que: (i) $\boldsymbol{X} = \boldsymbol{Y} \cup \boldsymbol{Z}$ e $\boldsymbol{Y} \cap \boldsymbol{Z} = \varnothing$, (ii) $\boldsymbol{Y}$ representa uma variável discreta R -dimensional e $\boldsymbol{Z}$ representa uma variável contínua $(N-R)$ -dimensional. Diz-se que X segue uma distribuição gaussiana condicional (LAURITZEN; WERMUTH, 1989; LAURITZEN, 1992; LAURITZEN, 1996; COWEL *et al*., 1999) se a densidade de probabilidade condicional conjunta para $\boldsymbol{Z}$, dados os estados de $\boldsymbol{Y}$, é uma distribuição normal $(N-R)$-dimensional:

$$f(z \mid y) \sim N(z; \boldsymbol{\mu}(\boldsymbol{y}), \boldsymbol{\Sigma}(\boldsymbol{y})) \tag{2.19}$$

para todo $\boldsymbol{Y} = \boldsymbol{y}$ tal que $p(\boldsymbol{y}) > 0$, $\boldsymbol{\mu}(\boldsymbol{y})$ é o vetor de médias, $\boldsymbol{\Sigma}(\boldsymbol{y})$ é a matriz de covariância definida positiva.

Considere-se um MGP para a variável mista $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$, onde $\boldsymbol{Y} = (Y_1,...,Y_R) = (X_1,...,X_R)$ e $\boldsymbol{Z} = (Z_1,...,Z_{N-R}) = (X_{R+1},...,X_N)$. Um *componente conectado* da estrutura s do MGP é definido como o maior subconjunto de $\boldsymbol{Z}$, tal que qualquer par de nós no componente conectado está conectado por um caminho que não inclui nós discretos no grafo s. Seja $Cc(s)_1,...,Cc(s)_g$ a partição de $\boldsymbol{Z}$, tal que cada $Cc(s)_i$ é um componente conectado de s. Seja $CcPa(s)_i$ o conjunto de nós

discretos pais dos nós contidos no componente conectado $Cc(s)_i$. Uma rede gaussiana condicional é um MGP para $\boldsymbol{X}$ onde:

- Os nós discretos não têm nós pais contínuos em s ,
- O conjunto de pais para cada nó no componente conectado $Cc(s)_i$ é exatamente $CcPa(s)_i$ para todo i , e
- $(CcPa(s)_i, Cc(s)_i)$ segue uma distribuição gaussiana condicional.

Este tipo de MGP foi introduzido por primeira vez em (LAURITZEN; WERMUTH, 1989) e desenvolvido posteriormente nos trabalhos de (LAURITZEN,1992; GEIGER; HECKERMAN, 1994; LAURITZEN,1996; COWEL *et al*.,1999). Recentemente, as CGNs têm sido aplicadas satisfatoriamente em problemas de agrupamento de dados (PEÑA, 2001).

A fatoração gráfica da distribuição de probabilidade conjunta para $\boldsymbol{X}$ codificada por uma CGN adquire a seguinte forma:

$$p(x \mid \boldsymbol{\theta_s}, s^h) = p(x_1,...,x_N \mid \boldsymbol{\theta_s}, s^h) = \prod_{i=1}^{R} p(x_i \mid \boldsymbol{pa}(s)_i, \boldsymbol{\theta_i}, s^h) \prod_{\boldsymbol{i=R+1}}^{\boldsymbol{N}} f(x_i \mid \boldsymbol{pa}(s)_i, \boldsymbol{\theta_i}, s^h) . \tag{2.20}$$

A Equação (2.20) é um caso particular de (2.18), na qual foi considerada a natureza mista da variável $\boldsymbol{X}$. Tipicamente a distribuição de probabilidade local para cada variável discreta $X_i \in \boldsymbol{Y}$ é considerada como uma distribuição multinomial, uma para cada valor de $\boldsymbol{Pa}(s)_i$.

Assumindo que X_i ($X_i \in \boldsymbol{Y}$) aceita r_i valores distintos denotados por $(x_i^1,...,x_i^{r_i})$, e que $\boldsymbol{Pa}(s)_i$ pode ser algum dos q_i estados distintos denotados por $\boldsymbol{pa}(s)_i^1,..., \boldsymbol{pa}(s)_i^{q_i}$, sendo $q_i = \prod_{X_e \in Pa(s)_i} r_e$, para todo i tal que $i \leq R$. Portanto, a distribuição multinomial para X_i no primeiro produtório de (2.20) consiste no

conjunto de probabilidades da forma:

$$p(x_i^k \mid \boldsymbol{pa}(s)_i^j, \boldsymbol{\theta_i}, s^h) = \theta_i^{jk} \tag{2.21}$$

onde θ_i^{jk} representa a probabilidade condicional de que X_i toma seu k-ésimo valor dado que $\boldsymbol{Pa}(s)_i$ toma seu j-ésimo valor, para todo k. Além disso, cumpre-se que $\sum_{k=1}^{r_i} \theta_i^{jk} = 1$, para todo i e j tal que $i \leq R$.

Para codificar uma distribuição Gaussiana condicional para $\boldsymbol{X}$, a função de densidade de probabilidade para cada variável contínua X_i ($X_i \in \boldsymbol{Z}$) deve ser um modelo de regressão linear condicionado em cada estado do conjunto de pais discretos de X_i, isto é, $\boldsymbol{Pa}(s^{(\boldsymbol{Y},X_i)})_i$, onde $s^{(\boldsymbol{Y},X_i)}$ é o sub-grafo induzido de s pelo conjunto de nós correspondentes às variáveis aleatórias $(\boldsymbol{Y}, X_i)$, para todo $i > R$. Seja q_i o número de estados que pode tomar $\boldsymbol{Pa}(s^{(\boldsymbol{Y},X_i)})_i$, os quais são denotados como: $\boldsymbol{pa}(s^{(\boldsymbol{Y},X_i)})_i^1, ..., \boldsymbol{pa}(s^{(\boldsymbol{Y},X_i)})_i^{q_i}$, na qual $q_i = \prod_{X_e \in Pa(s^{(\boldsymbol{Y},X_i)})_i} r_e$, para todo $i > R$. Seja também $\boldsymbol{Pa}(s^{\boldsymbol{Z}})_i$ o conjunto de pais contínuos de X_i, então o modelo de regressão linear para X_i segue a forma:

$$f(x_i \mid \boldsymbol{pa}(s^{(\boldsymbol{Y},X_i)})_i^j, \boldsymbol{pa}(s^{\boldsymbol{Z}})_i, \boldsymbol{\theta_i}, s^h) \sim N(x_i; m_i^j + \sum_{X_k \in \boldsymbol{Pa}(s^{\boldsymbol{Z}})_i} b_{ki}^j (x_k - m_k^j), v_i^j) \tag{2.22}$$

onde os parâmetros das funções de densidade de probabilidade para cada variável X_i são dados por $\boldsymbol{\theta_i} = (\boldsymbol{\theta_i^j})_{j=1}^{q_i}$ sendo $\boldsymbol{\theta_i^j} = (m_i^j, \boldsymbol{b}_i^j, v_i^j)$ para todo j, na qual: m_i^j é a média incondicional de X_i quando $\boldsymbol{Pa}(s^{(\boldsymbol{Y},X_i)})_i$ toma seu j-ésimo estado; $\boldsymbol{b}_i^j = (b_{1i}^j, ..., b_{i-1i}^j)$ é um vetor, na qual, se $X_k \in \boldsymbol{Pa}(s^{\boldsymbol{Z}})_i$ então b_{ki}^j é o coeficiente linear que reflete a força da relação entre X_k e X_i, em outro caso $b_{ki}^j = 0$, quando $\boldsymbol{Pa}(s^{(\boldsymbol{Y},X_i)})_i$ toma seu j-ésimo estado; v_i^j é a variância condicional de X_i dado

$\boldsymbol{Pa}(s^{\boldsymbol{Z}})_i$, quando $\boldsymbol{Pa}(s^{(Y,X_i)})_i$ toma seu j-ésimo estado.

Desta forma, para definir uma CGN, é necessário especificar:

- Um DAG, o qual codifica o conjunto de independências condicionais entre as variáveis aleatórias do problema em questão. O DAG segue as restrições impostas dadas anteriormente para modelos CGN.
- Um conjunto de distribuições de probabilidades, uma para cada variável discreta do modelo gráfico, isto é, θ_i^{jk} em (2.21), para $i \leq R$.
- Um conjunto de funções de densidade de probabilidade para as variáveis contínuas do modelo gráfico, isto é, médias incondicionais m_i^j, vetores de coeficientes lineares $\boldsymbol{b}_i^j$, e variâncias condicionais v_i^j, para todos i e j, tal que $i > R$.

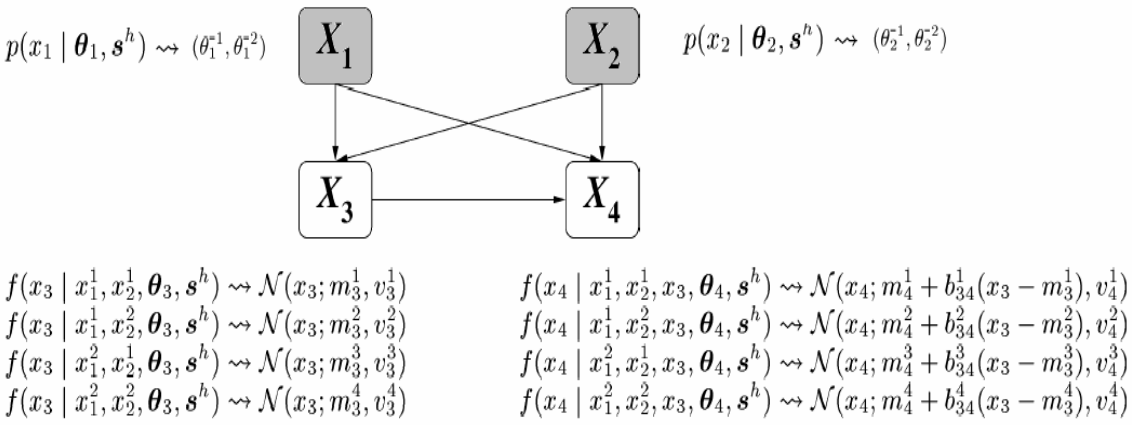


Figura 2.5: Exemplo de estrutura de modelo e distribuições de probabilidades locais para uma CGN. X_1 e X_2 são VA binárias, e X_3 e X_4 são VA contínuas.

A Figura 2.5 mostra um exemplo de estrutura e distribuições de probabilidades locais para uma CGN com 4 variáveis aleatórias $\boldsymbol{X} = (X_1, X_2, X_3, X_4)$, na qual, X_1 e X_2 são variáveis discretas binárias, e X_3 e X_4 são variáveis contínuas. A fatoração

gráfica da probabilidade conjunta para $\boldsymbol{X}$ é obtida segundo (2.20): $p(\boldsymbol{x} \mid \boldsymbol{\theta_s}, s^h) = p(x_1 \mid \boldsymbol{\theta_1}, s^h) \cdot p(x_2 \mid \boldsymbol{\theta_2}, s^h) \cdot f(x_3 \mid x_1, x_2, \boldsymbol{\theta_3}, s^h) \cdot f(x_4 \mid x_1, x_{2,} x_3, \boldsymbol{\theta_4}, s^h)$. Note-se a inexistência de independências condicionais induzidas pela estrutura do modelo de tipo $CI(X_i, (X_1, ...X_{i-1}) \setminus \boldsymbol{Pa}(s)_i \mid \boldsymbol{Pa}(s)_i)$.

2.2.3. REDES GAUSSIANAS CONDICIONAIS PARA AGRUPAMENTO DE DADOS

As redes Gaussianas condicionais são usadas no presente trabalho com o objetivo de realizar agrupamento de dados, para isso, são feitas as seguintes suposições: i) o banco de dados consiste de n casos $\boldsymbol{D} = \{x_1, ...x_n\}$, cada caso x_i é um vetor $x_i = (x_{i1}, ..., x_{iN+1}) = (c_i, y_{i1}, ..., y_{iN})$, no qual c_i indica a classe que gerou o caso, a qual é desconhecida e se tenta descobrir em problemas de agrupamentos. $y_i = (y_{i1}, ..., y_{iN})$ representa o vetor de N características observadas, chamados *atributos preditivos*; ii) É assumido que existem K classes em $\boldsymbol{D}$, as quais correspondem a diferentes processos físicos. A variável aleatória que descreve estes processos é denotada com $\boldsymbol{X} = (X_1, ..., X_{N+1})$, a qual pode ser dividida como $\boldsymbol{X} = (C, \boldsymbol{Y})$, isto é, a variável aleatória discreta da classe C, e uma variável aleatória contínua N-dimensional $\boldsymbol{Y} = (Y_1, ..., Y_N)$, a qual chama-se variável aleatória preditiva. É assumido que o mecanismo que gera cada caso em $\boldsymbol{D}$ trabalha em duas etapas: primeiro, um processo físico associado a uma classe é selecionado de acordo a uma distribuição de probabilidade, e segundo, uma instância é, de alguma forma, gerada de acordo à distribuição da probabilidade conjunta para $\boldsymbol{Y}$ do processo selecionado.

Uma CGN para agrupamento de dados é definida como na seção anterior (estrutura do modelo e conjunto de distribuições de probabilidade para essa estrutura) com a restrição adicional no grafo de que cada nó, que representa a

variável aleatória preditiva Y_i, depende (é filho) da variável aleatória discreta da classe C. Este requerimento é exigido pelas considerações feitas sobre o mecanismo que gera os dados.

A fatoração gráfica da distribuição da probabilidade conjunta para $\boldsymbol{X}$ codificada por uma CGN com a restrição imposta para agrupamento de dados adquire a seguinte forma (reescrita da Equação 2.20):

$$p(\boldsymbol{x} \mid \boldsymbol{\theta_s}, s^h) = p(c \mid \boldsymbol{\theta_s}, s^h) f(\boldsymbol{y} \mid c, \boldsymbol{\theta_s}, s^h) = p(c \mid \boldsymbol{\theta_C}, s^h) \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} f(y_i \mid c, \boldsymbol{pa}(s^Y)_i, \boldsymbol{\theta_i}, s^h). \quad (2.23)$$

Os parâmetros que definem a CGN são $\boldsymbol{\theta_s} = (\boldsymbol{\theta_C}, \boldsymbol{\theta_1}, ..., \boldsymbol{\theta_N})$, onde $\boldsymbol{\theta_C}$ são os parâmetros da distribuição de probabilidade (multinomial) para a variável aleatória da classe C, e $\boldsymbol{\theta_i}$ representa os parâmetros da função de densidade de probabilidade (modelo de regressão linear) para a variável aleatória contínua Y_i. A variável aleatória da classe pode tomar K valores denotados como $c^1, ..., c^K$, então, a distribuição multinomial consiste de um conjunto de probabilidades da forma:

$$p(c^g \mid \boldsymbol{\theta_C}, s^h) = \theta^g \quad (2.24)$$

na qual, $\theta^g > 0$, é a probabilidade de C tomar seu g-ésimo estado. Cumpre-se também que $\sum_{g=1}^{K} \theta^g = 1$, conseqüentemente, os parâmetros para C são $\boldsymbol{\theta_C} = (\theta^1, ..., \theta^K)$. Por outro lado, $f(y_i \mid c^g, \boldsymbol{pa}(s^Y)_i, \boldsymbol{\theta_i}, s^h)$ segue a seguinte forma para todo i, g e $\boldsymbol{pa}(s^Y)_i$:

$$f(y_i \mid c^g, \boldsymbol{pa}(s^Y)_i, \boldsymbol{\theta_i}, s^h) \sim N(y_i; m_i^g + \sum_{Y_k \in \boldsymbol{Pa}(s^{\boldsymbol{Z}})_i} b_{ki}^g (y_k - m_k^g), v_i^g) \quad (2.25)$$

em que os parâmetros das funções de densidade de probabilidade local para cada Y_i são dados por $\boldsymbol{\theta_i} = (\boldsymbol{\theta_i^g})_{g=1}^K$, sendo $\boldsymbol{\theta_i^g} = (m_i^g, \boldsymbol{b}_i^g, v_i^g)$ para todo g, na qual m_i^g

representa as médias incondicionais, $\boldsymbol{b}_i^g$ os vetores de coeficientes lineares, e v_i^g as variâncias condicionais, para todos i e g quando $\boldsymbol{Pa}(s^Y)_i$ toma seu $g-ésimo$ estado.

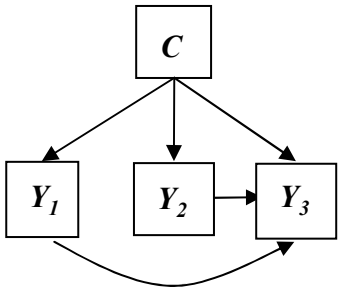
Estrutura do modelo	**Parâmetros do modelo**	
	Parâmetros	Distribuições de probabilidades
	$\boldsymbol{\theta_C} = (\theta^1, \theta^2)$	$p(c \mid \boldsymbol{\theta_C}, s^h)$
	$\boldsymbol{\theta_1} = (\boldsymbol{\theta_1^1}, \boldsymbol{\theta_1^2})$	
	$\boldsymbol{\theta_1^1} = (m_1^1, -, v_1^1)$	$f(y_1 \mid c^2, \boldsymbol{\theta_1}, s^h) \sim N(y_1; m_1^2, v_1^2)$
	$\boldsymbol{\theta_1^2} = (m_1^2, -, v_1^2)$	$f(y_1 \mid c^2, \boldsymbol{\theta_1}, s^h) \sim N(y_1; m_1^2, v_1^2)$
	$\boldsymbol{\theta_2} = (\boldsymbol{\theta_2^1}, \boldsymbol{\theta_2^2})$	
	$\boldsymbol{\theta_2^1} = (m_2^1, 0, v_2^1)$	$f(y_2 \mid c^1, \boldsymbol{\theta_2}, s^h) \sim N(y_2; m_2^1, v_2^1)$
	$\boldsymbol{\theta_2^2} = (m_2^2, 0, v_2^2)$	$f(y_2 \mid c^2, \boldsymbol{\theta_2}, s^h) \sim N(y_2; m_2^2, v_2^2)$
	$\boldsymbol{\theta_3} = (\boldsymbol{\theta_3^1}, \boldsymbol{\theta_3^2})$	
	$\boldsymbol{\theta_3^1} = (m_3^1, \boldsymbol{b}_3^1, v_3^1)$	$f(y_3 \mid c^1, y_1, y_2, \boldsymbol{\theta_3}, s^h) =$ $N(y_3; m_3^1 + b_{13}^1(y_1 - m_1^1) + b_{23}^1(y_2 - m_2^1), v_3^1)$
	$\boldsymbol{\theta_3^2} = (m_3^2, \boldsymbol{b}_3^2, v_3^2)$	$f(y_3 \mid c^2, y_1, y_2, \boldsymbol{\theta_3}, s^h) =$ $N(y_3; m_3^2 + b_{13}^2(y_1 - m_1^2) + b_{23}^2(y_2 - m_2^2), v_3^2)$

Figura 2.6: Exemplo de um modelo CGN para agrupamento de dados.

Na Figura 2.6 é mostrado um exemplo de CGN para classificação não supervisionada (PEÑA, 2001). Pode-se observar que o grafo segue a restrição imposta para classificação não supervisionada de que cada variável preditiva Y_i é filha da variável discreta da classe C.

Embora qualquer CGN com a restrição dada anteriormente pode ser usada para classificação não supervisionada, existem 3 tipos de especial interesse no presente trabalho: a *Naive Bayes - NB* (DUDA; HART,1973), a *Extended Naive Bayes - ENB* (PAZZANI,1996a; PAZZANI,1996b) e a *Tree Augmented Naive Bayes -*

TANB (FRIEDMAN; GOLDSZMIDT,1996). Estas CGNs apresentam um bom compromisso entre eficiência e efetividade no processo de aprendizado da estrutura do modelo. Outra vantagem é que a inferência nestas redes é relativamente mais fácil do que em redes densamente conectadas, onde a dificuldade é ainda mais agravada quando o número de casos ou atributos é grande. A seguir são apresentados brevemente estes três tipos de CGN.

***Naive Bayes* (NB)**

Este tipo de CGN é freqüentemente usado em problemas de classificação e de agrupamento de dados. Seu nome vem da consideração ingênua (*naive*) de que todas as variáveis aleatórias preditivas são condicionalmente independentes dada a variável da classe C. Conseqüentemente, a distribuição de probabilidade conjunta para $\boldsymbol{X} = (C, \boldsymbol{Y})$ que um modelo NB codifica, pode ser fatorada como:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta_s}, s^h) = p(c \mid \boldsymbol{\theta_s}, s^h) f(\boldsymbol{y} \mid c, \boldsymbol{\theta_s}, s^h) = p(c \mid \boldsymbol{\theta_C}, s^h) \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} f(y_i \mid c, \boldsymbol{\theta_i}, s^h) . \qquad (2.26)$$

Embora as considerações feitas na NB sejam freqüentemente irreais, pois as variáveis preditivas normalmente apresentam certas dependências, a NB apresenta resultados surpreendentemente bons em muitos domínios de aplicação (MICHIE *et al*.,1994).

Na Figura 2.7 é mostrado um exemplo de estrutura do modelo para um classificador NB com 3 variáveis aleatórias preditivas: $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \boldsymbol{Y_3}$, as quais, como pode ser observado, são unicamente filhas da variável da classe $\boldsymbol{C}$.

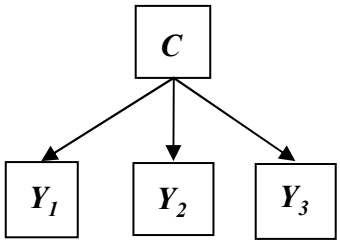


Figura 2.7: Estrutura do modelo para um classificador NB com 3 variáveis aleatórias preditivas.

Extended Naive Bayes (ENB)

Este modelo é muito similar à NB, a diferença é que o número de nós na ENB pode ser menor que o número de variáveis preditivas, isto porque algumas variáveis preditivas podem ser agrupadas como um único nó, chamado de super-nó. Para que um super-nó possa ser formado, cada variável no super-nó deve ter relacionamentos (arcos) com o resto das variáveis no super-nó e não possuir relacionamentos com as variáveis fora do super-nó, dada a variável da classe C.

A distribuição de probabilidade conjunta para $\boldsymbol{X} = (C, \boldsymbol{Y})$ que um modelo ENB codifica, pode ser fatorado como:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta_s}, s^h) = p(c \mid \boldsymbol{\theta_s}, s^h) f(\boldsymbol{y} \mid c, \boldsymbol{\theta_s}, s^h) = p(c \mid \boldsymbol{\theta_C}, s^h) \prod_{i=1}^{r} f(z_i \mid c, \boldsymbol{\theta_i}, s^h) \qquad (2.27)$$

na qual r é o número de super-nós denotados como $\boldsymbol{Z_i}$ onde se cumpre que: $\boldsymbol{Z_i} \subset \boldsymbol{Y}$, $\cup_{i=1}^{r} \boldsymbol{Z_i} = \boldsymbol{Y}$ e $\boldsymbol{Z_i} \cap \boldsymbol{Z_j} = \varnothing$ para todo $i \neq j$. $\boldsymbol{\theta_i}$ é o conjunto de parâmetros das distribuições de probabilidade para as variáveis agrupadas baixo o super-nó $\boldsymbol{Z_i}$.

Um exemplo de um ENB para agrupamento de dados com 6 variáveis preditivas é mostrado na Figura 2.8. Como pode ser observado, foram encontrados 3 super-nós $\boldsymbol{Z_1}, \boldsymbol{Z_2}, \boldsymbol{Z_3}$. O novo modelo construído com estes super-nós é um modelo NB simples, com cada nó representando uma estrutura local para as variáveis

aleatórias agrupadas, as quais não induzem nenhuma independência condicional entre estas dada a variável classe C .

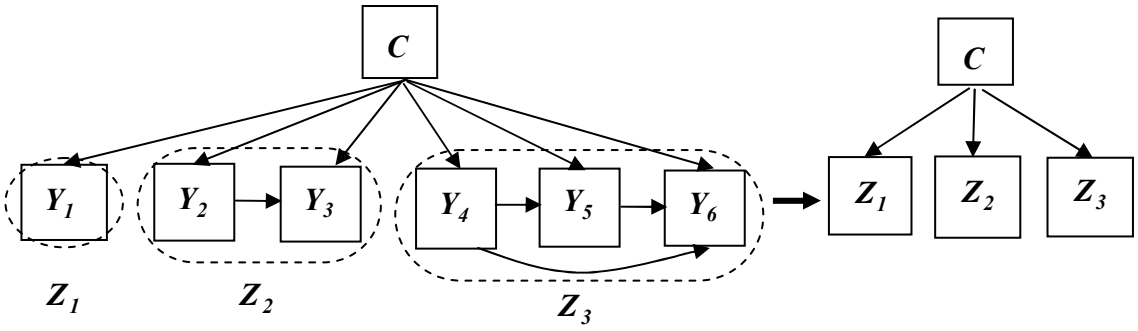


Figura 2.8: Estrutura do modelo para um classificador ENB com 6 variáveis preditivas agrupados em 3 super-nós e seu correspondente modelo NB.

Os modelos ENB para agrupamento de dados podem ser colocados em um lugar intermédiario entre modelos NB e modelos com todas as variáveis preditivas totalmente relacionadas, mantendo assim a simplicidade dos primeiros e a efetividade e expressividade dos últimos. Uma outra vantagem é encontrada na etapa de aprendizado estrutural pois a busca da estrutura de modelo ótima é realizada sobre o espaço de estruturas de modelo de tipo ENB, evitando assim a busca sobre o grande espaço de estruturas irrestritas e conseqüentemente reduzindo o esforço computacional desta etapa (PEÑA, 2001).

Tree Augmented Naive Bayes (TANB)

Este tipo de CGN é definido da seguinte forma:

i) cada variável preditiva tem a variável classe C como pai, e

ii) cada variável preditiva tem no máximo outra variável preditiva como pai.

Um exemplo de um modelo TANB é mostrado na Figura 2.9. onde pode ser

verificada a propriedade de que cada variável preditiva tem no máximo outra variável preditiva como pai.

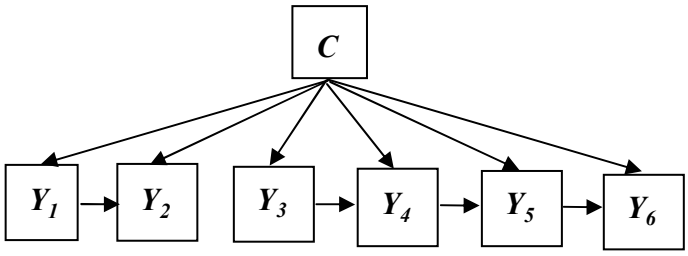


Figura 2.9: Estrutura do modelo para um classificador TANB com 6 variáveis preditivas.

Ao igual que os modelos ENB, os modelos TANB implicam vantagens computacionais no aprendizado da estrutura do modelo comparado com aprendizado de estruturas de modelos irrestritos. Também os modelos TANB apresentam uma maior expressividade quando comparados com modelos NB, mantendo a simplicidade destes últimos. Uma outra vantagem deste tipo de modelos é a facilidade de interpretação do modelo aprendido devido a que cada variável preditiva unicamente pode estar conectada a outra variável preditiva, sendo fácil encontrar as (in)dependências entre as variáveis, o que pode ser uma tarefa complicada em modelos densamente conectados (PEÑA, 2001).

Deve ser notado que não todas as independências condicionais codificadas por um modelo ENB podem ser representadas por um modelo TANB e vice-versa. Por exemplo, as independências condicionais codificadas num modelo TANB dado podem ser completamente representadas por um modelo ENB equivalente somente quando a longitude de cada caminho entre os nós das variáveis preditivas do modelo TANB é no máximo um. Por outro lado, todas as independências condicionais codificadas num modelo ENB dado podem ser representadas por um modelo TANB equivalente somente quando cada super-nó no modelo ENB agrupa

no máximo dois variáveis preditivas. Assim, a diferente expressividade dos modelos TANB e ENB pode implicar que os primeiros sejam mais adequados em certos problemas enquanto os segundos podem ser adequados em alguns outros problemas (PEÑA, 2001).

2.2.4. INFERÊNCIA EM REDES GAUSSIANAS CONDICIONAIS

Uma vez construída uma representação probabilística através de um modelo CGN, uma das tarefas mais importantes consiste em obter estimativas de probabilidades de eventos relacionados aos dados à medida que novas informações ou evidências sejam conhecidas. Esse processo é denominado inferência.

A inferência em MGP em geral e em CGN em particular permite, mediante o cálculo das probabilidades *a posteriori*, responder a uma série de "consultas" sobre um domínio de dados, a partir de nova informação (evidência) conhecida. As probabilidades de interesse que se tenta determinar no presente trabalho correspondem aos graus de crença das hipóteses feitas sobre um dado de ter sido gerado por diversas hipotéticas classes (processos físicos) dados os valores dos atributos preditivos (evidência).

A importância da inferência no presente trabalho não fica restrita ao cálculo de probabilidades de pertinência dos dados aos grupos, sua funcionalidade é aproveitada também no processo da estimativa de parâmetros do modelo, onde a inferência é utilizada para estimar os valores desconhecidos da variável classe C para assim ajustar os parâmetros do modelo num processo iterativo.

Existem basicamente dois tipos de métodos de inferência: os métodos exatos e os métodos aproximados. Nos métodos exatos, as probabilidades dos nós são calculadas sem outro erro senão o de arredondamento, inerente às limitações de

cálculos computacionais. Já os métodos aproximados utilizam diversas técnicas de simulação para obter valores aproximados das probabilidades, e são utilizados em casos em que os algoritmos exatos não são aplicáveis, ou o custo computacional é elevado.

No presente trabalho é utilizado o algoritmo de inferência exata "*árvore de junção",* proposto originalmente em (JENSEN *et al*.,1990) e posteriormente adaptado para modelos CGN em (LAURITZEN, 1992). Este algoritmo foi escolhido devido a sua eficiência computacional comparado com os outros algoritmos exatos (PASKIN, 2003).

Algoritmo de árvore de junção

Ao se obter uma evidência, ou seja, quando é conhecido um subconjunto de variáveis $E \subset \boldsymbol{X}$ com valores associados $X_i = e_i$ para $X_i \in E$, é preciso considerar se existe mais de um caminho entre o nó (nós) com a evidência e aquele cuja probabilidade deve ser atualizada pela inferência no modelo CGN. Para isso é necessária uma estrutura de controle que determine qual estratégia usar para propagar crenças ou probabilidades. O algoritmo de árvore de junção é um método geral de propagação de crenças em MGP, a qual utiliza uma estrutura de controle chamada *árvore de junção,* na qual os nós são determinados por subconjuntos de variáveis da estrutura do modelo original chamados *cliques*. Mediante esta estrutura, o método propaga as evidências, calculando probabilidades locais (com pequeno número de variáveis), evitando assim expressões globais (grande número de variáveis).

Um procedimento bastante usado na construção da árvore de junção para modelos CGN foi desenvolvido em (LAURITZEN, 1992). Nesse procedimento são

necessárias umas séries de transformações a partir do modelo CGN original para a obtenção da árvore de junção. Para uma melhor explicação do procedimento, é usado um exemplo fictício de um sistema incinerador de lixo tomado de (LAURITZEN, 1992). A Figura 2.10 mostra a estrutura do modelo extraída deste problema, onde os nós pretos indicam variáveis discretas, sendo F: estado do filtro {intacto, defeituoso}, W: tipo de lixo {industrial, caseiro} e B: regime de incineração {estável, instável}. Os nós brancos indicam variáveis contínuas, onde M_{in}: quantidade de metal no lixo, M_{out}: quantidade de metais pesados emitidos, D: quantidade de poeira emitida, C: quantidade de CO_2 emitido, e L: penetrabilidade da luz na poeira.

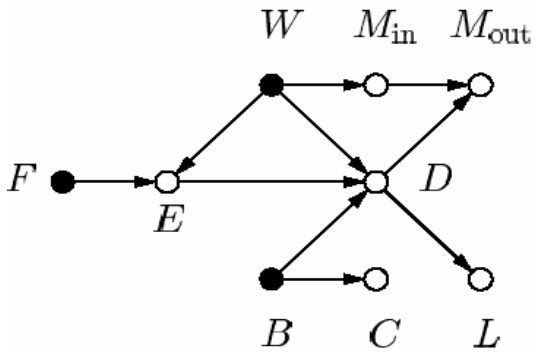


Figura 2.10: Estrutura do modelo de um sistema fictício de incineração de lixo (LAURITZEN, 1992) para explicar o procedimento de construção da árvore de junção.

Os passos para a construção da árvore de junção podem ser resumidos na construção das seguintes estruturas intermediarias (LAURITZEN, 1992):

a. Construção do grafo moralizado. Para construir o grafo moralizado primeiro se elimina a orientação dos arcos da estrutura original e logo se adiciona um arco não orientado (se inexistente) entre pares de nós pais comuns. Na Figura 2.11 é mostrado o grafo moralizado resultante para o exemplo de sistema de incineração

de lixo, onde os arcos em vermelho são os arcos adicionados.

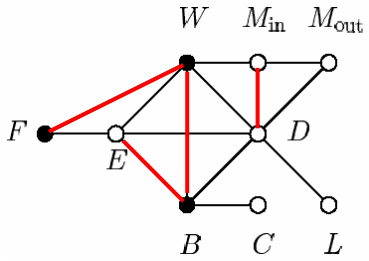


Figura 2.11: Grafo moralizado do sistema de sistema de incineração de lixo.

b. Formação de um grafo triangular. Consiste na introdução de arcos no grafo moralizado com a finalidade de evitar ciclos (caminhos fechados) com mais de três nós. Os arcos são denominados cordas e o grafo resultante, grafo triangular ou cordal. Existem diversas maneiras de se triangularizar um grafo, sendo ótima a que utiliza o mínimo possível de cordas. Um algoritmo ótimo é apresentado em (KJAERUL, 1990). No caso do sistema incinerador de lixo, o grafo moralizado (Figura 2.11) já é um grafo triangular, razão pela qual não é adicionado nenhum arco.

c. Formação de um grafo marcado decomponível. Grafos marcados são aqueles que possuem nós mistos que podem ser separados em nós discretos (Y) e nós contínuos (Z) como no caso das CGN. Para definir um grafo marcado decomponível é necessário introduzir a noção de decomposição em grafos marcados.

Uma *decomposição* de um grafo marcado não orientado G é uma partição (A,B,C) do seu conjunto de nós V, tal que: i) C separa A de B, ii) C é um subconjunto *completo* de V, isto é, cada par de nós em C é conectado por um arco iii) $C \subseteq Y \vee B \subseteq Z$. Quando uma decomposição (A,B,C) é encontrada em G diz-se que (A,B,C) decompõe G em componentes $G_{A\cup C}$ e $G_{B\cup C}$.

Um grafo marcado G é dito *decomponível* se este é completo, ou se existe uma decomposição (A,B,C) que decompõe G em sub-grafos $G_{A\cup C}$ e $G_{B\cup C}$ que são também decomponíveis.

Para que um grafo marcado seja decomponível, este tem que ser triangularizado (passo prévio), e não possuir caminhos entre dois nós discretos passando unicamente por nós contínuos, não sendo vizinhos os nós discretos. Para isso é necessário adicionar arcos entre aqueles nós discretos não vizinhos que possuem os caminhos mencionados.

Na Figura 2.12 é mostrado o grafo marcado decomponível do exemplo de sistema de incineração de lixo, onde foi adicionado o arco entre B e F para eliminar o caminho proibido (B,E,F) e tornar o grafo marcado em decomponível.

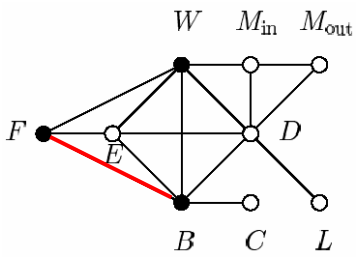


Figura 2.12: Grafo marcado decomponível do sistema de incineração de lixo.

d. Criação da árvore de junção. Uma árvore de junção é uma organização em forma de árvore de grupos de nós selecionados do grafo marcado decomponível chamados *cliques.* Um *clique* em um grafo não dirigido G é um subconjunto de nós de G que é completo e máximo. Completo significa que cada par de nós é conectado por um arco. Máximo significa que cada *clique* não está contido em um subconjunto completo maior. A Figura 2.13 mostra alguns *cliques* identificados no grafo marcado decomponível do sistema de incineração de lixo.

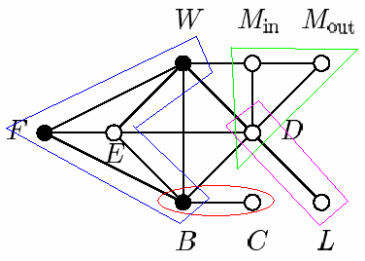


Figura 2.13: Alguns *cliques* identificados no grafo marcado decomponível do sistema de incineração de lixo.

A árvore de junção é formada por *cliques* do grafo marcado decomponível, na qual se cumpre a propriedade de que para cada par de *cliques* V_i e V_k, todos os *cliques* no caminho entre eles contêm $V_i \cap V_k$. Os arcos entre *cliques* vizinhos são rotulados com as variáveis comuns que os une, este conjunto de variáveis comuns é chamado separador.

No caso específico de modelos CGN, devido à assimetria entre as variáveis discretas $(\boldsymbol{Y})$ e as contínuas ($\boldsymbol{Z}$), é necessário uma condição adicional para que o esquema de propagação trabalhe apropriadamente. Esta condição identifica um *clique* V_r como raiz (raiz forte) na árvore de junção, si se cumpre que para qualquer par de *cliques* vizinhos V_i, V_j na árvore com V_i mais próximo a V_r do que V_j, se satisfaz que $(V_j \setminus V_i) \subseteq \boldsymbol{Z} \vee (V_j \cap V_i) \subseteq \boldsymbol{Y}$. Isto significa que quando um separador entre dois *cliques* vizinhos não é só discreto, o *clique* mais longe da raiz possui somente variáveis contínuas além do separador.

Na Figura 2.14 é mostrada a árvore de junção para o exemplo de sistema incinerador de lixo. Onde os retângulos indicam os separadores. O *clique* {W,E,B,F} pode ser usado como raiz da árvore. Por exemplo, {W,M_{in},D} tem unicamente a variável contínua M_{in} além do separador {W,D}.

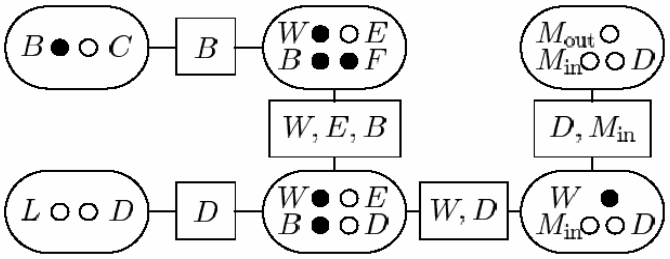


Figura 2.14: árvore de junção do sistema de incineração de lixo.

Inferência com árvores de junção em modelos CGN

Para descrever o processo de inferência considera-se que tanto a estrutura do modelo, quanto os parâmetros do modelo CGN já foram definidos previamente. A forma de se encontrar estes parâmetros é descrita posteriormente na parte de aprendizagem de redes gaussianas condicionais.

A inferência em uma árvore de junção é descrita utilizando a noção de potencial, que para o caso de modelos CGN é chamado CG-potencial. Um CG-potencial é definido sobre um *clique* ou separador da árvore de junção, mapeando cada instanciação dos valores das variáveis que o formam em um número real. Um CG-potencial é qualquer função ϕ da seguinte forma:

$$\phi(\boldsymbol{x}) = \phi(i, \boldsymbol{y}) = \chi(i) \exp\{g(i) + \boldsymbol{h}(i)^T \boldsymbol{y} - 0.5 \boldsymbol{y}^T \boldsymbol{K}(i) \boldsymbol{y}\} \tag{2.28}$$

em que, i denota o estado das variáveis discretas no CG-potencial, $\boldsymbol{y}$ o valor das variáveis contínuas e $\chi(i) = \{0,1\}$. A forma da função de um CG-potencial lembra uma distribuição gaussiana condicionada (na sua forma canônica), a diferença é que a matriz $\boldsymbol{K}(i)$ é considerada simétrica, mas não necessariamente positiva definida. Um CG-potencial é definido por suas características canônicas $(g, \boldsymbol{h}, \boldsymbol{K})$. Se $\boldsymbol{K}(i)$ é positiva definida, então o CG-potencial é uma distribuição gaussiana condicional que pode ser definida também por suas características de momentos ($\boldsymbol{\mu}(i), \boldsymbol{\Sigma}(i)$), onde:

$\boldsymbol{\mu}(i) = \boldsymbol{K}(i)^{-1}\boldsymbol{h}(i)$, $\boldsymbol{\Sigma}(i) = \boldsymbol{K}(i)^{-1}$.

Em um CG-potencial podem ser definidas as seguintes operações:

- ***Extensão:*** se $(g, \boldsymbol{h}, \boldsymbol{K})$ são as características de um CG-potencial ϕ definido nas variáveis $(i, \boldsymbol{y})$, algumas vezes é necessário estender o CG-potencial para um maior espaço de variáveis $(i, j, \boldsymbol{y}, \boldsymbol{z})$, este novo CG-potencial é denotado como $\bar{\phi}$, sendo as suas características canônicas $(\bar{g}(i), \bar{\boldsymbol{h}}(i), \bar{\boldsymbol{K}}(i))$, com $\bar{g}(i) = g(i)$, $\bar{\boldsymbol{h}}(i) = \begin{pmatrix} \boldsymbol{h}(i) \\ \boldsymbol{0} \end{pmatrix}$ e $\bar{\boldsymbol{K}}(i) = \begin{pmatrix} \boldsymbol{K}(i) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$

- ***Multiplicação e divisão:*** estas operações são definidas uma vez que os potenciais têm sido estendidos ao mesmo espaço de variáveis. A multiplicação é a soma das suas características canônicas: $(g_1, \boldsymbol{h_1}, \boldsymbol{K_1}) * (g_2, \boldsymbol{h_2}, \boldsymbol{K_2}) = (g_1 + g_2, \boldsymbol{h_1} + \boldsymbol{h_2}, \boldsymbol{K_1} + \boldsymbol{K_2})$
  A divisão é similarmente definida como a subtração das características canônicas,

- ***Marginalização:*** Dependendo do tipo de variável que se quer marginalizar existem vários casos. **Primeiro**, marginalizar variáveis contínuas corresponde à integração, seja $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{pmatrix}$, $\boldsymbol{h} = \begin{pmatrix} \boldsymbol{h_1} \\ \boldsymbol{h_2} \end{pmatrix}$, $\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K_{11}} & \boldsymbol{K_{12}} \\ \boldsymbol{K_{21}} & \boldsymbol{K_{22}} \end{pmatrix}$, onde $\boldsymbol{y_1}$ é de dimensão p e $\boldsymbol{y_2}$ de dimensão q. A integral sobre $\boldsymbol{y_1}$: $\int \phi(i, \boldsymbol{y_1}, \boldsymbol{y_2}) d\boldsymbol{y_1}$ é finita e igual ao CG-potencial $\tilde{\phi}$ com as seguintes características canônicas: $\tilde{g}(i) = g(i) + \{p \log(2\pi\} - \log(\det \boldsymbol{K_{11}}(i)) + h_1(i)^T \boldsymbol{K_{11}}(i)^{-1} h_1(i)\}/2$, $\tilde{\boldsymbol{h}}(i) = \boldsymbol{h_2}(i) - \boldsymbol{K_{21}}(i)\boldsymbol{K_{11}}(i)^{-1}\boldsymbol{h_1}(i)$, $\tilde{\boldsymbol{K}}(i) = \boldsymbol{K}_{22}(i) - \boldsymbol{K}_{21}(i)\boldsymbol{K}_{11}(i)^{-1}\boldsymbol{K}_{12}(i)$. **Segundo**, marginalizar variáveis discretas j, onde $\boldsymbol{K}$ não depende de j, isto é $\boldsymbol{h}(i, j) \equiv \boldsymbol{h}(i)$ e $\boldsymbol{K}(i, j) \equiv \boldsymbol{K}(i)$, resulta em um CG-potencial marginal com características canônicas: $\tilde{g}(i) = \log \sum_{j : \mathcal{X}(i,j)=1} \exp g(i, j)$, $\tilde{\boldsymbol{h}}(i) = \boldsymbol{h}(i)$ e $\tilde{\boldsymbol{K}}(i) = \boldsymbol{K}(i)$.

Terceiro, marginalizar variáveis discretas j, onde $\boldsymbol{K}$ depende em j, não resulta em um CG-potencial. A marginalização é somente definida para $\boldsymbol{K}(i,j)$ positiva definida, portanto é melhor defini-lo em termos das características de momentos. O marginal resultante é $\widetilde{\phi}$ com características de momentos $\{p,\boldsymbol{\mu},\boldsymbol{\Sigma}\}$, na qual $\widetilde{p}(i)=\sum_j p(i,j)$, $\widetilde{\boldsymbol{\mu}}(i)=\sum_j \boldsymbol{\mu}(i,j)p(i,j)/\widetilde{p}(i)$ e $\widetilde{\boldsymbol{\Sigma}}(i)=\sum_j \boldsymbol{\Sigma}(i,j)p(i,j)/\widetilde{p}(i)+\sum_j (\boldsymbol{\mu}(i,j)-\boldsymbol{\mu}(i))^T(\boldsymbol{\mu}(i,j)-\boldsymbol{\mu}(i))p(i,j)/\widetilde{p}(i)$. **Quarto**, no caso de variáveis mistas, primeiro são marginalizadas as variáveis contínuas para logo marginalizar as discretas.

Cada *clique* na árvore de junção é chamado *universo de crença.* O conjunto de todas as variáveis é chamado *universo total.* A coleção de todos os universos de crença é denotado por C e a coleção dos separadores por S. Cada *clique* $V \in C$ está associado com um CG-potencial ϕ_V e cada separador $S \in S$ está associado com um CG-potencial ϕ_S. Estes CG-potenciais são denominados *potencias de crença*. O *sistema de crença conjunta* ϕ_U do universo total é definido como:

$$\phi_U = \frac{\prod_{V\in C}\phi_V}{\prod_{S\in S}\phi_S} \tag{2.29}$$

na qual, ϕ_U é proporcional à densidade de probabilidade conjunta de todas as variáveis.

No inicio, a árvore de junção tem que ser iniciada de acordo com os parâmetros especificados no modelo CGN, isso é necessário para assegurar que o sistema de crença conjunta ϕ_U seja a distribuição de probabilidade conjunta codificada pela CGN. Uma das formas de iniciar a árvore de junção é atribuir cada nó A do DAG original a um dos *cliques* V da árvore de junção, tal que

$(A \cup Pa(A)) \subseteq V$. Assim, o CG-potencial ϕ_V de cada *clique* V é iniciado com o produto dos CG-potenciais (distribuições condicionais) dos nós atribuídos ao *clique*, fazendo a operação de extensão se necessário. Os CG-potenciais dos separadores são iniciados com valor 1, assim como os *cliques* sem nós atribuídos.

Depois de iniciada a árvore de junção, são necessários os seguintes passos para realizar o processo de inferência (LAURITZEN, 1992):

- Introdução da evidência em todos os *cliques* e separadores da árvore de junção. Isto implica na modificação dos potenciais de crença iniciais.
- Execução da operação de coleção de evidência desde a raiz da árvore de junção. Esta operação é chamada *CollectEvidence*. As mensagens são formadas marginalizando primeiro as variáveis redundantes contínuas, e logo sumarizando as variáveis discretas.
- Uma vez que a raiz tenha absorvido toda a informação disponível, esta deve ser distribuída aos restantes universos de crença na árvore de junção, esta operação de distribuição é chamada *DistributeEvidence*.

Uma vez realizados os passos anteriores, pode ser calculada a distribuição de probabilidade atualizada de qualquer variável de interesse. Simplesmente se tem que marginalizar a variável em qualquer universo de crença que a contenha.

2.2.5. APRENDIZAGEM DE REDES GAUSSIANAS CONDICIONAIS

Em algumas situações é possível construir toda a rede a partir do conhecimento de um especialista, porém, dependendo do domínio a ser modelado, este pode ser um processo difícil e demorado, sobretudo se o número de variáveis é grande. Considerando isso e o fato que os dados são cada vez mais acessíveis e baratos, é que atualmente há um grande interesse no desenvolvimento e

aperfeiçoamento de métodos que aprendam as estruturas e os parâmetros a partir dos dados.

No presente trabalho assume-se que a estrutura do modelo já foi especificada por algum especialista ou escolhido arbitrariamente, como os modelos apresentados na Seção 2.2.3. Portanto, o foco desta seção é apresentar um método de aprendizado de parâmetros em redes CGN com estrutura de modelo já definida. Particularmente o presente trabalho é focado em aprendizado de parâmetros com dados incompletos, isto é, algumas das variáveis não são observadas ou estão ocultas em algumas instâncias dos dados. Este interesse se deve ao fato de que agrupamento de dados é um caso particular de aprendizado com dados incompletos, em que somente uma variável (classe) é desconhecida em todas as instâncias.

O algoritmo *Expectation Maximization - EM* (DEMPSTER *et al*.,1977) é um método para estimar funções de máxima verossimilhança a partir de dados incompletos. Quando os dados são incompletos utilizam-se os casos em que foram observadas as variáveis para aprender a predizer seus valores quando não observados. Também pode ser utilizado para variáveis cujos valores nunca foram observados, fazendo certas considerações sobre a forma geral da distribuição de probabilidade dessas variáveis. As características deste algoritmo permitem usá-lo em problemas de aprendizagem de CGN a partir de dados incompletos, como no caso do presente trabalho.

EM para aprendizagem de parâmetros em CGN

O algoritmo EM para aprendizagem de parâmetros em MGP quando a estrutura é conhecida e todas as variáveis são discretas foi proposto em

(LAURITZEN, 1995). Posteriormente em (MCMICHAEL *et al.*, 1999) estenderam este algoritmo para aprendizado de paramétrios em modelos CGN. A seguir é detalhado este algoritmo.

Considere-se $D = \{\boldsymbol{d}_1,...,\boldsymbol{d}_e,...,\boldsymbol{d}_n\}$ o conjunto de dados incompletos (evidências), onde cada dado é o vetor $\boldsymbol{d}_e = (d_{ie})_{i=1}^{N}$, sendo d_{ie} o valor da i-ésima variável do dado $\boldsymbol{d}_e$. Assumindo dados condicionalmente independentes, a função de verossimilhança de uma rede CGN com N variáveis e estrutura de modelo s é:

$$L = \sum_{i=1}^{N} \sum_{e=1}^{n} \log p(d_{ie} \mid d_{ie}^{+}, \boldsymbol{\theta_i}) \tag{2.30}$$

na qual, d_{ie}^{+} são os valores das variáveis pais $\boldsymbol{Pa}(s)_i$ da i-ésima variável do dado $\boldsymbol{d}_e$, $\boldsymbol{\theta_i}$ são os parâmetros locais da i-ésima variável na rede. Por outro lado denotemos a w_{ie}^{jk} como uma variável indicadora que vale 1 no caso que a i-ésima variável do dado $\boldsymbol{d}_e$ tenha valor k e o estado das variáveis pais d_{ie}^{+} seja j, caso contrário vale 0.

Se o i-ésimo termo de (2.30) corresponde a uma variável discreta, então a função de verossimilhança para esta variável é:

$$L_i = \sum_{e=1}^{n} \sum_{j=1}^{q_i} \sum_{i=1}^{r_i} w_{ie}^{jk} \log \theta_i^{jk} \tag{2.31}$$

da Seção 2.2.2, r_i é o número de valores distintos da variável X_i, q_i é o número de estados distintos das variáveis pais $\boldsymbol{Pa}(s)_i$, e cada parâmetro θ_i^{jk} representa a probabilidade condicional de que X_i tome seu k-ésimo valor dado que $\boldsymbol{Pa}(s)_i$ toma seu j-ésimo valor, isto é, $\theta_i^{jk} = p(x_i^k \mid \boldsymbol{pa}(s)_i^j, \boldsymbol{\theta_i}, s)$. O valor esperado do indicador w_{ie}^{jk} é a probabilidade condicional $p(x_i^k, \boldsymbol{pa}(s)_i^j \mid d_e, \boldsymbol{\theta'}, s)$, onde $\boldsymbol{\theta'}$ indica os

parâmetros estimados na iteração EM previa. O valor esperado $\overline{N}_i^{jk}$ do número de casos em que a variável X_i toma seu valor k, e $\boldsymbol{Pa}(s)_i$ toma seu valor j, também chamado como *estatísticas suficientes*, é dado por:

$$\overline{N}_i^{jk} = \sum_{e=1}^{n} \sum_{j=1}^{q_i} \sum_{i=1}^{r_i} p(x_i^k, \boldsymbol{pa}(s)_i^j \mid d_e, \boldsymbol{\theta'}, s) \tag{2.32}$$

No caso de dados discretos faltosos ou ocultos, $p(x_i^k, \boldsymbol{pa}(s)_i^j \mid \boldsymbol{d}_e, \boldsymbol{\theta'}, s)$ pode ser calculada usando o algoritmo de árvore de junção. Um resultado comum em distribuições normais é que a estimativa de máxima verossimilhança pode ser encontrada igualando as mínimas estatísticas suficientes com seus valores esperados, assim o passo de maximização (M) para variáveis discretas é sumarizado por:

$$\hat{\theta}_i^{jk} = \frac{\overline{N}_i^{jk}}{\sum_{k=1}^{r_i} \overline{N}_i^{jk}} \tag{2.33}$$

No caso que o i-ésimo termo em (2.30) corresponda a uma variável contínua X_i, as variáveis pais podem ser divididas em variáveis contínuas denotados por Z_i e variáveis discretas Y_i. O valor da variável contínua Z_i no dado $\boldsymbol{d}_e$ é denotado como z_{ie}. A função de densidade de probabilidade para esta variável contínua é $N(x_{ie}; m_i^j + \sum_{Z_i} b_{ki}^j (z_{ie} - m_k^j), v_i^j)$ (2.22), re-escrevendo de outra forma, esta distribuição pode ser expressada como $N(x_{ie}; B_i^j \times \tilde{z}_{ie}, v_i^j)$, na qual $\tilde{z}_{ie} = [z_{ie}{}^T, 1]^T$, $B_i^j = [(A_i^j)^T, (m_i^j)^T]$, sendo A_i^j a matriz de regressão formada pelos coeficientes de regressão linear b_{ki}^j. Então a função de verossimilhança para esta variável X_i é:

$$L_i = -0.5 \sum_{e=1}^{n} \sum_{j=1}^{q_i} w_{ie}^{jk} (\log \left| 2\pi\, v_i^j \right| + (x_{ie} - B_i^j \times \tilde{z}_{ie})^T (v_i^j)^{-1} (x_{ie} - B_i^j \times \tilde{z}_{ie})) \tag{2.34}$$

pode ser mostrado que o segundo termo desta equação fatoriza como:

$$L_i = -0.5(\sum_{j=1}^{q_i} N_i^j \log\left|2\pi\, v_i^j\right| + Tr\{(v_i^j)^{-1}[(B_i^j - S^j_{x_i,\tilde{z}_i} \times (S^j_{\tilde{z}_i,\tilde{z}_i})^{-1})$$

$$\times S^j_{\tilde{z}_i,\tilde{z}_i}(B_i^j - S^j_{x_i,\tilde{z}_i} \times (S^j_{\tilde{z}_i,\tilde{z}_i})^{-1})^T + S^j_{x_i,x_i} - S^j_{x_i,\tilde{z}_i}(S^j_{\tilde{z}_i,\tilde{z}_i})^{-1} S^j_{x_i,\tilde{z}_i}]\}) \qquad (2.35)$$

na qual, $N_i^j = \sum_{e=1}^{n} w_{ie}^j$, e $S^j_{a,b} = \sum_{e=1}^{n} w_{ie}^j a_e (b_e{}^T)$, sendo a_e e b_e variáveis simuladas. A estimativa de máxima verossimilhança para B_i^j é:

$$\hat{B}_i^j = S^j_{x_i,\tilde{z}_i} \times (S^j_{\tilde{z}_i,\tilde{z}_i})^{-1} \qquad (2.36)$$

por outro lado, a matriz de covariância é estimada derivando com respeito a $(v_i^j)^{-1}$, obtendo-se:

$$(\hat{v}_i^j)^{-1} = \frac{S^j_{x_i,x_i} - S^j_{x_i,\tilde{z}_i}(S^j_{\tilde{z}_i,\tilde{z}_i})^{-1}(S^j_{x_i,\tilde{z}_i})^T}{N_i^j}. \qquad (2.37)$$

Sumarizando, no passo *Expectation* é usado o algoritmo árvore de junção para estimar as variáveis faltosas ou ocultas. Na etapa *Maximization* são calculadas as estimativas dos parâmetros de máxima verossimilhança dadas em (2.33), (2.36) e (2.37) de acordo com o tipo de variável. O processo é repetido até chegar a um número máximo de iterações ou até que o logaritmo da verossimilhança entre duas iterações sucessivas seja menor que um limiar.

Capítulo 3

MATERIAIS E MÉTODOS

Neste capítulo são descritos os métodos de agrupamento implementados, os bancos de dados onde foram testados e os experimentos realizados no presente trabalho. Na Seção 3.1 são descritos os bancos de dados usados, os quais são divididos em: bancos de dados de validação, e um banco de dados genotípicos, o qual constitui o problema de aplicação deste trabalho. Na Seção 3.2 é descrito o método de agrupamento hierárquico Bayesiano. Na Seção 3.3 é descrito o segundo método implementado: o método de agrupamento baseado em redes gaussianas condicionais. O planejamento dos experimentos é descrito na Seção 3.4.

3.1. BANCOS DE DADOS

Nesta seção são descritos os bancos de dados usados para testar os métodos de agrupamento implementados. Com o objetivo de validar os métodos e obter uma idéia do desempenho dos mesmos, realizou-se primeiro uma análise de agrupamentos sobre 3 bancos de dados já classificados, os quais são chamados *bancos de dados de validação*. Nestes bancos de dados é conhecida a informação da classe porém esta não é usada para realizar o agrupamento, somente é usada para avaliar a proximidade dos grupos encontrados com respeito às classes

verdadeiras. Após desta validação inicial, os métodos foram usados em um problema de aplicação real: a taxonomia de uma coleção brasileira de estirpes de bactérias fixadoras de nitrogênio pertencentes ao gênero *Bradyrhizobium.* Este banco de dados é composto por dados genotípicos resultantes da análise de certas regiões do RNA ribossômico (abreviado por rRNA) extraído das estirpes.

3.1.1. BANCOS DE DADOS DE VALIDAÇÃO

Synthetic-2000

Este banco de dados foi construído artificialmente com 2000 pontos de 4 dimensões gerados aleatoriamente de uma mistura de 4 distribuições de probabilidade normais (componentes), as quais são as classes a serem encontradas. Os parâmetros destas distribuições são mostrados na Tabela 3.1 para cada uma das componentes $G_1,...,G_4$.

Tabela 3.1 – Parâmetros das 4 distribuições gaussianas usadas para gerar os 2000 pontos de Synthetic-2000.

Parâmetros	Componentes			
	G1	G2	G3	G4
Proporção	0.25	0.24	0.26	0.25
Média X_1	0.0584	0.7276	0.4014	0.3124
Média X_2	0.8979	0.4873	0.5106	0.2555
Média X_3	0.9803	0.6801	0.0512	0.6078
Média X_4	0.6578	0.3650	0.9957	0.4415
Variância	0.0909	0.0937	0.1026	0.0845

Na Figura 3.1 é mostrado o gráfico dos 2000 pontos usando as 3 primeiras coordenadas, onde cada ponto foi pintado de acordo com o componente que o

gerou. Como pode ser observado, os 4 componentes, que constituem as 4 classes a serem descobertas, estão visualmente bem separadas. Este banco é usado para testar o comportamento dos métodos de agrupamento em dados com classes claramente separadas.

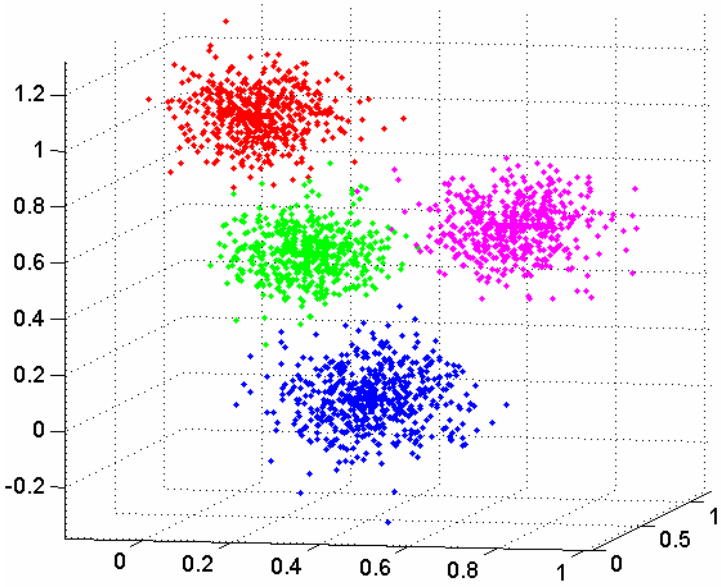


Figura 3.1: Distribuição dos pontos do banco de dados Synthetic-2000 considerando as 3 primeiras coordenadas. Cada cor indica um componente.

Synthetic-1000

Este banco de dados foi construído artificialmente com 1000 pontos em 3 dimensões gerados de uma mistura de 5 componentes. Os parâmetros destes componentes são mostrados na Tabela 3.2. Na Figura 3.2 é mostrado o gráfico dos 1000 pontos integrantes, onde cada ponto foi pintado de acordo com o componente que o gerou. Como pode ser observado, existem alguns componentes que estão medianamente sobrepostos. Este banco é usado para avaliar os métodos em dados com classes sobrepostas.

Tabela 3.2 – Parâmetros das 5 distribuições normais usadas para gerar os 1000 pontos de Synthetic-1000.

Parâmetros	**Componentes**				
	G1	G2	G2	G4	G5
Proporção	0.19	0.21	0.21	0.21	0.19
Média X_1	0.4501	0.7082	0.7557	0.2355	0.4811
Media X_2	0.3983	0.1062	0.1562	0.7252	0.8636
Media X_3	0.4121	0.3780	0.6606	0.2536	0.8326
Variância	0.0963	0.0932	0.0905	0.0929	0.1052

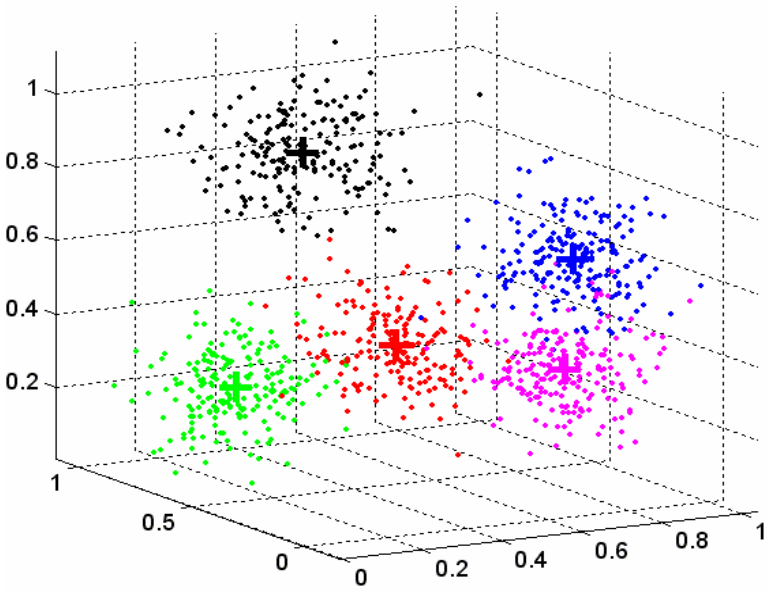


Figura 3.2: Distribuição dos pontos do banco de dados Synthetic-1000. Cada ponto foi pintado de acordo com o componente que o gerou.

Iris

Este banco de dados é amplamente conhecido na área de mineração de dados e aprendizado de máquina (DASARATHY, 1980). É composto por 150 registros divididos igualmente em 3 classes: virgínica, versicolor e setosa (esta linearmente separável das outras duas), as quais são tipos da flor conhecida por Iris. Os registros são definidos em termos de quatro atributos numéricos que trazem as informações

de comprimento e largura de sépala e comprimento e largura de pétala. A Figura 3.3 mostra o gráfico dos 150 pontos usando os 3 primeiros atributos, onde pode ser observado que duas classes estão sobrepostas. Ao contrário dos bancos de dados anteriores, este é um banco real, onde as classes podem não seguir necessariamente distribuições de probabilidades normais. Com este banco de dados se tenta avaliar o comportamento dos métodos de agrupamento em dados reais com classes sobrepostas e que não seguem necessariamente distribuições normais.

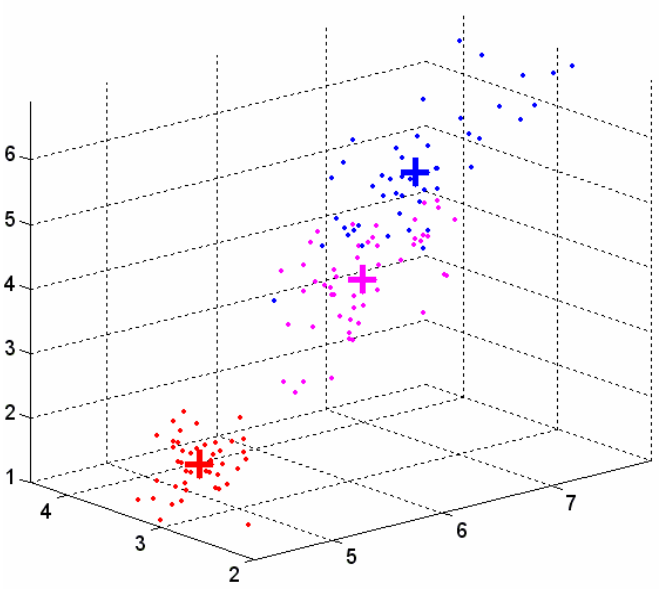


Figura 3.3: Pontos do banco de dados Íris considerando os 3 primeiros atributos. Cada ponto foi pintado de acordo a sua classe.

3.1.2. BANCO DE DADOS DE APLICAÇÃO: ESTIRPES DE BRADYRHIZOBIUM

Para um melhor entendimento deste banco de dados, é apresentado primeiro o contexto do problema no qual este banco de dados está envolvido para logo descrever o banco de dados.

O nitrogênio (N) é um dos elementos mais importante para os seres vivos, formando parte das moléculas orgânicas vitais. Este elemento existe em abundância na atmosfera terrestre (80% do ar) na forma de N_2, porém, é um dos elementos

mais escassos do solo, interferindo no crescimento das plantas (HUNGRIA *et al*., 1997).

No Brasil os teores de nitrogênios do solo não são elevados (entre 0.03% a 0.05%) o que leva à necessidade de enriquecer o solo com fontes alternativas como fertilização e fixação biológica. A fertilização é um processo dispendioso já que requer muita energia na produção de fertilizantes, tornando-os caros. Por outro lado, a fixação biológica é o sistema natural de transformação do N_2 em formas assimiláveis pelas plantas através de bactérias que assimilam o nitrogênio diretamente do ar (HUNGRIA *et al*., 1997).

Os rizóbios são um tipo de bactéria que assimila o nitrogênio somente quando em simbiose com plantas hospedeiras (leguminosas). A simbiose consiste em que a planta cria um lugar especial para os rizóbios viverem, formando os chamamos nódulos radiculares e fornecendo carboidratos que os rizóbios usam para obter energia e assim poder reduzir o N_2 do ar em compostos nitrogenados que são transferidos para a planta hospedeira e o solo. O *Bradyrhizobium* (do grego: bradus=lento; logo, bactérias de crescimento lento) é um gênero de rizóbios, cujas espécies *Bradyrhizobium japonicum* e *Bradyrhizobium elkanii* nodulam a soja. Essas espécies são divididas em estirpes, variando entre elas a quantidade de nitrogênio fixado. Algumas estirpes fixam todo o nitrogênio necessário para a produção normal da planta e outras não fixam praticamente nenhum nitrogênio (HUNGRIA *et al*., 1997).

A taxonomia das bactérias pertencentes ao gênero *Bradyrhizobium* é ainda pouco refinada. Tradicionalmente foram classificadas com base na habilidade de nodular uma espécie específica de planta hospedeira (JORDAM, 1938 [1] apud

[1] JORDAM, D. C. (1938). Rhizobiaceae Conn.

KRIEG; HOLT,1984). Posteriormente a análise dos genes ribossomais mostrou que várias espécies eram completamente diferentes tendo mais parentesco com bactérias não fixadoras de N_2 (GARRITY; HOLT, 2001). Devido a estes resultados conflitantes é que se escolheu o banco de dados de estirpes de *Bradyrhizobium* como problema de aplicação dos métodos implementados no presente trabalho.

O banco de dados de estirpes de *Bradyrhizobium* consiste em dados genéticos de uma coleção brasileira de 128 estirpes que exibem características fenotípicas de espécies *japonicum* e *elkanii*. Estas estirpes foram isoladas de 33 espécies de leguminosas tropicais. Os dados foram obtidos pelo laboratório de Biotecnologia do Solo da Empresa Brasileira de Pesquisa Agropecuária – Centro Nacional de Pesquisa de Soja (Embrapa Soja), em Warta, distrito de Londrina, Paraná. Para maiores detalhes do procedimento de obtenção deste banco de dados, revisar (GERMANO *et al*., 2006).

Neste banco de dados, cada estirpe é descrita por 9 imagens de eletroforese de Gel, as quais correspondem ao resultado da análise do rRNA pela técnica RFLP-PCR (*Restriction Fragment Length Polymorphism - Polymerase Chain Reaction*) (DAVISON, 2006). A seguir descrevem-se brevemente os passos envolvidos nesta técnica:

- ***Extração de DNA***: Aqui se utilizam processos de rompimento de células, centrifugação e substâncias que são capazes de desnaturar e retirar as proteínas que estão acopladas ao DNA ou RNA.
- ***Reação em cadeia de DNA Polimerase (PCR):*** Neste passo, uma seqüência específica de DNA é amplificada em ciclos repetidos de desnaturação, hibridação e extensão. A quantidade de DNA é dobrada em cada ciclo.
- ***Polimorfismo de tamanho de fragmentos de DNA:*** Aqui se faz o tratamento do

DNA com enzimas de restrição, as quais são proteínas que reconhecem uma seqüência de nucleotídeos específica (sítios de restrição) e a digerem, cortando o DNA em diferentes fragmentos. O número e tamanho dos fragmentos são estabelecidos pelo número de sítios de restrição reconhecidos pela enzima no DNA.

- ***Eletroforese:*** Aqui os fragmentos do DNA são separados com base em seus tamanhos. A eletroforese usa o principio que cada fragmento possui carga elétrica negativa devido ao grupo fosfato. O processo de eletroforese consiste em colocar a amostra de fragmentos de DNA sobre um gel feito de agarose ou poliacrilamida e aplicar um campo elétrico ao longo do gel. Cada fragmento se move na direção do lado positivo do campo. Este movimento produz uma série de *bandas*, as quais são aglomerações de fragmentos com similar tamanho. A posição das bandas é inversamente proporcional ao tamanho dos fragmentos. Isto significa que as bandas mais longe da origem estão compostas por fragmentos de menor tamanho. Normalmente em um mesmo gel são colocadas varias amostras de DNA de organismos distintos, produzindo uma série de canaletas, que são as distribuições de bandas dos organismos examinados numa determinada região do DNA cortado por alguma enzima de restrição. Ao final do processo de eletroforese é aplicado um corante para visualizar as bandas no gel. Estes corantes apresentam fluorescência quando excitados com luz ultravioleta (UV). O produto final é uma fotografia do gel sob radiação UV que serve para análises posteriores. A Figura 3.4 mostra uma foto exemplo resultante do processo de eletroforese.

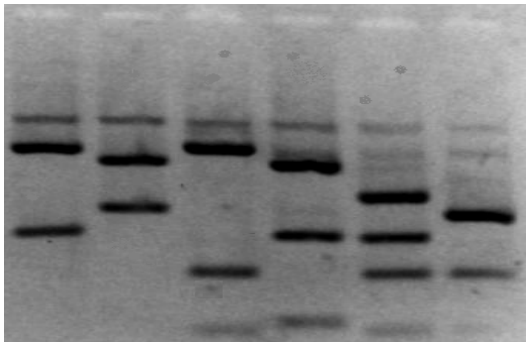


Figura 3.4: Exemplo de foto resultante do processo de eletroforese em gel, na qual foram analisadas 6 amostras produzindo 6 canaletas.

No banco de dados de estirpes de *Bradyrhizobium*, cada estirpe foi analisada em 3 regiões ribossomais: 16S, 23S e IGS. Para cada uma dessas regiões foram usadas três enzimas de restrição, portanto, cada estirpe é descrita por nove imagens de canaletas de eletroforese segundo a Tabela 3.3.

Tabela 3.3 - Relação de enzimas de restrição utilizadas e regiões ribossomais analisadas na obtenção do banco de dados de estirpes de *Bradyrhizobium*

Canaleta	Enzima de restrição	Região Ribossomal
1	Cfo I	16S
2	Dde I	16S
3	Msp I	16S
4	Hae III	23S
5	Hha I	23S
6	Hinf I	23S
7	Dde I	IGS
8	Hae III	IGS
9	Msp I	IGS

As imagens das canaletas, as quais serão chamadas simplesmente canaletas, foram extraídas das fotografias de gel mediante um processo manual usando um programa de edição de imagens. Estas imagens encontram-se em formato tiff em escala de cinza com 8 bits por pixel.

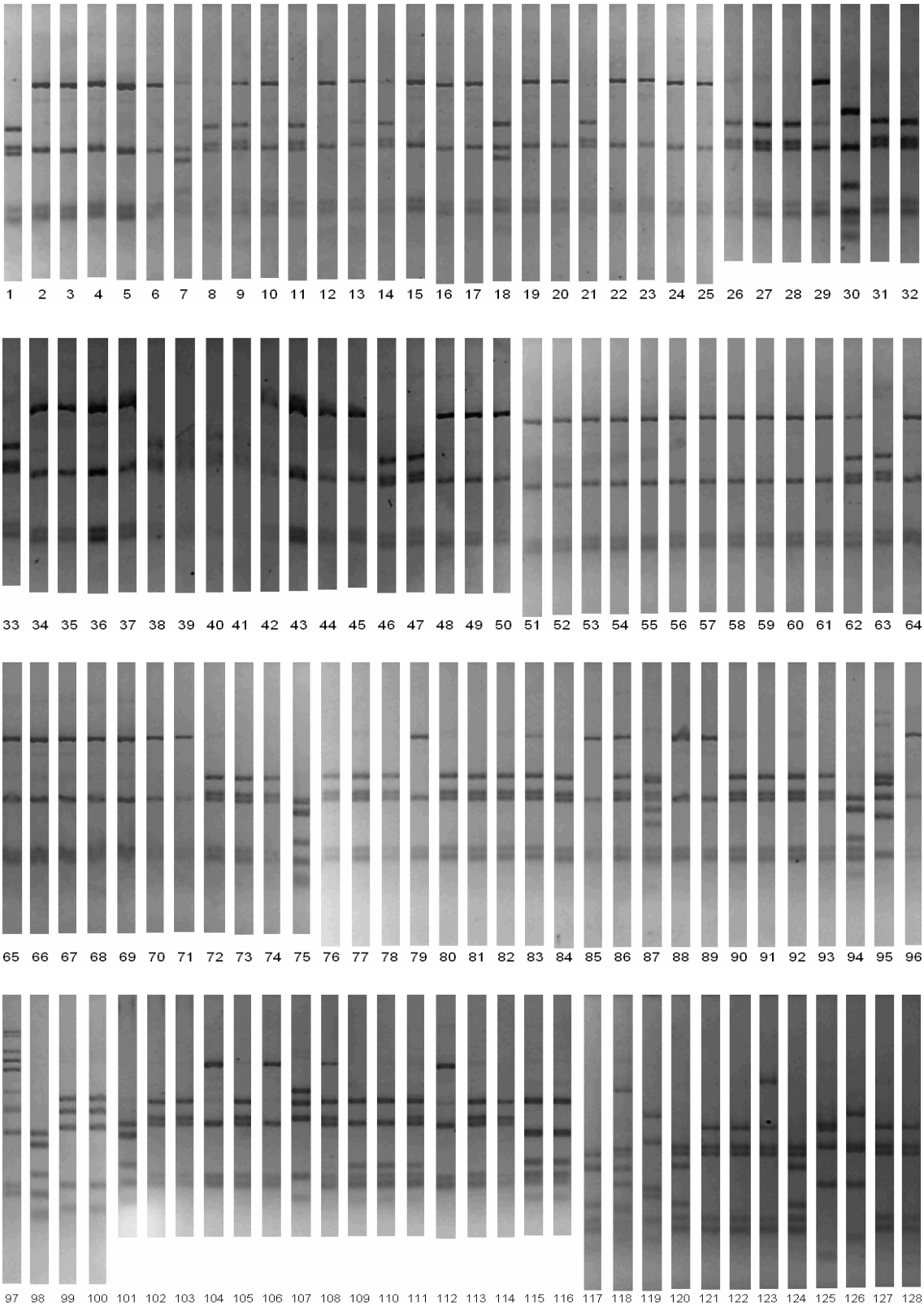


Figura 3.5: Subconjunto de canaletas do banco de estirpes de *Bradyrhizobium* usadas para a avaliação experimental dos métodos. Estas canaletas correspondem a todas as estirpes da coleção (128), analisadas na região ribossomal rRNA 16S e fragmentadas com a enzima de restrição Cfo I.

No presente trabalho foi usado somente um subconjunto de dados do banco de estirpes de *Bradyrhizobium* pois o foco principal era avaliar os métodos implementados e interpretar os resultados sem muita dificuldade. Este subconjunto é composto pelas canaletas de todas as estirpes correspondentes à região ribossomal 16S fragmentadas com a enzima de restrição Cfo I. A escolha da região ribossomal 16S foi porque ela tem sido a região recomendada para traçar filogenias e fazer taxonomias (WANG *et al*., 1999) visto que apresenta variabilidade suficiente para distinguir entre espécies e inferir progressões evolucionárias. Na Figura 3.5 são mostradas as canaletas compreendidas neste subconjunto de dados.

3.1.2.1. Pré-processamento das canaletas

Cada canaleta foi pré-processada com o objetivo de facilitar o processo de agrupamento e melhorar os resultados. A informação relevante das imagens de gel se encontra na distribuição das suas bandas, as quais são regiões da imagem onde a intensidade é notavelmente mais forte do que a intensidade do resto da imagem (intensidade de fundo). Estas bandas indicam a concentração de fragmentos de DNA de tamanho parecido. Muitas vezes o processo de identificação das bandas torna-se complicado devido à existência de uma série de influências físicas imersas no processo de eletroforese que causam perturbações na imagem resultante, entre elas se encontram:

- Presença de pequenas manchas na imagem devido a uma distribuição não uniforme do gel no processo de eletroforese;
- Deformações e falta de definição das bandas;
- Variabilidade da iluminação ao longo da imagem.

O objetivo do pré-processamento é minimizar estes inconvenientes e identificar as bandas de forma clara. O resultado do pré-processamento é o *eletroferograma*, o qual é uma seqüência ou sinal discreto que representa a intensidade dos pixels da canaleta ao longo do seu comprimento (posição), onde os picos desta seqüência indicam a presença de bandas.

Denotando R como a matriz de tons de cinza de uma determinada canaleta, formada por elementos (r_{ij}) que representam a intensidade do pixel na linha i e coluna j. As dimensões de R são denotadas com h e w, sendo h o número de linhas e w o número de colunas. O eletroferograma pode ser obtido de acordo com o seguinte procedimento:

1. Calcula-se a seqüência s correspondente à média das colunas de R:

$$s[i] = \frac{1}{w}\sum_{j} r_{ij} \tag{3.1}$$

2. Desloca-se a seqüência anterior para abaixo em uma quantidade th, isto é $t[i] = s[i] - th$, sendo th um limiar calculado empiricamente como $th = \mu + 0.5 \times \sigma$, na qual μ e σ são a média e o desvio padrão da e $\boldsymbol{s}$.

3. Estima-se a seqüência de eletroferograma $\mathbf{d}$ como:

$$d[i] = \begin{cases} \dfrac{t[i]}{Max\ (\mathbf{t})}, & t[i] > 0 \\ 0, & t[i] \le 0 \end{cases} \tag{3.2}$$

O passo 1 tem sentido desde que a informação relevante encontra-se no comprimento da imagem, este passo também reduz o efeito das pequenas manchas tornando-se insignificante na média da largura da imagem. O passo 2 desloca o sinal de tal forma que os picos (bandas) fiquem no lado positivo. Esta informação relevante é controlada por um limiar, o qual é dependente do primeiro e segundo

momento do sinal (μ e σ). Encontrou-se empiricamente que todo o que está acima de $\mu + 0.5 \times \sigma$ representa as bandas. O passo 3 encontra o eletroferograma considerando somente a parte positiva da seqüência previa, normalizando-a com respeito a seu máximo valor. A parte negativa do sinal é anulada devido a que é considerada como ruído de fundo da imagem. Na Figura 3.6 é mostrado um exemplo de pré-processamento para uma canaleta.

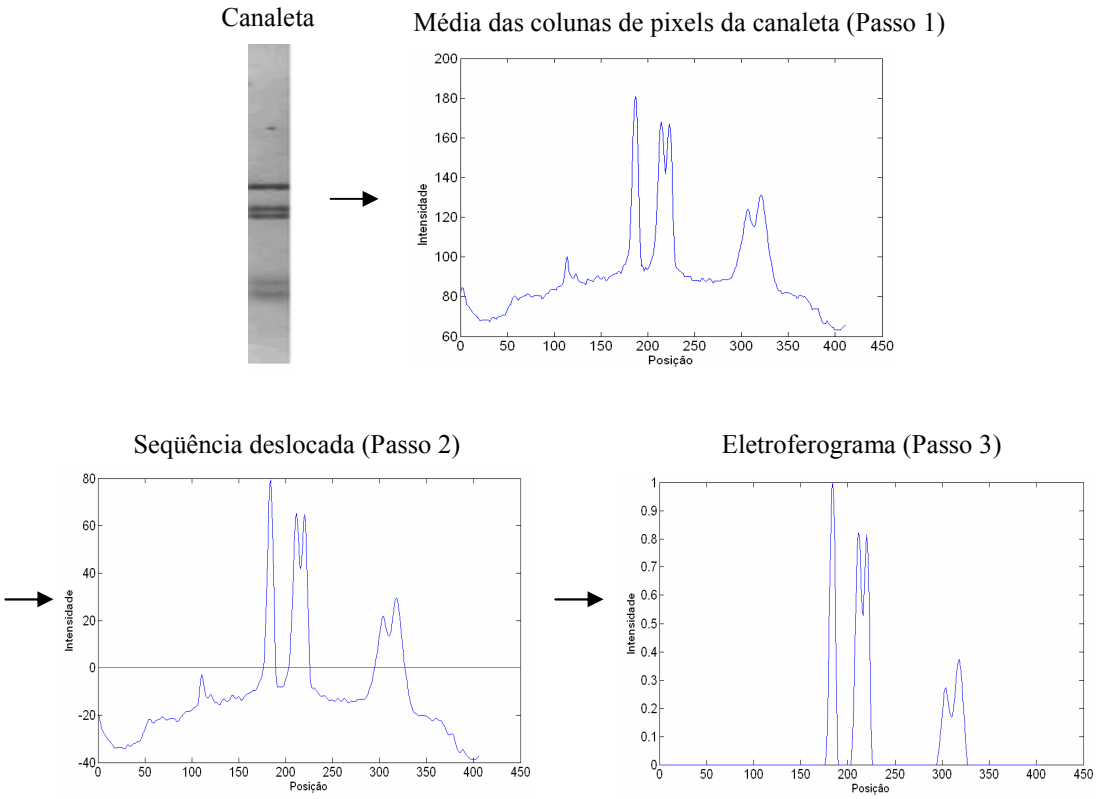


Figura 3.6: Exemplo de pré-processamento de uma imagem de canaleta de eletroforese de gel.

3.2. MÉTODO DE AGRUPAMENTO HIERÁRQUICO BAYESIANO

Na Seção 2.1.2 foi apresentado o algoritmo de agrupamento HBC. Este algoritmo constitui a base do primeiro método de agrupamento implementado neste trabalho, o qual é descrito nesta seção. O método basicamente segue os passos do

HBC mostrados no pseudocódigo da Figura 2.3. Lembrando brevemente, HBC começa formando uma partição inicial com grupos unitários, em cada interação são fundidos dois grupos, aqueles que apresentem a maior probabilidade de estarem fundidos. Foi mostrado que esses grupos são aqueles que maximizam o fator $\frac{SC(c_x \cup c_y)}{SC(c_x)SC(c_y)}$ (denotado como $U(c_x, c_y)$) sobre o espaço de pares de grupos existentes c_x e c_y. Para calcular este fator é necessário calcular as verossimilhanças dos grupos candidatos ($SC(c_x)$ e $SC(c_y)$) e do hipotético grupo fundido ($SC(c_x \cup c_y)$).

A verossimilhança de um grupo c_i é calculada como o produtório das probabilidades de pertinência de cada elemento do grupo: $p(d_j \mid c_i)$ (probabilidade do elemento d_j ser agrupado no grupo c_i). O cálculo desta probabilidade elementar é dependente da forma como são modelados os dados e os grupos, e constitui a diferencia principal do método implementado neste trabalho com respeito ao método HBC original, o qual foi especificado para agrupar dados de tipo documentos de texto.

A seguir é mostrada a forma como são modelados os dados e os grupos e como são calculadas as probabilidades de pertinência elementares para cada um dos bancos de dados analisados.

Cálculo das probabilidades de pertinência elementares nos bancos de dados de validação: Synthetic-2000, Synthetic-1000 e Íris.

Aqui é usado o procedimento apresentado em (VILLANUEVA; CHRIST; MACIEL, 2007; CHRIST; VILLANUEVA; MACIEL, 2007). A característica comum dos

bancos de validação é que eles são dados vetoriais, isto é, cada dado d_j é um vetor que representa os atributos do elemento j. Para calcular as probabilidades elementares de pertinência assume-se que os elementos dentro de cada grupo foram produzidos por uma distribuição normal multivariada (Gaussiana). A escolha de modelos gaussianos é justificada pela simplicidade na estimativa dos parâmetros e pelo sucesso em várias aplicações, inclusive com dados não necessariamente gaussianos (MURTAGH; RAFTERY, 1984; BANFIELD; RAFTERY, 1993; DASGUPTA; RAFTERY, 1998). Com esta consideração, a probabilidade de cada dado d_j de ser agrupado no grupo c_i é calculada como:

$$p(d_j \mid c_i) = N(d_j; \boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}) \tag{3.3}$$

onde $\boldsymbol{\mu}_{c_i}$ é o vetor média e $\boldsymbol{\Sigma}_{c_i}$ a matriz de covariância do grupo c_i. Estes parâmetros são computados como parâmetros de máxima verossimilhança:

$$\boldsymbol{\mu}_{c_i} = \frac{1}{|c_i|}\sum_{d_j \in c_i} d_j \ , \ \boldsymbol{\Sigma}_{c_i} = \frac{1}{|c_i|-1}\sum_{j=1}^{|c_i|}(d_j - \boldsymbol{\mu}_{c_i})(d_j - \boldsymbol{\mu}_{c_i})^T \tag{3.4}$$

Para estimar os parâmetros do modelo Gaussiano, cada grupo deve ter um número mínimo de elementos (maior que a dimensionalidade dos dados) para evitar matrizes de covariância singulares. Portanto, o processo de aglomeração não pode ser iniciado com grupos unitários. Assim, para iniciar a aglomeração é necessário fornecer uma partição inicial com grupos granulares (pequenos, mas não unitários) e homogêneos. A homogeneidade é exigida devido a que a qualidade do dendrograma resultante é influenciada pela homogeneidade dos grupos iniciais, uma vez que o método somente funde grupos. Para criar as partições iniciais sugere-se o uso de algoritmos de agrupamentos como K-Means (DUDA *et al.*, 2001), ou ART2 (CARPENTER; GROSSBERG, 1987) devido a seu conhecido desempenho, sua

simplicidade e por permitir o controle da homogeneidade dos grupos mediante o ajuste do parâmetro de vigilância ρ na ART2 ou o número de grupos em K-Means.

Cálculo das probabilidades de pertinência elementares no banco de estirpes de Bradyrhizobium.

Contrariamente aos bancos de dados de validação, o banco de estirpes de Bradyrhizobium não se encontra na forma de vetores de características ou coordenadas, razão pela qual, o cálculo das probabilidades de pertinência elementares é diferente do descrito para os bancos de dados de validação. Os eletroferogramas resultantes do pré-processamento das canaletas (Seção 3.1.2.1) constituem os dados de entrada a serem agrupados mediante o presente método.

Considere-se a canaleta j com eletroferograma d_j, para calcular a sua probabilidade de pertencer ao grupo c_i, isto é $p(d_j \mid c_i)$, é assumido que a canaleta é resultado de um processo aleatório, onde a variável aleatória é a posição P na canaleta em que é observado material genético (bandas). A probabilidade de se encontrar material genético numa posição particular k da canaleta dado seu eletroferograma, $p(P = k \mid d_j)$, pode ser calculada como a fração de material encontrado nessa posição, indicado pelo k-ésimo valor do eletroferograma ($d_j[k]$), com respeito ao total de material representado no eletroferograma, assim :

$$p(P = k \mid d_j) = \frac{d_j[k]}{\sum_i d_j[i]} \tag{3.5}$$

O conjunto de probabilidades obtidas por (3.5) para todas as posições k é a distribuição de probabilidades de P, a qual é considerada como o modelo probabilístico do dado. O cálculo desta distribuição de probabilidades pode ser

interpretado como uma normalização do eletroferograma com respeito a sua área, a qual representa a quantidade total de material genético na canaleta.

O modelo probabilístico de um grupo c_i é determinado como a média dos modelos probabilísticos dos seus dados integrantes, calculado como:

$$p(P=k \mid c_i) = \frac{\sum_{d_j \in c_i} d_j[k]}{\sum_{d_j \in c_i} \sum_{i} d_j[i]} \tag{3.6}$$

Para calcular a probabilidade elementar $p(d_j \mid c_i)$, considera-se todos os eventos $P=k$ (encontrar material genético na posição k). Condicionando sobre estes eventos e aplicando a lei de probabilidade total, tem-se:

$$p(d_j \mid c_i) = \sum_k p(d_j \mid c_i, P=k)\, p(P=k \mid c_i) \tag{3.7}$$

considerando que existe independência condicional entre os dados e o grupo dado o evento $P=k$ e aplicando o teorema de Bayes, tem-se:

$$p(d_j \mid c_i) = p(d_j) \sum_k \frac{p(P=k \mid d_j)\, p(P=k \mid c_i)}{p(P=k)} \tag{3.8}$$

desta última equação, o termo $p(d_j)$ é a probabilidade marginal do dado e é desconsiderada no cálculo, já que para propósitos de maximização ela é sempre constante para qualquer partição. Os outros termos correspondem aos modelos probabilísticos do dado d_j (3.5), do grupo c_i (3.6) e do banco de dados inteiro, a qual é calculada com (3.6) fazendo $c_i = D$.

Ao invés dos bancos de dados de validação, o processo de aglomeração no banco de estirpes de Bradyrhizobium é iniciado desde uma partição inicial com grupos unitários. Isto devido à forma como se encontram representados os dados

(seqüências), podendo-se formular um modelo probabilístico para cada dado ou grupo unitário. O algoritmo HBC junta em cada iteração os dois grupos c_x, c_y com maior valor do fator $U(c_x, c_y)$, a distância representada no dendrograma (altura) em que são unidos esses dois grupos é $\log(U(c_x, c_y))$. O logaritmo é usado para uma melhor visualização do dendograma acentuando mais as deferências entre pequenas e grandes distâncias, facilitando assim a identificação de grupos. O método de agrupamento Hierárquico Bayesiano - HBC foi implementado no *software* MATLAB®[1]

3.3. MÉTODO DE AGRUPAMENTO DE DADOS BASEADO EM REDES GAUSSIANAS CONDICIONAIS

Como foi visto no Capítulo 2, as CGN são uma classe de MGP que pode lidar com dados discretos e contínuos. A adequação destes modelos para classificação não supervisionada é feita introduzindo uma variável aleatória discreta chamada variável classe, sendo as variáveis preditivas dependentes da variável classe. Para poder aprender e fazer inferência com este tipo de modelo é necessário que os dados, também chamados evidências, se encontrem na forma de vetores de características ou coordenadas. Como foi visto anteriormente, os bancos de dados de validação já se encontram no formato tabular requerido, entretanto, o banco de dados de estirpes de *Bradyrhizobium* é dado na forma de seqüências discretas (eletroferogramas), os quais claramente, não são da forma requerida pelas CGN. Conseqüentemente, para fazer análise de agrupamentos do banco de estirpes de *Bradyrhizobium* com o método baseado em modelos CGN, é necessário primeiro

[1] http://www.mathworks.com/

transformar os eletroferogramas disponíveis na forma de vetores de coordenadas que os modelos CGNs possam entender.

3.3.1. TRANSFORMAÇÃO DOS ELETROFEROGRAMAS EM VETORES DE COORDENADAS

Os eletroferogramas obtidos na etapa de pré-processamento são transformados em vetores de coordenadas mediante a técnica de escalamento multidimensional - MDS (*Multidimensional Scaling*) (KRUSKAL; WISH, 1978). Esta técnica encontra uma representação na forma de vetores de coordenadas a partir de uma matriz de distância calculada entre os dados. Basicamente a técnica realiza sucessivas manipulações da matriz de distância com o objetivo de encontrar um conjunto de pontos em um espaço euclidiano cuja matriz de distância calculada neles seja próxima da matriz de distância original.

Para calcular a matriz de distância entre os eletroferogramas foi escolhido o coeficiente de correlação de Pearson (OOYEN, 2001), o qual mede a associação linear entre duas seqüências sem depender da unidade de medida. Quanto maior o valor do coeficiente, maior a similaridade entre as seqüências. Para dois eletroferogramas d_i e d_j, o coeficiente de correlação é definido:

$$\delta_{ij} = \frac{1}{T}\sum_k \left(\frac{d_i[k]-\overline{d_i}}{\sigma_{d_i}}\right)\left(\frac{d_j[k]-\overline{d_j}}{\sigma_{d_j}}\right). \qquad (3.9)$$

onde T é o tamanho dos eletroferogramas, $\overline{d}$ e σ são a média e o desvio padrão dos eletroferogramas respectivamente.

Os elementos da matriz de distância R são formados a partir dos coeficientes de correlação calculados entre todos os pares de eletroferogramas, isto é $R=(r_{ij})$, onde $r_{ij}=1-\delta_{ij}$.

A implementação de MDS usada neste trabalho foi a que se encontra disponível no *toolbox* estatístico de MATLAB® mediante a função **cmdscale,** a qual devolve uma matriz de vetores cujas coordenadas estão ordenadas em forma decrescente de acordo a sua importância, ou seja, as primeiras coordenadas representam a maior variância dos dados.

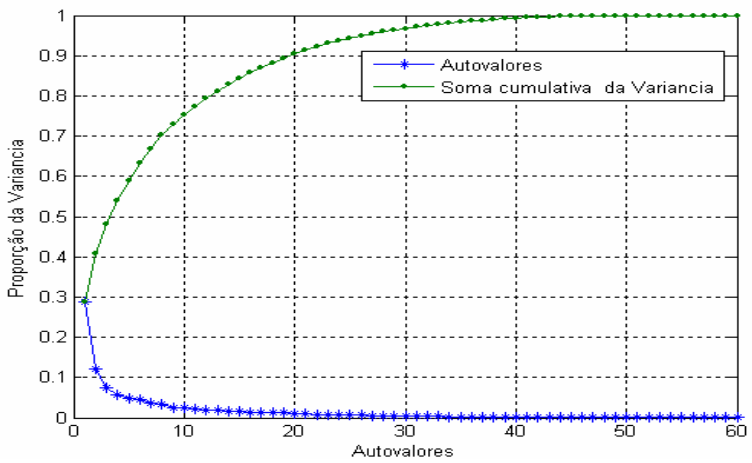


Figura 3.7: Análise de autovalores da matriz de vetores resultantes de MDS.

Para selecionar o número adequado de coordenadas a serem consideradas no agrupamento, foi realizada uma análise de autovalores da matriz de vetores resultantes. A Figura 3.7 mostra a fração da variância que representa cada autovalor, onde os primeiros autovalores correspondem às primeiras coordenadas. Com este análise foram escolhidas as 10 primeiras coordenadas, as quais representam acima do 75 % da variância dos dados.

3.3.2. AGRUPAMENTO COM NAIVE-BAYES

Para realizar o agrupamento foi escolhido o modelo Naive-Bayes (NB) estudado no Capítulo 2. Embora as considerações feitas neste tipo de modelo são irrealistas, este tem apresentado bons resultados em muitos domínios de aplicação

(MICHIE *et al*.,1994) apesar de sua simplicidade. Essa foi a principal razão para sua utilização no presente trabalho.

O número de nós filhos da variável classe é igual ao número de atributos ou coordenadas dos bancos de dados. No caso dos bancos de validação foram usados todos os atributos para construir os modelos. Já no banco de estirpes de *Bradyrhizobium* e com base no resultado da análise de autovalores, foram usadas somente as 10 primeiras coordenadas dos vetores resultantes da transformação MDS dos eletroferogramas. O motivo pelo qual não são usadas todas as coordenadas é porque o modelo resultaria excessivamente complexo sem um ganho importante nos resultados. Na Figura 3.8 é mostrada a estrutura do modelo para o banco de dados Íris. Os modelos para os outros bancos de dados são parecidos, variando unicamente o número de nós filhos. Em todos os bancos de dados, as variáveis preditivas são de natureza contínua e a variável classe é de natureza discreta, ração pela qual é justificado o uso de modelos CGN.

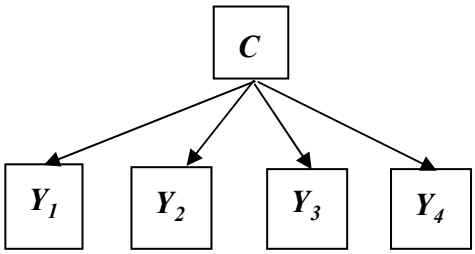


Figura 3.8: Estrutura de modelo NB para fazer análise de agrupamentos do banco de dados Iris.

Para a implementação dos modelos NB foi usado o *software Bayes Net Toolbox - BNT,* um *toolbox* de código aberto escrito para MATLAB® por Kevin Murphy. Para uma boa documentação sobre este *software* consultar (MURPHY, 2001). Este software tem mostrado uma ampla aceitação na comunidade acadêmica

e foi base de novas implementações como o *Probabilistic Networks Library - PNL*[1] de Intel (ERUHIMOV *et. al.*,2003).

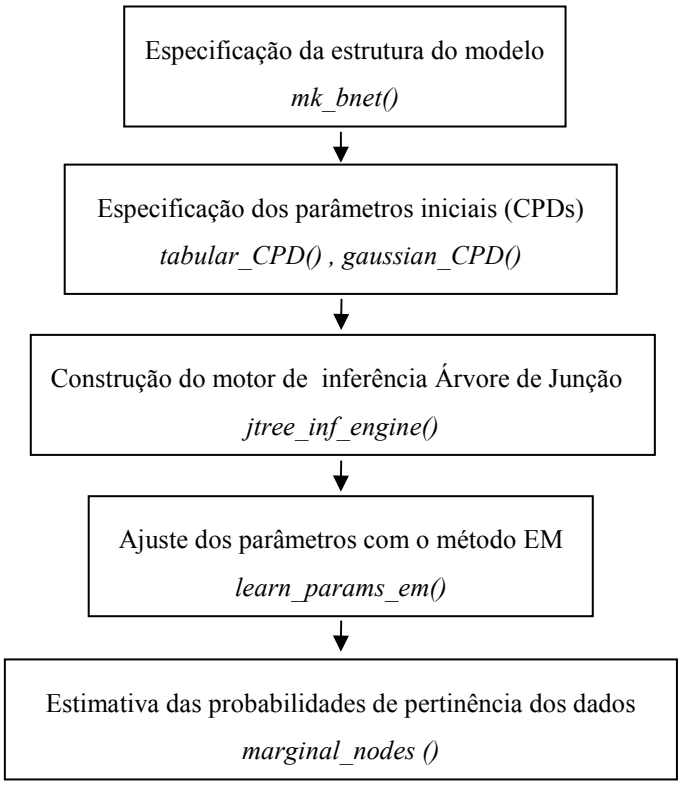


Figura 3.9: Etapas do método de agrupamento usando modelos NB. As funções indicadas estão implementadas em BNT.

Na Figura 3.9 é mostrado um diagrama de blocos indicando as etapas do método de agrupamento com modelos NB. São indicadas também as funções principais de cada etapa, as quais se encontram implementadas no *toolbox* BNT. A seguir é descrita cada uma das etapas integrantes do método.

Especificação da estrutura do modelo: Aqui é especificada a estrutura do modelo da CGN. Esta estrutura é fornecida na forma de matriz de adjacência, a qual codifica o grafo da estrutura. Adicionalmente é necessário especificar o tamanho e tipo de cada nó. Para o nó discreto da classe, o tamanho representa a quantidade de

[1] http://www.intel.com/technology/computing/pnl/

estados possíveis que pode tomar, ou seja, o número de grupos. Para nós contínuos, o tamanho representa o número de variáveis agrupadas no nó. No caso dos modelos NB implementados neste trabalho, todos os nós contínuos representam unicamente um atributo ou variável preditiva, portanto seu tamanho sempre é 1. Nos experimentos, o tamanho da variável classe é variado sistematicamente para descobrir o número ótimo de grupos. Na Figura 3.10 é mostrada a matriz de adjacência do modelo NB para o banco de dados Iris, onde o valor do elemento da linha i e coluna j indica se existe um arco entre os nós i e j. As matrizes de adjacência para os outros modelos NB são construídas da mesma maneira.

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figura 3.10: Matriz de adjacência do modelo NB para o banco de dados Iris.

Especificação dos parâmetros iniciais (CPDs): No *toolbox* BNT, os parâmetros são representados por objetos CPDs (*Conditional Probability Distribution*), os quais definem a distribuição de probabilidade de um nó, dados seus pais. Em nós discretos, o CPD é uma tabela armazenada em forma de array multidimensional. Se um nó discreto não possui pais, o CPD é um vetor que armazena suas probabilidades *a priori*. É possível também especificar a distribuição *a priori* dos parâmetros de nós discretos, os tipos suportados são *dirichlet*[1] e *entropic*. Em nós contínuos (Gaussianos) com pais discretos (como o NB), os parâmetros são a média e a matriz de covariância por cada valor das variáveis discretas, portanto o CPD fica

[1] A distribuição de Dirichlet é a família conjugada da distribuição multinomial, a qual assume-se seguem as variáveis discretas.

definido com o vetor de médias e as matrizes de covariância armazenadas em uma *array* multidimensional. BNT não suporta distribuições *a priori* para os parâmetros de nós contínuos.

Especificamente nos modelos NB do presente trabalho, o nó classe C não tem pais, portanto seu CPD é um vetor que indica a distribuição das probabilidades *a priori* de cada classe. O valor destas probabilidades é inicialmente assumido com valores uniformes e seguindo uma distribuição paramétrica *a priori* de tipo *Dirichlet*. À medida que o modelo vai aprendendo dos dados (etapa de aprendizado) estas probabilidades vão fugindo da distribuição uniforme para uma distribuição que representa melhor os dados. A força com que a distribuição inicial uniforme dos parâmetros prevalece na etapa de aprendizado é controlada pelo parâmetro TAE (*tamanho de amostra equivalente)* da distribuição de *Dirichlet*. Por exemplo um valor TAE = 10 significa que o conhecimento *a priori* que as classes são uniformemente distribuídas é baseado na observação equivalente de 10 amostras. Nos experimentos o valor de TAE é variado de acordo com o experimento executado (seção seguinte).

No caso das variáveis contínuas (variáveis preditivas), o vetor de médias tem o mesmo tamanho que o nó classe (nó pai) o qual é dependente do teste executado. Igualmente cada matriz de covariância é em realidade um escalar desde que todas as variáveis preditivas são unidimensionais, portanto estes escalares representam as variâncias de cada variável para cada estado da variável classe. Tanto o vetor de médias e as variâncias são iniciados em valores aleatórios para logo serem ajustados com base aos dados na etapa de aprendizado.

Construção do motor de inferência árvore de junção: A construção do motor de inferência é realizada aqui devido a que na seguinte etapa (aprendizado dos

parâmetros) este é requerido para completar os valores da variável classe. O pacote BNT tem implementado uma variedade de algoritmos para realizar inferência, entre esses se destaca o algoritmo árvore de junção, o qual foi escolhido neste trabalho devido a sua eficiência computacional quando comparado com outros algoritmos exatos (PASKIN, 2003). O algoritmo implementado em BNT segue os mesmos passos descritos na Seção 2.2.4. Na Figura 3.11 é mostrada a árvore de junção resultante para o modelo NB do banco de dados Íris, na qual podem ser observados 4 *cliques* $\{V_1,...,V_4\}$ e 3 separadores $\{S_1,...,S_3\}$. Similarmente, as árvores de junção para os outros modelos NB foram construídas produzindo estruturas similares à mostrada na Figura 3.11, com a única variação do número de *cliques* e separadores conectados ao clique V_2. A quantidade destes *cliques* e separadores é função do número de variáveis preditivas que descrevem o banco de dados. Todos os *cliques* contêm a variável classe e uma variável preditiva, entretanto, os separadores contêm unicamente a variável classe. O *clique* raiz V_1 é aquele que contém a última variável preditiva.

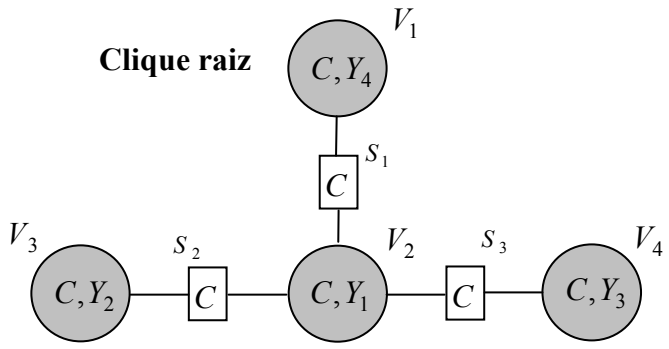


Figura 3.11: Árvore de junção para o modelo NB do banco de dados Íris.

Após da construção da árvore de junção são definidos os CG-potenciais em cada *clique* e cada separador, para isso cada nó é atribuído a um *clique* onde ele e

seus pais são membros (Seção 2.2.4). Os CG-potenciais dos *cliques* são definidos como os produtos dos CG-potenciais dos seus nós atribuídos. O CG-potencial de cada nó é calculado com os parâmetros do seu CPD previamente iniciado (passo anterior). Os CG-potenciais dos separadores são iniciados com valor 1 devido a que não possuem nós atribuídos. A Tabela 3.4 mostra a atribuição de nós em *cliques* para o modelo NB do banco de dados Íris. A atribuição de nós em *cliques* nos outros modelos NB é similar, variando somente em número de nós e *cliques*.

Tabela 3.4 - Atribuição de nós em cliques na árvore de junção do modelo NB para o banco de dados Íris.

Nó	C	Y_1	Y_2	Y_3	Y_4
Clique	V_1	V_1	V_2	V_3	V_4

Ajuste dos parâmetros com o método EM: Neste passo são atualizados os parâmetros em base aos dados disponíveis ou evidências. O aprendizado é feito usando a implementação do algoritmo EM disponível no BNT. Este algoritmo foi descrito no Capítulo 2 o qual usa a árvore de junção para estimar o estado da variável oculta classe em cada dado. Na etapa E (*Expectation*) são introduzidos na árvore de junção os valores das variáveis preditivas (evidência) de cada dado para assim calcular as probabilidades marginais sobre a variável classe (probabilidades de pertinências às classes) e com essas probabilidades calcular as estatísticas suficientes (Seção 2.2.5). Na etapa M (*Maximization*) são calculadas as estimativas de parâmetros de máxima verossimilhança dadas por (2.33), (2.36) e (2.37). Os coeficientes de regressão linear não são calculados devido a que no modelo NB não existe relacionamento entre variáveis preditivas. O processo é repetido até atingir um número máximo de iterações ou até que o logaritmo da verossimilhança entre duas iterações sucessivas seja menor que um limiar. Nos experimentos, o número

máximo de iterações foi colocado em 100 e o limiar de verossimilhança em 10^{-20}.

Estimativa das probabilidades de pertinência dos dados: Uma vez que os parâmetros foram ajustados, o modelo foi usado para estimar as probabilidades de pertinência dos dados com respeito a cada grupo, representado pela variável classe. Ao igual que o passo de ajuste de parâmetros, as probabilidades de pertinência são obtidas calculando as probabilidades marginais sobre o nó classe, após de introduzir os valores dos atributos do dado nas respectivas variáveis preditivas da árvore de junção. Para poder comparar os resultados do modelo NB com os resultados de outros métodos de agrupamento, os dados são atribuídos ao grupo com a maior probabilidade de pertinência, gerando assim uma partição ou agrupamento. É importante ressaltar que a informação dos graus de pertinência com que um dado está sendo ligado aos diversos grupos é perdida no momento de atribuir o dado ao grupo de maior probabilidade. Portanto as probabilidades de pertinência constituem uma informação mais completa que a simples partição dos dados, sendo esta uma das principais vantagens dos modelos probabilísticos.

3.4. EXPERIMENTOS

Os experimentos foram divididos em duas partes, primeiramente foram planejados experimentos sobre os bancos de dados de validação com o objetivo de avaliar os métodos implementados, realizando uma comparação do desempenho dos mesmos com respeito a outros métodos de agrupamentos conhecidos. Após esta validação inicial, os métodos implementados foram utilizados para realizar análise de agrupamentos sobre o banco de dados de estirpes de *Bradyrhizobium* com o objetivo de encontrar uma taxonomia das estirpes. A seguir são detalhados os dois tipos de experimentos realizados.

Experimentos nos bancos de dados de validação

Como foi descrito anteriormente, nestes bancos de dados é conhecido o rótulo da classe de cada dado. Esta informação é usada aqui com a finalidade de testar a qualidade dos agrupamentos realizados pelos métodos implementados. Considerando que o banco de dados tem m classes verdadeiras e que pode ser expressado na forma de pares (d_j, l_j), onde l_j indica a etiqueta verdadeira do dado d_j que pode tomar valores $\{cl_1, cl_2, ..., cl_m\}$, a qualidade de um agrupamento ou partição $C = \{c_1, c_2, ...c_k\}$ com k grupos é medida com o *índice de pureza*, o qual é definido como:

$$pureza = \frac{\sum_{c_i \in C} \max(N_{c_i}(cl_1), N_{c_i}(cl_2), ..., N_{c_i}(cl_m))}{n} \qquad (3.10)$$

na qual, $N_{c_i}(cl_j)$ denota o número de elementos com etiqueta cl_j dentro do grupo c_i. O índice de pureza indica a percentagem de elementos de uma partição que possuem a etiqueta majoritária em cada grupo.

No caso do método de agrupamento hierárquico Bayesiano, foram geradas 2 partições iniciais para cada banco de dados com os métodos K-Means e ART2. Ambas partições iniciais consistiram de 40 grupos para o banco de dados Synthetic-2000, 30 grupos para o banco de dados Synthetic-1000 e 6 grupos homogêneos para o banco de dados Íris. O método hierárquico foi avaliado em 3 aspectos:

- ***Qualidade do dendrograma gerado***, isto foi realizado observando a *pureza média*, ou seja, a média das purezas calculadas em cada partição do dendrograma. Foram também computadas e comparadas as purezas de dendrogramas gerados por outros métodos hierárquicos não probabilísticos como o: *Single Linkage, Complete Linkage e Average Linkage.*

- ***Número de grupos ótimos***, aqui se usou o critério BIC (Bayesian Information Criterion) (SCHWARZ,1978). Este critério é usado para avaliar a expressividade de cada partição do dendrograma e assim escolher o agrupamento ótimo dos dados (a partição com o maior BIC). O BIC para uma partição C_k é definido como:

$$BIC_{C_k} = 2L_{C_k} - m_{C_k} \log(n) \tag{3.11}$$

  na qual, L_{C_k} é o logaritmo da verossimilhança da partição C_k, m_{C_k} é o número de parâmetros independentes da partição e n é o número de dados.
- ***Qualidade da partição ótima***, aqui a partição ótima foi comparada contra partições geradas por outros algoritmos particionais como o K-Means e o EM. A comparação se realizou usando o índice de pureza definido anteriormente.

No caso do método de agrupamento baseado em redes gaussianas condicionais, a avaliação consistiu em 2 aspectos:

- ***Aproveitamento do conhecimento a priori***. Isto foi realizado com o objetivo de determinar quanto é melhorado o agrupamento dos dados com este tipo de modelo à medida que é introduzida a informação a priori disponível. Nos bancos de dados de validação é conhecido que os elementos estão uniformemente distribuídos nas classes existentes. Isto se traduz em uma distribuição inicial da variável classe de tipo uniforme. A força com que este conhecimento a priori prevalece quando os dados são introduzidos no modelo é controlado pelo parâmetro TAE. Quanto maior TAE, maior é a confiança no conhecimento a priori. O banco de dados Synthetic-2000 foi testado com valores de TAE={0,1,100,2000,20000}, o banco de dados Synthetic-1000 foi testado com valores de TAE={0,1,100,1000,10000} e o banco de dados Íris foi testado com valores de TAE={0,1,10,100,1000,10000}. A avaliação aqui consistiu na

observação dos índices de pureza para os distintos valores de TAE.

- ***Determinação do particionamento ótimo***. Isto se realizou com dois critérios estatísticos: o BIC definido em (3.11) e o AIC (Akaike Information Criterion) (AKAIKE,1974), o qual é definido para um agrupamento C_k como:

$$AIC_{C_k} = 2L_{C_k} - 2m_{C_k} \tag{3.12}$$

em que L_{C_k} é o logaritmo da verossimilhança da partição C_k e m_{C_k} é o número de parâmetros independentes da partição. O número de grupos (número de estados que pode tomar a variável classe) foi variado nos diferentes testes tomando valores entre 2 e 7 para os bancos Synthetic-2000 e Synthetic-1000 e entre 2 e 6 para o banco de dados Íris.

Cada experimento consistiu em estabelecer um valor para o TAE e para o número de grupos seguido pela execução de 10 testes, isto é, a execução de todas as etapas do método (Figura 3.9). O resultado de cada experimento é o agrupamento realizado pelo melhor teste, ou seja, o teste que corresponde ao modelo aprendido com a maior verossimilhança. A justificativa de realizar vários testes por experimento se baseia na tentativa de aprender modelos com parâmetros ótimos globais, o qual não é garantido no método de aprendizado EM. Os experimentos foram realizados de tal forma que o TAE só muda de valor uma vez realizados os experimentos para todos os valores do número de grupos.

Experimentos no banco de estirpes de *Bradyrhizobium*

Neste banco de dados não se dispõe da informação da classe, razão pela qual não é possível avaliar a pureza dos agrupamentos gerados. Os experimentos realizados aqui têm por finalidade encontrar agrupamentos de bactérias considerando que os métodos terão um desempenho similar ao apresentado nos

bancos de dados de validação.

No caso do método hierárquico Bayesiano, foram utilizados os critérios BIC e AIC descritos anteriormente para encontrar o particionamento ótimo ou mais "natural" do dendrograma gerado.

No método de agrupamento baseado no modelo NB, foram executados experimentos com o objetivo de encontrar a partição ótima dos dados e testar a estabilidade dos agrupamentos variando a força do conhecimento *a priori*. Este conhecimento *a priori* corresponde à hipótese de que as classes são uniformemente distribuídas. Embora esta hipótese pode não ser certa, é usada aqui para avaliar a estabilidade dos agrupamentos baixo diversos níveis de confiança (TAE) neste simulado conhecimento *a priori*. A estratégia de testes foi similar à seguida nos bancos de dados de validação, com o TAE tomando valores {0,1,10,100,1000,10000} e o número de grupos variando entre 2 e 10. Igualmente, por cada valor de TAE foram realizados experimentos para todos os valores do número de grupos, sendo o resultado de cada experimento o melhor de 10 testes realizados. Os critérios BIC e AIC foram usados similarmente para determinar o número ótimo de grupos.

Capítulo 4

RESULTADOS E DISCUSSÕES

Este capítulo está estruturado em duas partes. Na primeira parte são apresentados e discutidos os resultados dos experimentos realizados nos bancos de dados de validação. Na segunda parte são apresentados e discutidos os resultados da análise de agrupamentos realizado no banco de estirpes de *Bradyrhizobium* com os métodos implementados.

4.1. RESULTADOS NOS BANCOS DE DADOS DE VALIDAÇÃO

4.1.1. SYNTHETIC-2000

Resultados do método hierárquico Bayesiano.

Com este método de agrupamento foram criados dois dendrogramas: o primeiro, a partir de uma partição inicial de 40 grupos gerado pelo algoritmo K-Means com 99.8% de pureza (3.10), o qual foi chamado HBC-Kmeans, e o segundo, com uma partição inicial também de 40 grupos gerada pelo algoritmo ART2 com 98.5% de pureza, chamado HBC-ART2. Na Figura 4.1 é mostrada a evolução da pureza nas 40 partições de ambos dendrogramas junto com a pureza das 40 últimas

partições de dendrogramas gerados pelos algoritmos *Single Linkage*, *Complete Linkage* e *Average Linkage.* Uma comparação das purezas médias dos dendrogramas é mostrada na Figura 4.2, onde é observado que quase todos os métodos apresentaram purezas similares exceto o método *Single Linkage* que apresentou uma pureza média significativamente inferior. Estes resultados são concordantes com o fato de que os grupos em Synthetic-2000 estão bem separados.

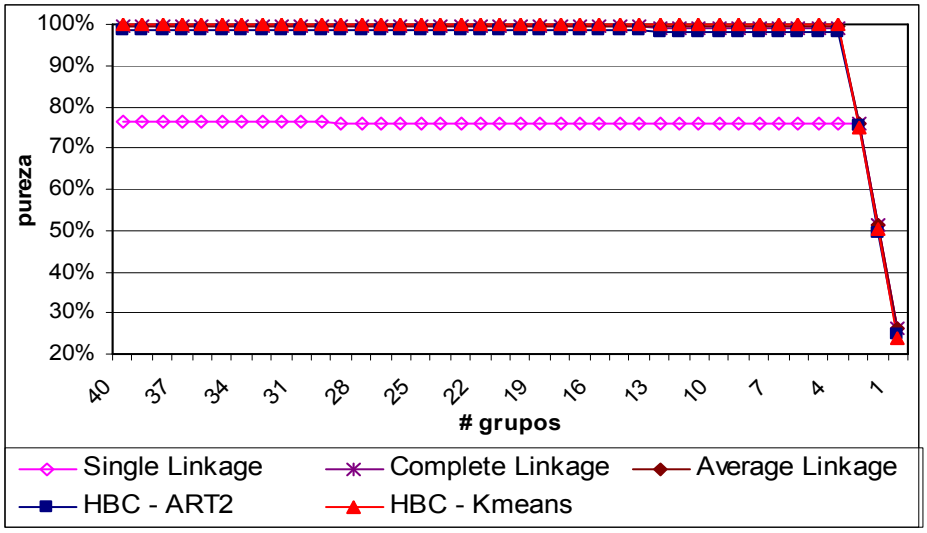


Figura 4.1: Evolução da pureza dos dendrogramas gerados em Synthetic-2000 com patições iniciais K-means e ART2 e comparados com dendrogramas gerados pelos métodos *Single Linkage*, *Complete Linkage* e *Average Linkage*

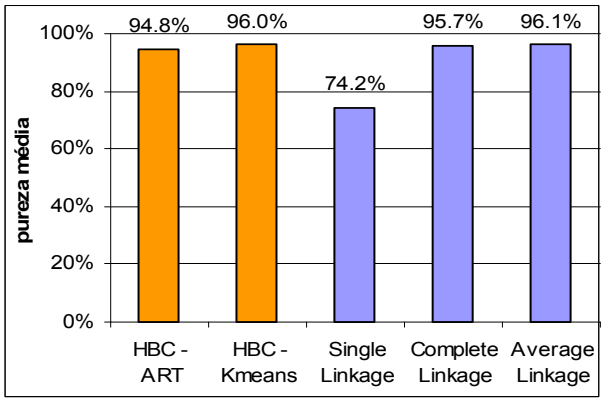


Figura 4.2: Pureza média de dendrogramas gerados em Synthetic-2000 pelo método hierárquico Bayesiano e os métodos hierárquicos tradicionais.

Observando a Figura 4.2 foi encontrado que a pureza do dendrograma gerado com a partição inicial ART2 é levemente inferior do que o gerado com a partição inicial K-Means. Para encontrar o número ótimo de grupos foi escolhido este último dendrograma e calculado o critério BIC em cada partição. A Figura 4.3 mostra a curva do BIC resultante onde pode ser observado que a partição ótima corresponde ao agrupamento de 4 grupos.

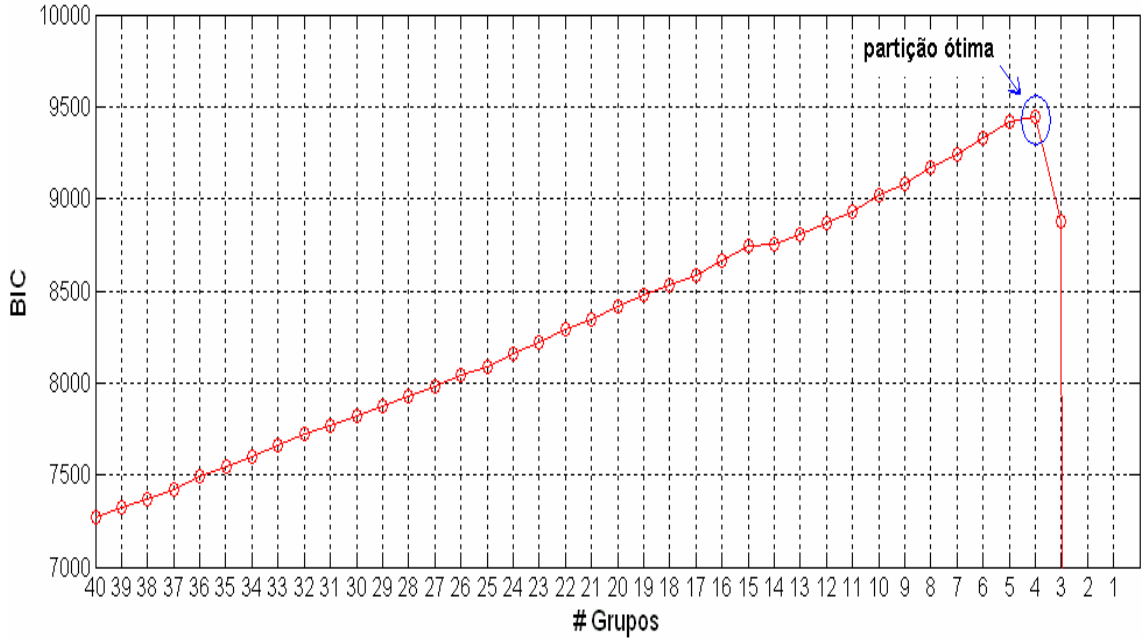


Figura 4.3: Evolução do critério BIC no processo de aglomeração do dendrograma HBC-Kmeans no banco de dados Synthetic-2000.

Na Figura 4.4 é comparada a pureza da partição ótima obtida com o critério BIC contra outras partições de 4 grupos realizados pelos algoritmos particionais K-Means e EM. Aqui é observado que todos os métodos geraram agrupamentos com purezas similares e muito próximas de 100%, sendo este resultando concordante com o esperado, pois o banco de dados Synthetic-2000 tem grupos claramente separados.

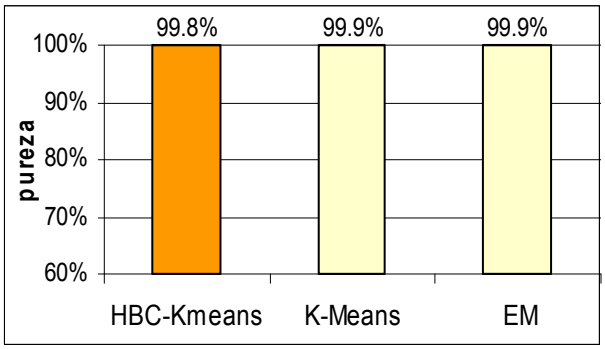


Figura 4.4: Pureza da partição ótima do método hierárquico Bayesiano contra partições de 4 grupos gerados por outros métodos particionais no banco de dados Synthetic-2000.

Resultados do método baseado no modelo NB.

Na Figura 4.5 é mostrada a pureza dos agrupamentos realizados pelo modelo NB (atribuindo os dados ao grupo com maior probabilidade de pertinência). É observado que as curvas de pureza para os diferentes valores de TAE estão sobrepostas, o que significa que a pureza dos agrupamentos é independente do valor de TAE, ou também, que o conhecimento *a priori* (que as classes são uniformes) não tem influência na formação dos grupos. Isto pode ser devido a que as classes estão bem separadas e que inclusive com um pobre conhecimento *a priori* das classes se podem obter agrupamentos comparáveis com os obtidos com um forte conhecimento *a priori*.

Nas Figuras 4.6 e 4.7 são mostradas as curvas dos critérios AIC e BIC respectivamente para diferentes valores de TAE como função do número de grupos. É observado que nos dois critérios e em todas as curvas de TAE o número ótimo de grupos é 4, o qual é o número certo de grupos. Este resultado é ainda mais claro para valores altos de TAE, o que significa que o conhecimento *a priori* disponível ajuda na determinação da partição ótima.

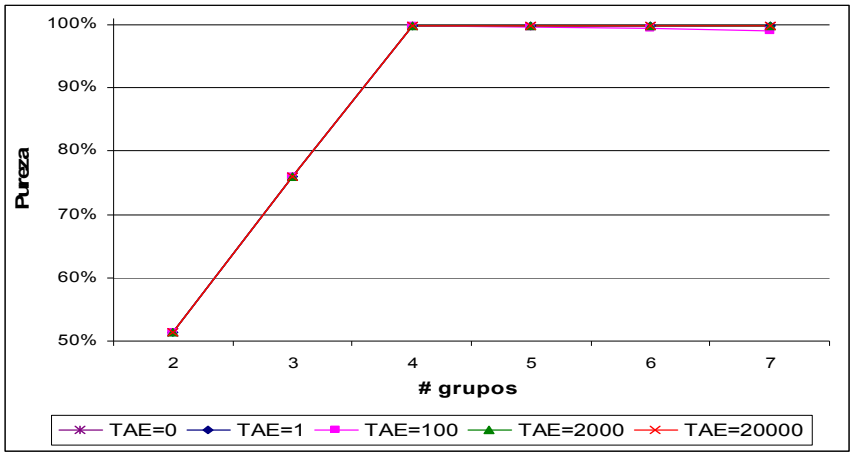


Figura 4.5: Pureza das partições feitas pelo modelo NB em Synthetic-2000.

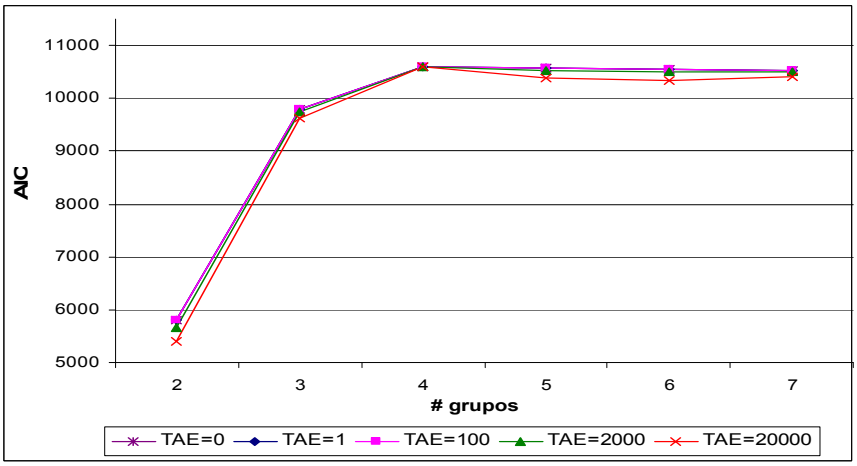


Figura 4.6: Critério AIC para diferentes valores de TAE como função do número de grupos no modelo NB - Synthetic-2000.

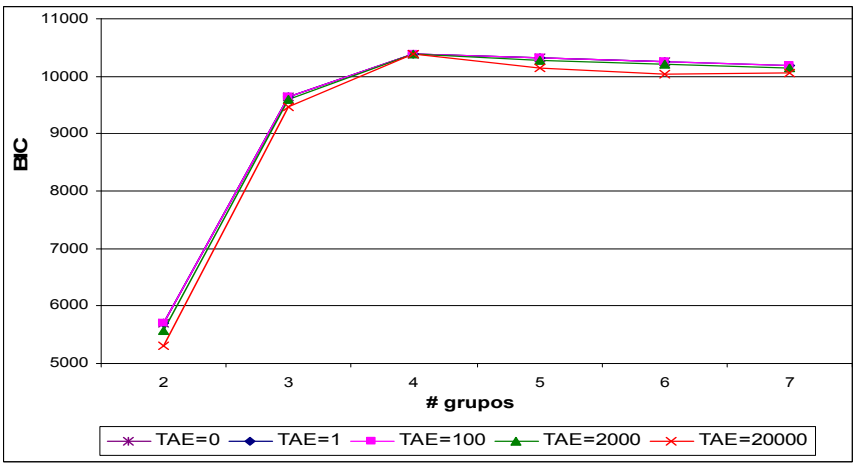


Figura 4.7: Critério BIC para diferentes valores de TAE como função do número de grupos no modelo NB - Synthetic-2000.

4.1.2. SYNTHETIC-1000

Resultados do método hierárquico Bayesiano.

Assim como em Synthetic-2000, foram criados dois dendrogramas a partir de duas partições iniciais, a primeira com 30 grupos gerados pelo algoritmo K-Means com 94% de pureza e cujo dendrograma resultante foi chamado HBC-Kmeans, e a segunda, também com 30 grupos criados pelo algoritmo ART2 com 85% de pureza e cujo dendrograma resultante foi chamado HBC-ART2. Na Figura 4.8 é mostrada a evolução da pureza nas 30 partições de cada dendrograma junto com as purezas das 30 últimas partições de dendrogramas gerados pelos algoritmos *Single Linkage*, *Complete Linkage* e *Average Linkage.* Uma comparação das purezas médias dos dendrogramas é mostrada na Figura 4.9, onde é observado que o dendrograma HBC-Kmeans tem maior pureza média que o HBC-ART2, isto devido a que sua partição inicial também foi melhor. O HBC-Kmeans também apresentou uma pureza similar ao melhor dos outros algoritmos hierárquicos analisados (*Average Linkage)*, mas observando as partições com o número certo de grupos (5 grupos) o HBC-Kmeans apresenta a maior pureza de todos os métodos. A análise do número ótimo de grupos foi realizada sobre o dendrograma HBC-Kmeans, para isso foi calculado o critério BIC em cada uma das suas partições. A Figura 4.10 mostra a curva do BIC resultante, onde pode ser observado que o máximo valor do BIC corresponde à partição de 5 grupos, não obstante, a partição de 4 grupos apresenta um BIC próximo do máximo. Este resultado pode ser explicado pela existência de 2 grupos medianamente sobrepostos no banco de dados (azul e rosa na Figura 3.2) que podem estar favorecendo de certa forma a hipótese da existência de 4 grupos.

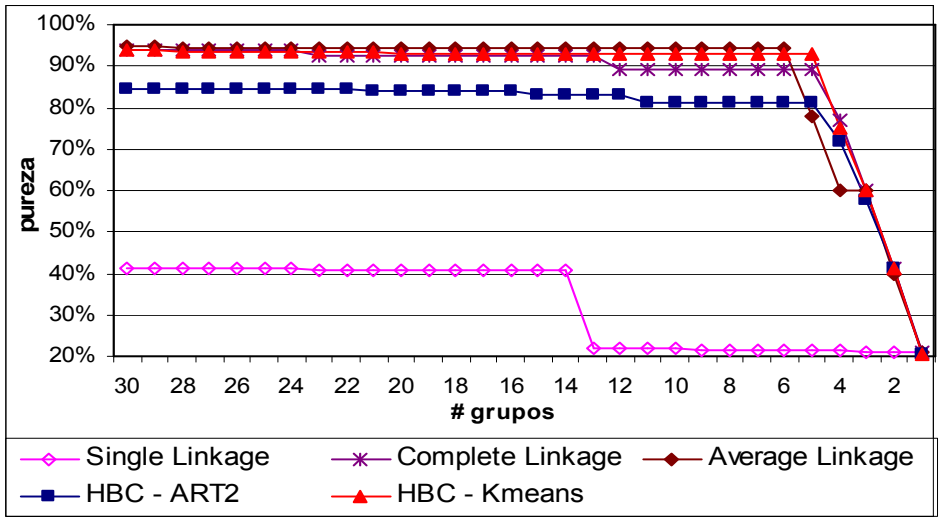


Figura 4.8: Evolução da pureza dos dendrogramas gerados em Synthetic-1000 pelo método hierárquico Bayesiano e métodos hierárquicos tradicionais.

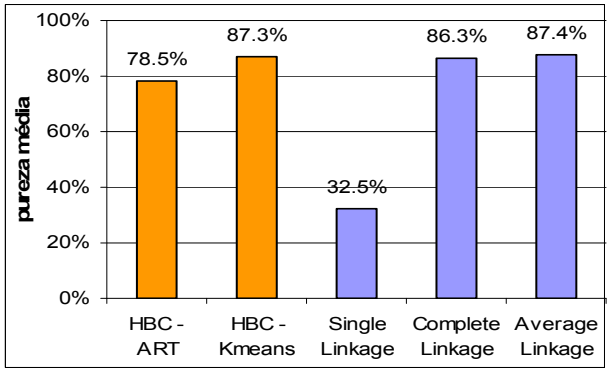


Figura 4.9: Pureza média de dendrogramas gerados em Synthetic-1000 pelo método hierárquico Bayesiano e métodos hierárquicos tradicionais.

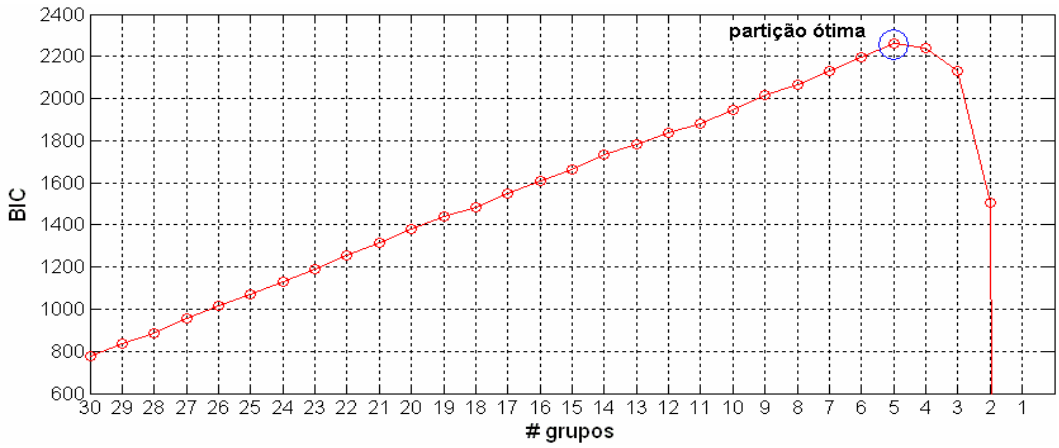


Figura 4.10: Evolução do critério BIC no processo de aglomeração do dendrograma HBC-Kmeans no banco de dados Synthetic-1000.

Na Figura 4.11 é comparada a pureza da partição ótima de 5 grupos contra outras partições de 5 grupos obtidos pelos algoritmos particionais K-Means e EM. Aqui é observado que os métodos particionais analisados geraram agrupamentos com purezas similares e ligeiramente superiores à partição ótima, mas a desvantagem destes métodos é que eles precisam da especificação do número de grupos para efetuar o agrupamento.

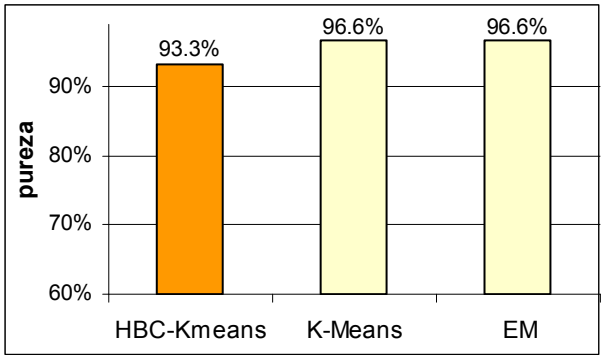


Figura 4.11: Pureza da partição ótima do método hierárquico Bayesiano contra partições de 5 grupos gerados por outros métodos particionais no banco de dados Synthetic-1000.

Resultados do método baseado no modelo NB.

Na Figura 4.12 é mostrada a pureza dos agrupamentos realizados pelo modelo NB (atribuindo os dados ao grupo com maior probabilidade de pertinência). Ao contrário dos resultados em Synthetic-2000, as purezas das distintas partições incrementam com valores altos de TAE (1000 e 10000), o que significa que quanto mais confiança se tem no conhecimento *a priori* (que as classes são uniformes) melhores são os agrupamentos resultantes. Isto pode ser devido a que as classes não estão bem separadas neste banco de dados e conseqüentemente a introdução de conhecimento *a priori* acerca das classes ajuda na obtenção de melhores agrupamentos.

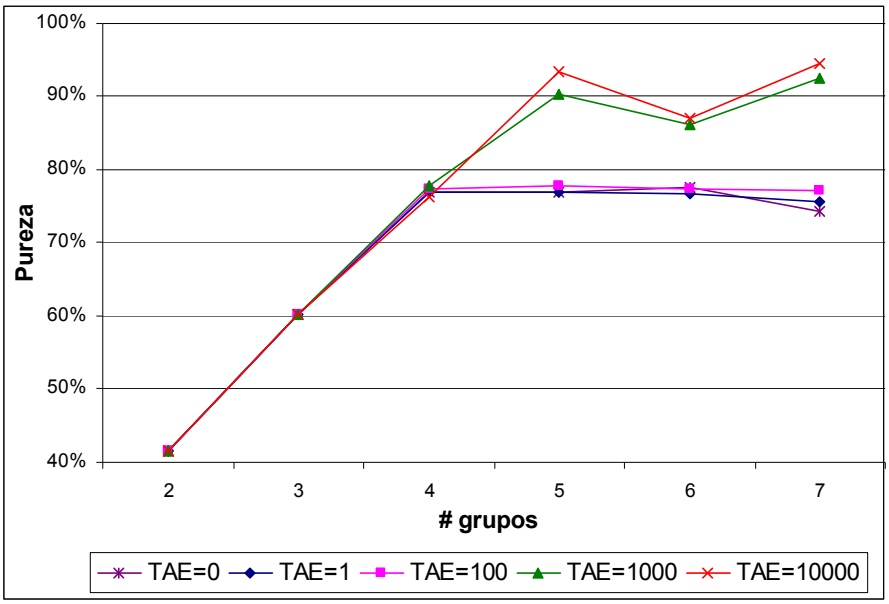


Figura 4.12: Pureza das partições feitas pelo modelo NB em Synthetic-1000.

Nas Figuras 4.13 e 4.14 são mostradas as curvas dos critérios AIC e BIC para diferentes valores de TAE como função do número de grupos. É observado que em ambos critérios o número ótimo de grupos é 4 para valores baixos de TAE (0, 1, 100) e 5 para valores altos de TAE (1000, 10000). Esta discordância pode ser explicada pela natureza *fuzzy* das classes, sendo que um pobre conhecimento *a priori* gera partições ótimas com número errado de grupos. Este resultado comprova o aproveitamento do conhecimento *a priori* no modelo NB na determinação do numero certo de grupos.

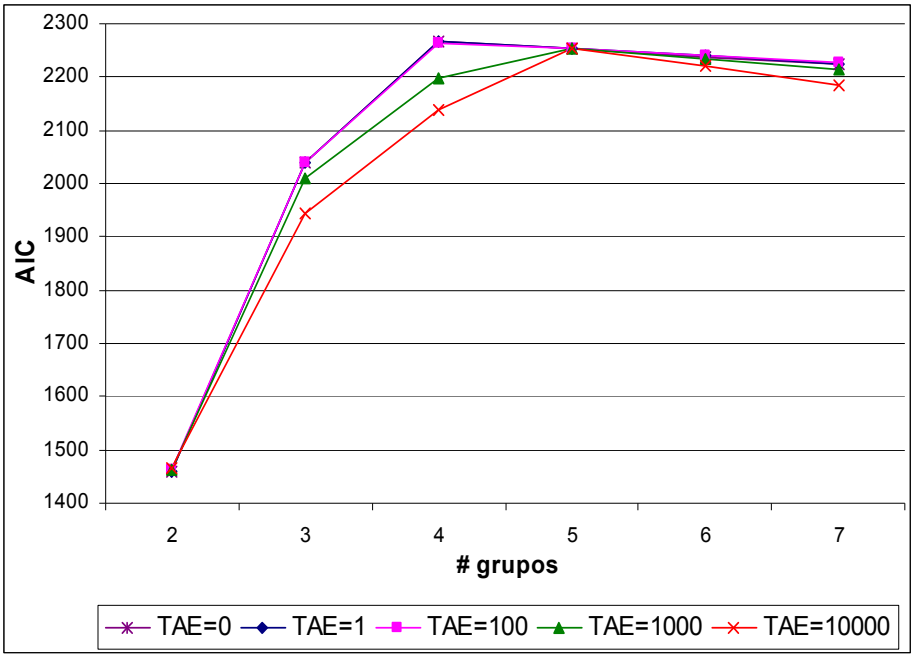


Figura 4.13: Critério AIC para diferentes valores de TAE como função do número de grupos no modelo NB - Synthetic-1000.

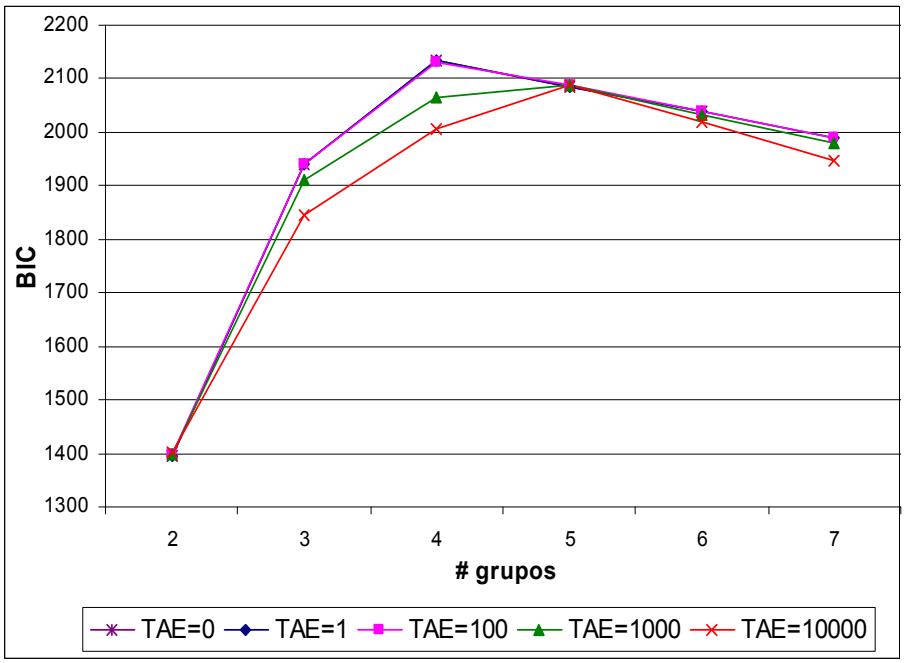


Figura 4.14: Critério BIC para diferentes valores de TAE como função do número de grupos no modelo NB - Synthetic-1000.

4.1.3. IRIS

Resultados do método hierárquico Bayesiano.

Similar aos bancos de dados anteriores, o método hierárquico Bayesiano foi usado para criar dois dendrogramas a partir de duas partições iniciais, a primeira, com 6 grupos gerados pelo algoritmo K-Means com 95.3% de pureza e cujo dendrograma resultante foi chamado HBC-Kmeans, e a segunda, também com 6 grupos criados pelo algoritmo ART2 com 96% de pureza e cujo dendrograma resultante foi chamado HBC-ART2. Na Figura 4.15 é mostrada a evolução da pureza nas 6 partições integrantes de cada dendrograma junto com as purezas das 6 últimas partições dos dendrogramas gerados pelos algoritmos *Single Linkage*, *Complete Linkage* e *Average Linkage.* Uma comparação das purezas médias dos dendrogramas é mostrada na Figura 4.16, onde é observado que os dendrogramas HBC-Kmeans e HBC-ART2 têm purezas médias similares e significativamente maiores que os outros dendrogramas.

A análise do número ótimo de grupos foi realizada sobre o dendrograma HBC-Kmeans, para isso foi calculado o critério BIC em cada uma das suas partições. A Figura 4.17 mostra a curva do BIC resultante, onde é observado que o máximo valor do BIC corresponde à partição de 3 grupos, o qual é o número certo de grupos. É observado também que a partição de 2 grupos apresenta o segundo maior valor do BIC, o qual está próximo do máximo. Isto é explicado pela conhecida existência de 2 grupos sobrepostos em Íris (virgínica e versicolor) que favorecem a hipótese da existência de somente 2 grupos no banco de dados.

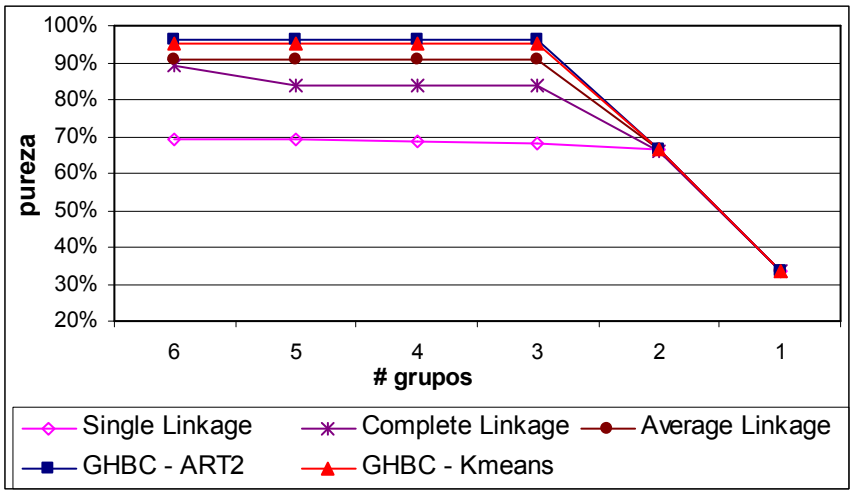


Figura 4.15: Evolução da pureza dos dendrogramas gerados em Iris pelo método hierárquico Bayesiano e métodos hierárquicos tradicionais.

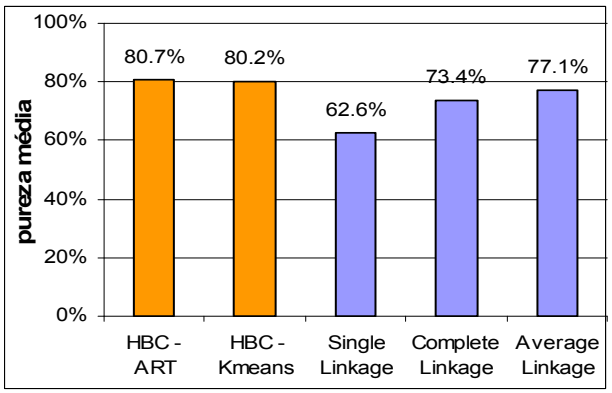


Figura 4.16: Pureza média de dendrogramas gerados em Iris pelo método hierárquico Bayesiano e métodos hierárquicos tradicionais.

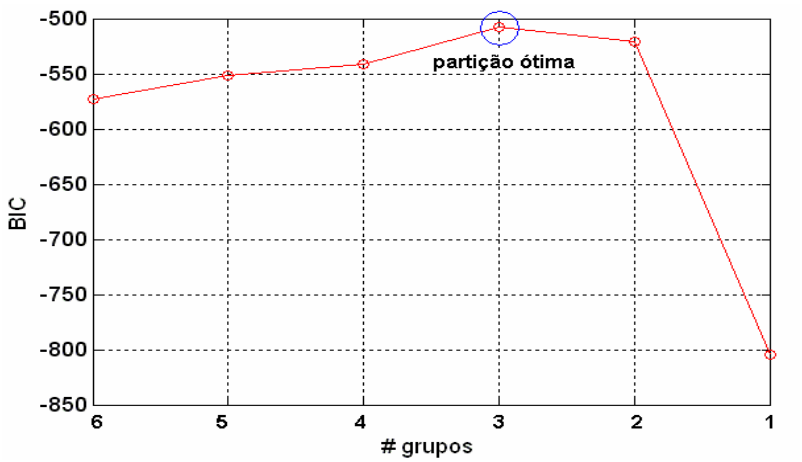


Figura 4.17: Evolução do critério BIC no processo de aglomeração do dendrograma HBC-Kmeans no banco de dados Iris.

Na Figura 4.18 é comparada a pureza da partição ótima contra outras partições de 3 grupos obtidos pelos algoritmos particionais K-Means e EM. Aqui é observado que a pureza da partição ótima é significativamente maior que os dendrogramas dos outros métodos analisados. Este resultado mostra a conveniência do método hierárquico Bayesiano em dados reais e que não necessariamente seguem uma distribuição gaussiana.

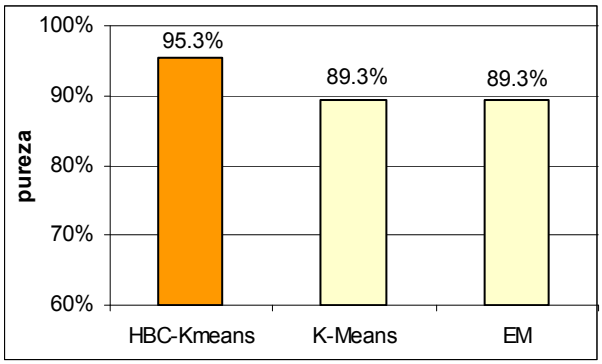


Figura 4.18: Pureza da partição ótima do método hierárquico Bayesiano contra partições de 3 grupos gerados por outros métodos particionais no banco de dados Iris.

Resultados do método baseado no modelo NB.

Na Figura 4.19 é mostrada a pureza dos agrupamentos realizados pelo modelo NB (atribuindo os dados ao grupo com maior probabilidade de pertinência). Similar aos resultados em Synthetic-1000, as purezas das distintas partições incrementam com o valor de TAE, o que significa que quanto mais confiança se tem no conhecimento *a priori* (que as classes são uniformes) melhores são os agrupamentos resultantes. Desta forma é mostrado o aproveitamento do conhecimento *a priori* existente no modelo NB para a obtenção de melhores agrupamentos (com grupos mais homogêneos).

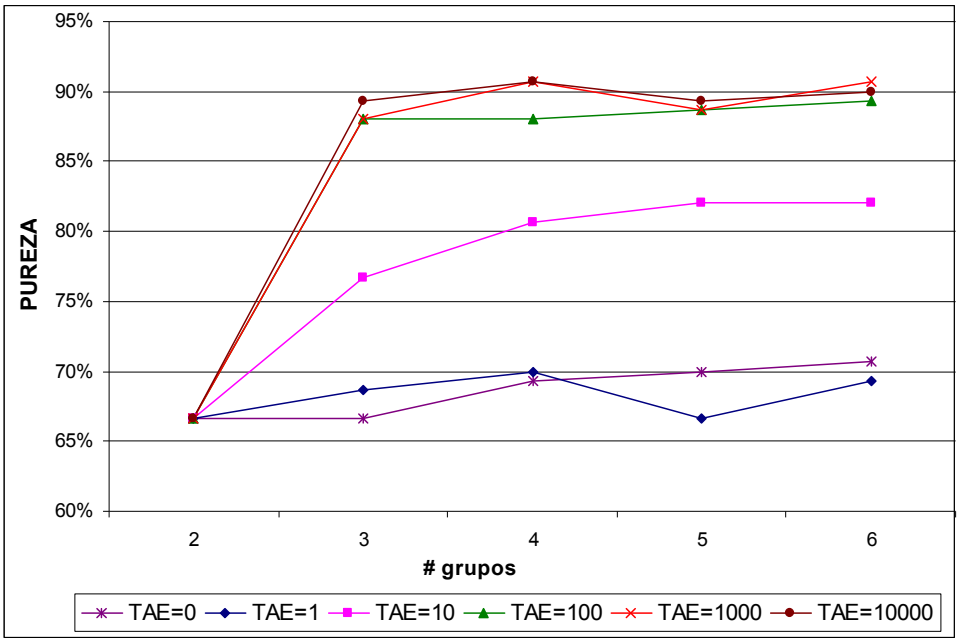


Figura 4.19: Pureza das partições feitas pelo modelo NB em Iris.

Nas Figuras 4.20 e 4.21 são mostradas as curvas dos critérios AIC e BIC para diferentes valores de TAE como função do número de grupos. É observado que no critério AIC, a partição ótima tem 2 grupos para valores baixos de TAE (0 - 100) e 3 grupos para valores altos de TAE (1000, 10000), já no critério BIC, todas as partições ótimas contêm 2 grupos independentemente do valor de TAE. Esta discordância é explicada pela alta sobreposição de duas classes em Íris, sendo esta sobreposição o suficientemente grande para que inclusive com um forte conhecimento *a priori* acerca da distribuição das classes não seja possível determinar o número certo de grupos com o critério BIC. Já com o critério AIC é possível determinar o número certo de grupos para altos valores de TAE devido ao fato que o critério AIC penaliza menos a formação de partições com uma alta quantidade de grupos.

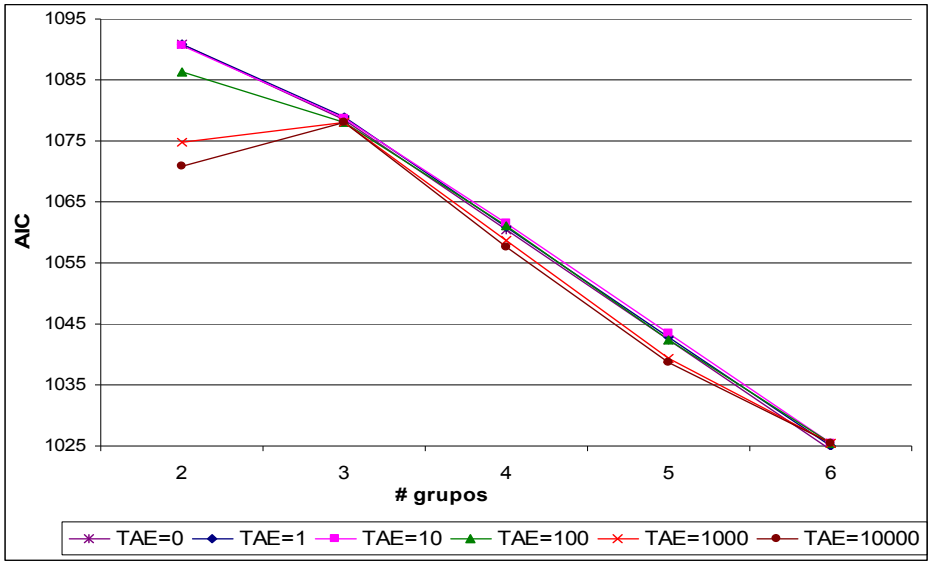


Figura 4.20: Critério AIC para diferentes valores de TAE como função do número de grupos no modelo NB - Iris.

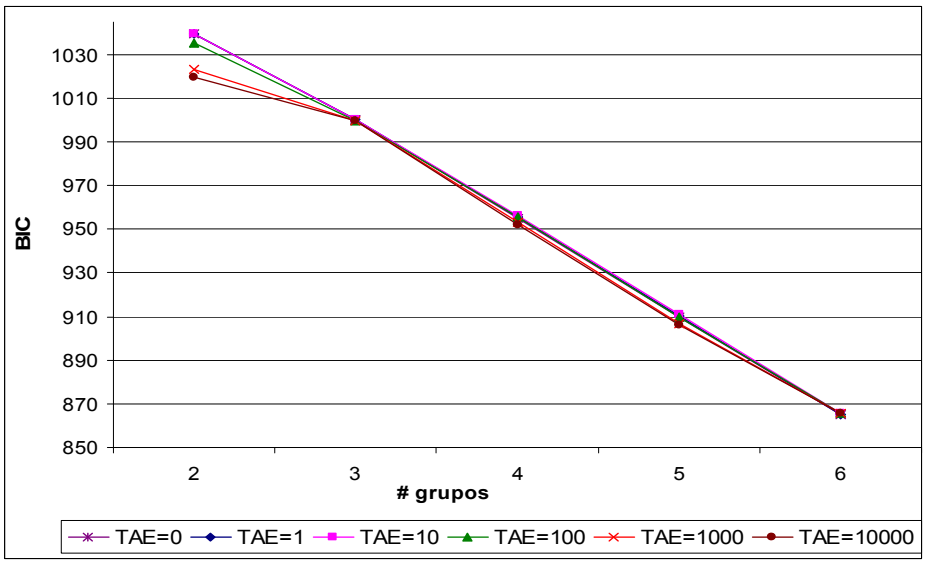


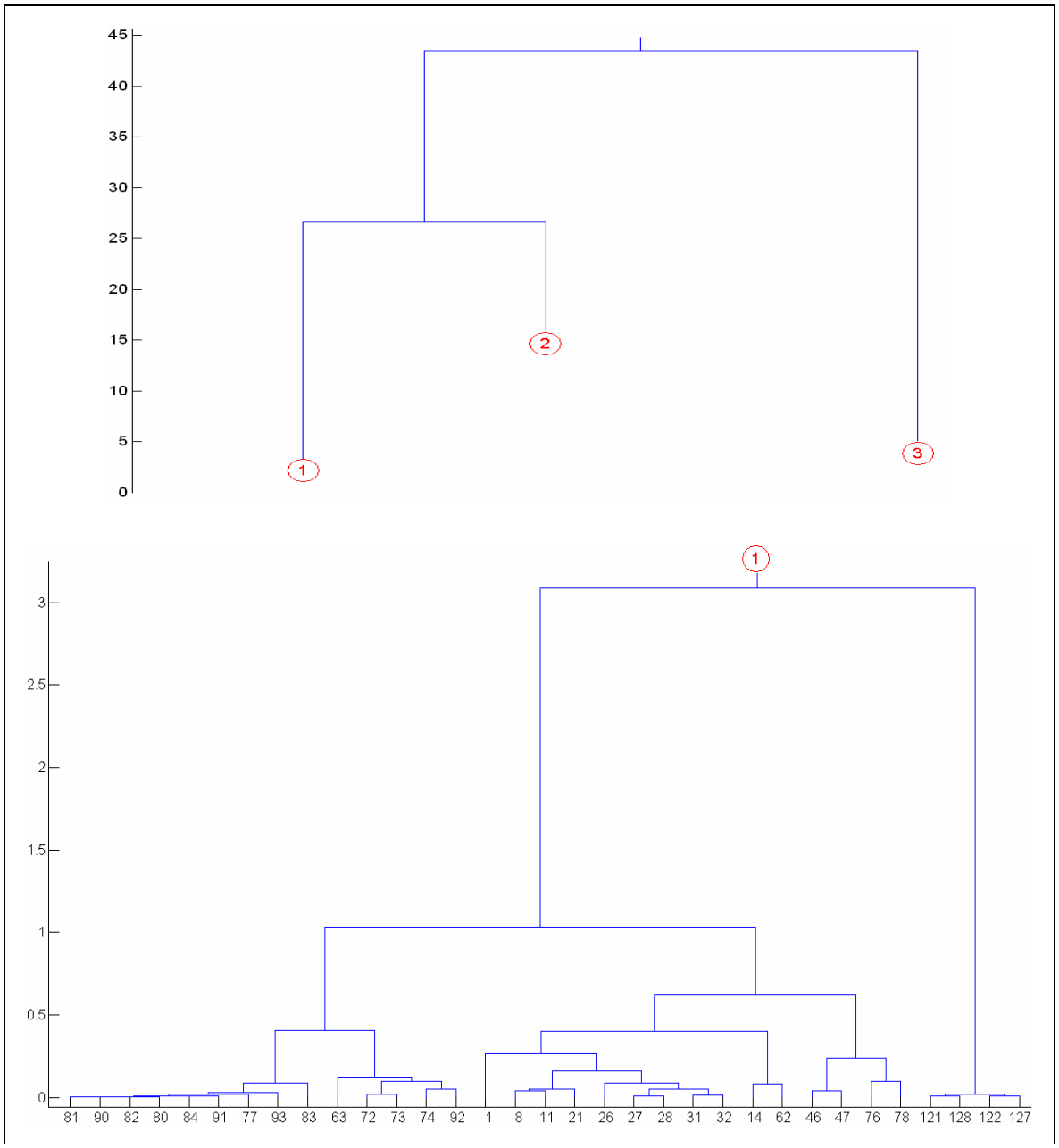
Figura 4.21: Critério BIC para diferentes valores de TAE como função do número de grupos no modelo NB - Iris.

4.2. RESULTADOS NOS BANCO DE ESTIRPES DE *BRADYRHIZOBIUM*

Resultados do método hierárquico Bayesiano.

O método hierárquico Bayesiano foi usado aqui para construir uma

representação em dendrograma das estirpes de Bradyrhizobium e encontrar o melhor agrupamento destas. Para isso foram usados como dados de entrada os eletroferogramas resultantes do pré-processamento das canaletas (Figura 3.5). O dendrograma resultante é mostrado na Figura 4.22, o qual, devido a seu tamanho. foi dividido em 4 partes: um dendrograma principal e 3 sub-dendrogramas. A escala de distâncias mostrada corresponde ao logaritmo do fator U (Seção 3.2).



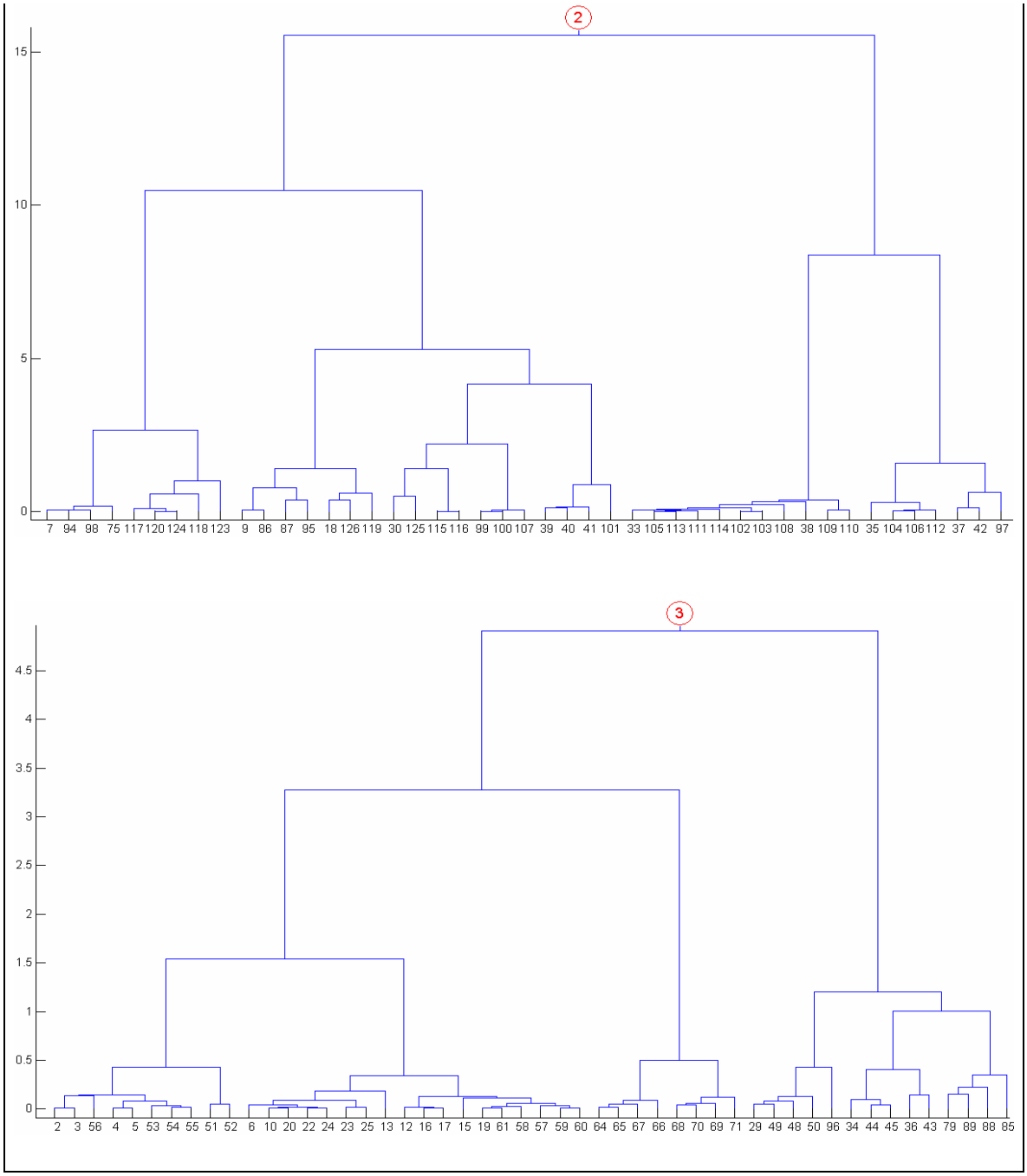


Figura 4.22: Dendrograma das estirpes de Bradyrhizobium construído pelo método hierárquico Bayesiano tendo como dados de entrada os eletroferogramas resultantes do pré-processamento das canaletas.

Na Figura 4.23 é mostrada a evolução da verossimilhança e dos critérios BIC e AIC calculados nas 30 últimas partições do dendrograma resultante. Como pode ser observada, a verossimilhança decresce à medida que se vão fundindo grupos, este

fato era esperado já que uma propriedade da verossimilhança é que esta sempre favorece partições com maior número de grupos (propriedade de *overfitting*), motivo pelo qual não pode ser usada para encontrar o número ótimo de grupos. É observado também que no critério AIC, a partição com o maior valor é a de 4 grupos, não obstante, existe uma alta proximidade entre o AIC desta partição e o AIC da partição de 9 grupos, pelo que não é muito clara a escolha da partição ótima. Já no critério BIC esta ambigüidade fica esclarecida a favor da partição de 4 grupos, isto devido a que o critério BIC tem uma maior penalidade para partições com alto número de grupos. Na Tabela 4.1 são mostrados os dados integrantes de cada grupo desta partição ótima.

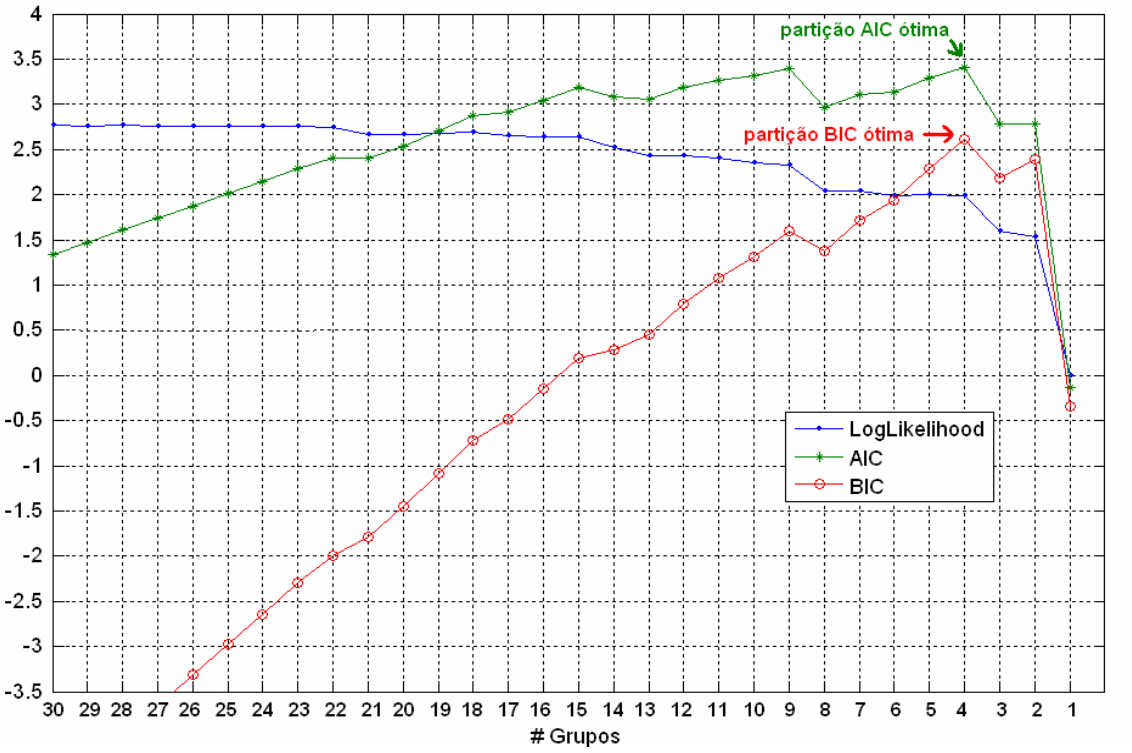


Figura 4.23: Evolução da verossimilhança e dos critérios BIC e AIC calculados nas 30 últimas partições do dendrograma construído pelo método hierárquico Bayesiano no banco de estirpes de *Bradyrhizobium*

Tabela 4.1: Dados integrantes dos grupos que compõem a partição BIC ótima do dendrograma construído pelo método hierárquico Bayesiano no banco de estirpes de Bradyrhizobium.

G1	33 35 37 38 42 97 102 103 104 105 106 108 109 110 111 112 113 114
G2	1 8 11 14 21 26 27 28 31 32 46 47 62 63 72 73 74 76 77 78 80 81 82 83 84 90 91 92 93 121 122 127 128
G3	2 3 4 5 6 10 12 13 15 16 17 19 20 22 23 24 25 29 34 36 43 44 45 48 49 50 51 52 53 54 55 56 57 58 59 60 61 64 65 66 67 68 69 70 71 79 85 88 89 96
G4	7 9 18 30 39 40 41 75 86 87 94 95 98 99 100 101 107 115 116 117 118 119 120 123 124 125 126

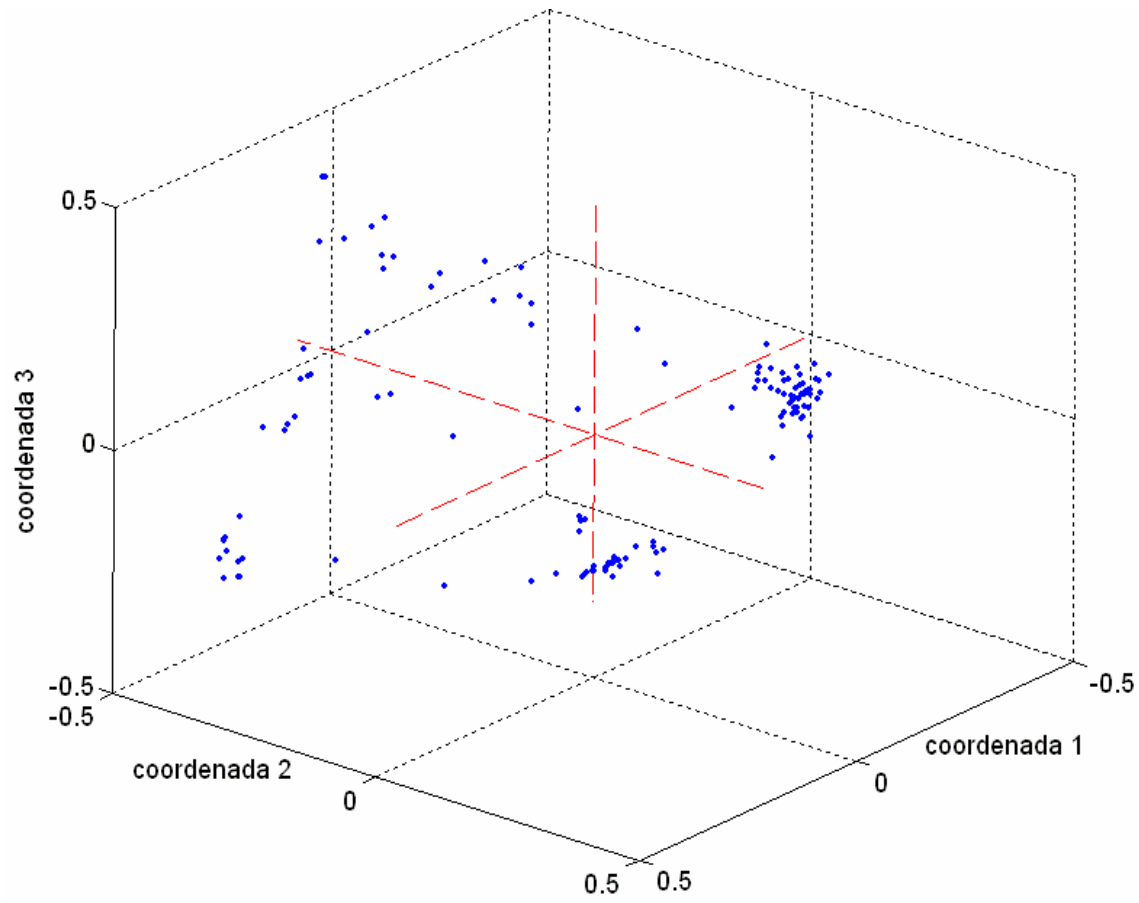


Figura 4.24: Pontos resultantes da transformação MDS dos eletroferogramas tomando as 3 primeiras coordenadas, representando o 50% da variância dos dados.

Resultados do método baseado no modelo NB.

Na Figura 4.24 são representadas as 3 primeiras coordenadas das 10 que compõem o banco de dados resultante da transformação dos eletroferogramas por MDS. De acordo com a análise de autovalores (Seção 3.3.1) aproximadamente o 50% da variância dos dados está contido nestas 3 primeiras coordenadas, o que significa que a representação em 3D perde muitos detalhes dos dados e é usada somente para dar uma idéia da distribuição dos mesmos.

O modelo NB foi implementado para agrupar o banco de dados transformado. Foram realizados vários experimentos seguindo a estratégia de testes descrita no capítulo anterior (Seção 3.4). As Figuras 4.25 e 4.26 mostram as curvas dos critérios estatísticos BIC e AIC para diferentes valores de TAE como função do número de grupos. Pode ser observado que no caso do critério AIC, existe uma alta ambigüidade na determinação do número ótimo de grupos, podendo ser selecionada qualquer das partições entre 7 e 10 grupos como a partição ótima, sendo a partição com 9 grupos a que destaca levemente com respeito às outras partições. É observado também que em partições com alto número de grupos (5-10) as curvas com valores grandes de TAE estão por embaixo das curvas com TAEs menores, sendo o contrário para partições com baixo número de grupos (2-4). Isto significa que quanto mais confiança se dá à hipótese que as classes são uniformes, menor é a aceitação de partições com alto número de grupos. Já com o critério BIC (Figura 4.26), a incerteza na determinação da partição ótima desaparece, sendo identificada claramente a partição de 4 grupos como a partição ótima em todas as curvas de TAE. Este resultado é ainda mais claro para valores altos de TAE (1000 e 10000). A discordância de ambos critérios na determinação da partição ótima é devida a que o critério BIC dá uma maior penalidade que o critério AIC para partições com um alto

número de grupos. A discussão de qual critério deve ser usado está fora do alcance deste trabalho. No caso do banco de estirpes de *Bradyrhizobium* e baseados em trabalhos prévios (GERMANO *et. al.*, 2006), não se tem motivos fortes para escolher partições com alto número de grupos (>7), razão pela qual foi escolhida como partição ótima a de 4 grupos indicada pelo critério BIC.

Na Tabela 4.2 são mostradas as probabilidades de pertinência dos dados nos distintos grupos para o modelo NB ótimo de 4 grupos e TAE = 100. Foi verificado que estas probabilidades são muito similares para os outros valores de TAE. Conseqüentemente, nos modelos NB ótimos de 4 grupos, as partições resultantes de atribuir os dados ao grupo de maior probabilidade são similares sem importar o valor de TAE. Na Figura 4.27 são mostrados os dados em 3D da partição ótima de 4 grupos para o TAE=100, onde cada cor indica um grupo, sendo os mesmos agrupamentos observados para os outros valores de TAE.

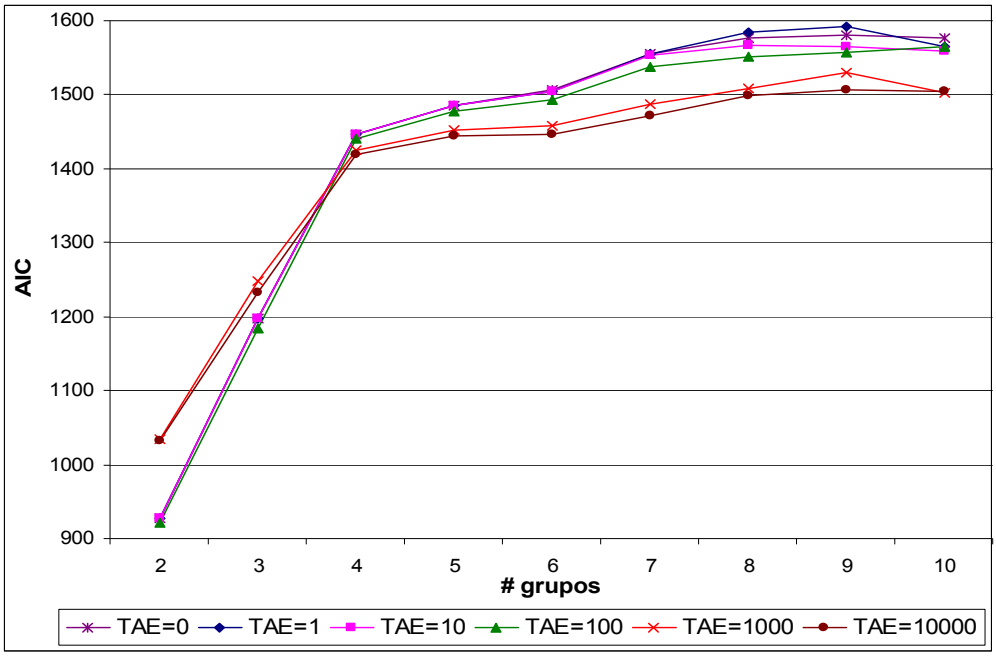


Figura 4.25: Critério AIC para diferentes valores de TAE como função do número de grupos no modelo NB – banco de estirpes de *Bradyrhizobium*.

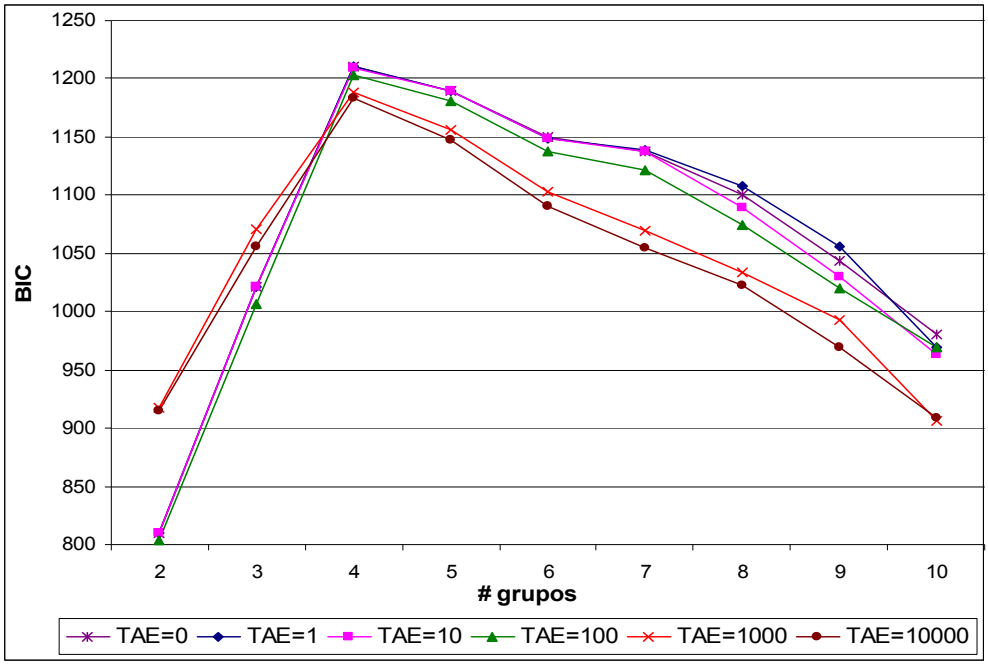


Figura 4.26: Critério BIC para diferentes valores de TAE como função do número de grupos no modelo NB – banco de estirpes de *Bradyrhizobium*.

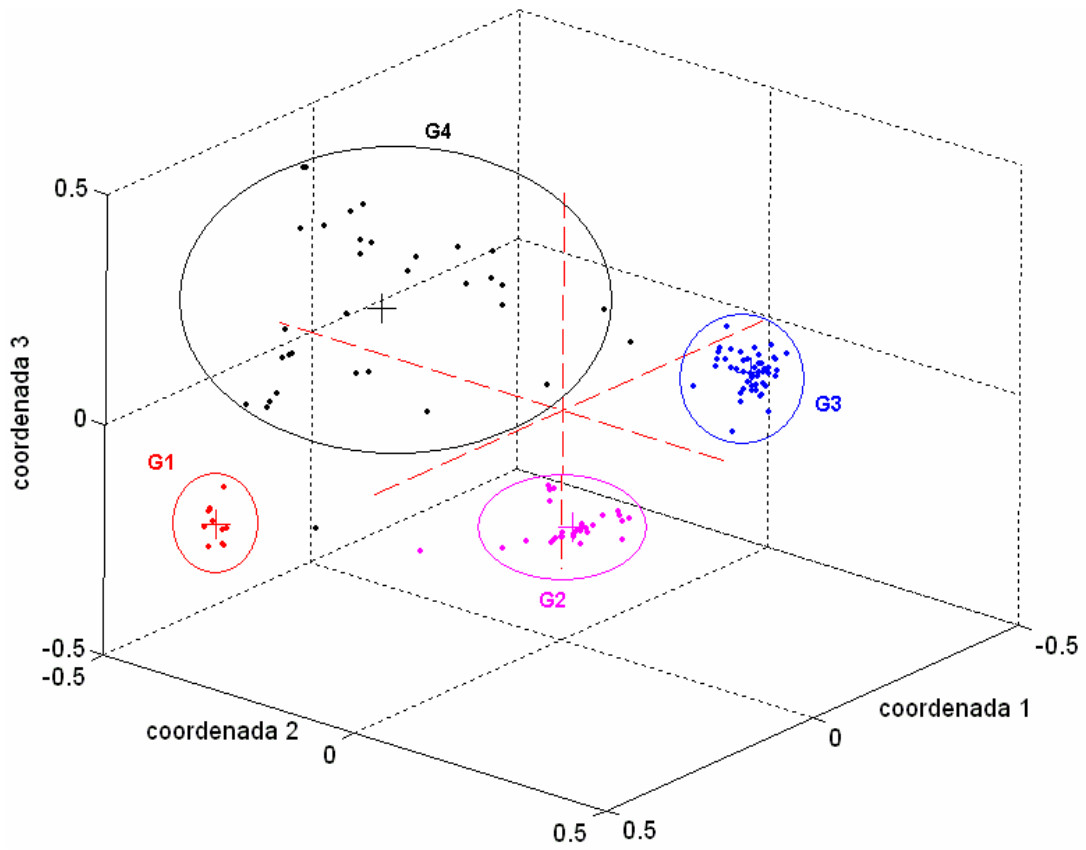


Figura 4.27: Agrupamento ótimo de 4 grupos gerado pelo método baseado no modelo NB no banco de estirpes de *Bradyrhizobium*

A Tabela 4.3 mostra a partição ótima de 4 grupos como resultado da atribuição dos dados aos grupos com maior probabilidade de pertinência (Tabela 4.2). Na Tabela 4.4 são mostradas as médias e as variâncias de cada grupo calculadas com base a seus dados atribuídos. Analisando estes resultados foi encontrado que os grupos G1 (vermelho), G2 (roxo) e G3 (azul) são bastante compactos e bem separados, apresentando uma covariância média igual a 0.032, 0.029 e 0.040 respectivamente. Já o grupo G4 (cor negro) apresenta uma variância média de 0.167, a qual é significativamente maior que os anteriores, o que é comprovado na Figura 4.27 com seus elementos bastante espalhados. Esta dispersão dos dados pode significar que as estirpes deste grupo não estão compartindo muita informação genética em comum, tal como os outros grupos, e portanto este grupo não representa nenhuma sub-espécie. Ou também pode significar que não existem dados suficientes para se inferir novas sub-espécies.

Comparando a partição ótima obtida pelo método hierárquico Bayesiano (Tabela 4.1) com a partição ótima gerada pelo método baseado no modelo NB (Tabela 4.3) é possível observar uma alta concordância em ambos agrupamentos. Os grupos G2 e G3 são exatamente iguais, entretanto, no grupo G1 do método hierárquico, existem 7 dados mais que o grupo G1 do método NB (35, 37, 42, 97,104,106,112). Isto significa uma discrepância de somente 5.4% entre ambos particionamentos. Estes resultados indicam uma alta probabilidade de que os grupos G2 e G3 representem alguma sub-espécie de estirpes de *Bradyrhizobium* com características comuns. No entanto, esta probabilidade é menor no grupo G1 devido à discrepância observada no agrupamento de 7 elementos por cada um dos métodos implementados.

Tabela 4.2: Probabilidades de pertinência resultantes do modelo NB ótimo de 4 classes construído no banco de estirpes de Bradyrhizobium.

Estirpe	G1	G2	G3	G4
1	0.000	1.000	0.000	0.000
2	0.000	0.000	1.000	0.000
3	0.000	0.000	1.000	0.000
4	0.000	0.000	1.000	0.000
5	0.000	0.000	1.000	0.000
6	0.000	0.000	1.000	0.000
7	0.025	0.000	0.000	0.975
8	0.000	1.000	0.000	0.000
9	0.000	0.000	0.000	1.000
10	0.000	0.000	1.000	0.000
11	0.000	1.000	0.000	0.000
12	0.000	0.000	1.000	0.000
13	0.000	0.000	1.000	0.000
14	0.000	1.000	0.000	0.000
15	0.000	0.000	1.000	0.000
16	0.000	0.000	1.000	0.000
17	0.000	0.000	1.000	0.000
18	0.000	0.000	0.000	1.000
19	0.000	0.000	1.000	0.000
20	0.000	0.000	1.000	0.000
21	0.000	1.000	0.000	0.000
22	0.000	0.000	1.000	0.000
23	0.000	0.000	1.000	0.000
24	0.000	0.000	1.000	0.000
25	0.000	0.000	1.000	0.000
26	0.000	1.000	0.000	0.000
27	0.000	1.000	0.000	0.000
28	0.000	1.000	0.000	0.000
29	0.000	0.000	1.000	0.000
30	0.000	0.000	0.000	1.000
31	0.000	1.000	0.000	0.000
32	0.000	1.000	0.000	0.000
33	1.000	0.000	0.000	0.000
34	0.000	0.000	1.000	0.000
35	0.050	0.000	0.000	0.950
36	0.000	0.000	1.000	0.000
37	0.060	0.000	0.000	0.940
38	1.000	0.000	0.000	0.000
39	0.003	0.000	0.000	0.997
40	0.001	0.000	0.000	0.999
41	0.017	0.000	0.000	0.983
42	0.050	0.000	0.000	0.950
43	0.000	0.000	1.000	0.000
44	0.000	0.000	1.000	0.000
45	0.000	0.000	1.000	0.000
46	0.000	1.000	0.000	0.000
47	0.000	1.000	0.000	0.000
48	0.000	0.000	1.000	0.000
49	0.000	0.000	1.000	0.000
50	0.000	0.000	1.000	0.000
51	0.000	0.000	1.000	0.000
52	0.000	0.000	1.000	0.000
53	0.000	0.000	1.000	0.000
54	0.000	0.000	1.000	0.000
55	0.000	0.000	1.000	0.000
56	0.000	0.000	1.000	0.000
57	0.000	0.000	1.000	0.000
58	0.000	0.000	1.000	0.000
59	0.000	0.000	1.000	0.000
60	0.000	0.000	1.000	0.000
61	0.000	0.000	1.000	0.000
62	0.000	1.000	0.000	0.000
63	0.000	1.000	0.000	0.000
64	0.000	0.000	1.000	0.000

Estirpe	G1	G2	G3	G4
65	0.000	0.000	1.000	0.000
66	0.000	0.000	1.000	0.000
67	0.000	0.000	1.000	0.000
68	0.000	0.000	1.000	0.000
69	0.000	0.000	1.000	0.000
70	0.000	0.000	1.000	0.000
71	0.000	0.000	1.000	0.000
72	0.000	1.000	0.000	0.000
73	0.000	1.000	0.000	0.000
74	0.000	1.000	0.000	0.000
75	0.017	0.000	0.000	0.983
76	0.000	1.000	0.000	0.000
77	0.000	1.000	0.000	0.000
78	0.000	1.000	0.000	0.000
79	0.000	0.000	1.000	0.000
80	0.000	1.000	0.000	0.000
81	0.000	1.000	0.000	0.000
82	0.000	1.000	0.000	0.000
83	0.000	1.000	0.000	0.000
84	0.000	1.000	0.000	0.000
85	0.000	0.000	1.000	0.000
86	0.000	0.000	0.000	1.000
87	0.013	0.000	0.000	0.987
88	0.000	0.000	1.000	0.000
89	0.000	0.000	1.000	0.000
90	0.000	1.000	0.000	0.000
91	0.000	1.000	0.000	0.000
92	0.000	1.000	0.000	0.000
93	0.000	1.000	0.000	0.000
94	0.016	0.000	0.000	0.984
95	0.029	0.000	0.000	0.971
96	0.000	0.000	1.000	0.000
97	0.176	0.000	0.000	0.824
98	0.012	0.000	0.000	0.988
99	0.001	0.000	0.000	0.999
100	0.001	0.000	0.000	0.999
101	0.000	0.000	0.000	1.000
102	1.000	0.000	0.000	0.000
103	1.000	0.000	0.000	0.000
104	0.041	0.000	0.000	0.959
105	1.000	0.000	0.000	0.000
106	0.041	0.000	0.000	0.959
107	0.000	0.000	0.000	1.000
108	1.000	0.000	0.000	0.000
109	1.000	0.000	0.000	0.000
110	1.000	0.000	0.000	0.000
111	1.000	0.000	0.000	0.000
112	0.052	0.000	0.000	0.948
113	1.000	0.000	0.000	0.000
114	1.000	0.000	0.000	0.000
115	0.000	0.000	0.000	1.000
116	0.000	0.000	0.000	1.000
117	0.000	0.000	0.000	1.000
118	0.000	0.000	0.000	1.000
119	0.000	0.000	0.000	1.000
120	0.000	0.000	0.000	1.000
121	0.000	1.000	0.000	0.000
122	0.000	1.000	0.000	0.000
123	0.017	0.000	0.000	0.983
124	0.000	0.000	0.000	1.000
125	0.000	1.000	0.000	0.000
126	0.000	0.000	0.000	1.000
127	0.000	1.000	0.000	0.000
128	0.000	1.000	0.000	0.000

Tabela 4.3: Atribuição dos dados aos diversos grupos da partição ótima de 4 grupos gerada pelo modelo NB no banco de estirpes de Bradyrhizobium.

G1	33 38 102 103 105 108 109 110 111 113 114
G2	1 8 11 14 21 26 27 28 31 32 46 47 62 63 72 73 74 76 77 78 80 81 82 83 84 90 91 92 93 121 122 125 127 128
G3	2 3 4 5 6 10 12 13 15 16 17 19 20 22 23 24 25 29 34 36 43 44 45 48 49 50 51 52 53 54 55 56 57 58 59 60 61 64 65 66 67 68 69 70 71 79 85 88 89 96
G4	7 9 18 30 35 37 39 40 41 42 75 86 87 94 95 97 98 99 100 101 104 106 107 112 115 116 117 118 119 120 123 124 126

Tabela 4.4: Parâmetros dos grupos integrantes da partição ótima de 4 grupos gerada pelo modelo NB no banco de estirpes de Bradyrhizobium.

Parâmetros	G1 (vermelho) # elems = 11		G2 (roxa) # elems = 34		G3 (azul) # elems = 50		G4 (negro) # elems = 33	
	média	variância	média	variância	média	variância	média	variância
Coord 1	0.24	0.047	0.31	0.057	-0.35	0.062	0.13	0.111
Coord 2	-0.49	0.045	0.27	0.038	0.09	0.051	-0.26	0.184
Coord 3	-0.32	0.021	-0.04	0.044	-0.03	0.019	0.19	0.171
Coord 4	-0.05	0.030	0.00	0.025	0.00	0.065	0.01	0.238
Coord 5	-0.04	0.037	0.00	0.032	0.00	0.055	0.01	0.228
Coord 6	0.02	0.043	0.01	0.024	0.01	0.036	-0.03	0.192
Coord 7	0.03	0.026	0.00	0.020	0.00	0.041	-0.01	0.158
Coord 8	-0.01	0.010	0.00	0.022	0.00	0.027	0.01	0.148
Coord 9	0.01	0.042	0.00	0.016	0.00	0.028	0.00	0.141
Coord 10	0.00	0.018	0.00	0.014	0.00	0.013	-0.01	0.102

Capítulo 5

CONCLUSÕES E SUGESTÕES

No presente trabalho foram implementados e testados dois métodos de agrupamento de dados visados para aplicações em taxonomia molecular. Tendo em vista os resultados obtidos, chegou-se às conclusões descritas a seguir.

O método de agrupamento hierárquico Bayesiano (HBC) demonstrou a capacidade de gerar dendrogramas de boa qualidade nos bancos de dados de validação. No caso dos bancos sintéticos, HBC foi similar ao melhor dos algoritmos hierárquicos analisados. No caso do banco de dados Íris, ele foi significativamente superior. HBC também demonstrou a capacidade de determinar o número certo de grupos em todos os bancos de validação mediante o critério BIC. As partições ótimas de HBC mostraram ser também de alta qualidade quando comparados com agrupamentos gerados por algoritmos particionais como K-Means e EM. Com estes resultados em mente, HBC foi usado para gerar uma representação em dendrograma do banco de dados genotípicos de estirpes de *Bradyrhizobium*, encontrando-se certa ambigüidade na seleção da partição ótima com o critério AIC. Já no critério BIC esta ambigüidade foi esclarecida a favor da partição de 4 grupos.

O método HBC mostrou não necessitar da escolha de alguma métrica de distância para gerar o dendrograma, isto porque a aglomeração é conduzida pelo

critério de máxima probabilidade *a posteriori* das partições. É sugerido para trabalhos futuros neste método o estudo da redução da complexidade computacional, a qual é quadrática com o número de dados.

No método de agrupamento baseado em redes gaussianas condicionais optou-se pelo modelo *Naive Bayes* (NB) devido a sua simplicidade. Os resultados nos bancos de validação foram aceitáveis, mostrando-se que o método aproveita de forma positiva o conhecimento *a priori* acerca da distribuição das classes para gerar agrupamentos com uma maior qualidade. Este conhecimento é também aproveitado positivamente para esclarecer a determinação do número ótimo de grupos, não obstante, em casos onde as classes são muito sobrepostas como o banco de dados Íris, este conhecimento *a priori* não foi suficiente para determinar o número certo de grupos com o critério BIC.

No caso do agrupamento do banco de estirpes de *Bradyrhizobium* com o método baseado em NB, foi necessário realizar uma transformação dos dados em vetores de coordenadas, o qual implicou a utilização de uma métrica de distância (correlação de Pearson). Esta transformação também implicou certa perda de informação em beneficio da simplicidade do modelo. Nos resultados, foi observada certa ambigüidade na determinação da partição ótima com o critério AIC, dando como partições possíveis entre 7 e 10 grupos. Já com o critério BIC a incerteza na determinação da partição ótima desapareceu, sendo identificada a partição de 4 grupos como a partição ótima, para todos os valores do parâmetro TAE. Um ponto importante observado nas partições com o número ótimo (BIC) de grupos foi que, após de atribuir os dados aos grupos com a maior probabilidade, os agrupamentos resultantes foram iguais para todos os valores de TAE. Isto implica que os

agrupamentos são bastante estáveis com respeito ao conhecimento *a priori* acerca das classes quando o número de grupos é igual ao ótimo (BIC).

Observando a partição (BIC) ótima do banco de estirpes de *Bradyrhizobium* gerada pelo método baseado em NB, foram encontrados 3 grupos (G1-G3) bastante compactos e separados. O quarto grupo apresentou uma alta variância em todas seus coordenadas, o que significa que as estirpes agrupadas nele não compartem muita informação genética em comum, como os outros grupos, e portanto este grupo possivelmente não representa nenhuma sub-espécie. Comparando esta partição ótima com a partição ótima do método hierárquico Bayesiano, foi encontrada uma alta concordância (94.6%) em ambos agrupamentos, com os grupos G2 e G3 exatamente iguais em ambas partições, e uma discrepância de 7 elementos nos grupos G1 e G4. Sugere-se para futuros trabalhos o uso de outras estruturas de modelo como a ENB, TANB, ou aprender as estruturas dos mesmos dados. Isto com a finalidade de comparar os resultados do NB com modelos mais expressivos.

Um outro ponto importante no método de agrupamento baseado em NB, foi na etapa de estimativa de parâmetros devido à dificuldade em conseguir bons aprendizados, isto é, modelos com verossimilhança alta. Esta dificuldade foi devida à natureza estocástica do algoritmo de aprendizado EM, o qual não garante parâmetros ótimos globais. Por outro lado este algoritmo apresentou o maior custo computacional (em tempo) do método. É sugerido para trabalhos futuros o estudo de outros algoritmos de aprendizado que superem estas limitações.

Finalmente, a biblioteca BNT foi de grande utilidade na implementação do classificador NB já que esta tem implementado um amplo conjunto de funções para modelos gráficos probabilísticos.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. (1974). A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v.19, n.6, p.716–723.

ANDERBERG, M. R. (1973). **Cluster Analysis for Applications**. New York: Academic.

BANFIELD, J. D.; RAFTERY, A. E. (1993). Model based gaussian and non-gaussian clustering. **Biometrics,** v.49, p.803-821.

BERKHIN P. (2002). **Survey of Clustering Data Mining Techniques**. In: Accrue Sotware.

BOUCKAERT, R. R. (1995). **Bayesian Belief Networks: From Construction to inference**. PhD Thesis, Faculteit Wiskunde en Informática. University of Utrecht - Utrecht.

CARPENTER, G. A.; GROSSBERG, S. (1987). ART2: self-organization of stable category recognition codes for analog input patterns. **Applied Optics,** v.26, n.23, p.4919-4930.

CASTILLO, E.; GUTIERREZ, J. M.; HADI, A. S. (1997). **Expert Systems and Probabilistic Network Models**. Springer-Verlag.

CHRIST, R. E.; VILLANUEVA, E. R.; MACIEL, C. D. (2007). Gaussian Hierarchical Bayesian Clustering algorithm. **Proceedings of The seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)**. Rio de Janeiro, 2007. In press.

COWELL, R. G.; DAWID, A. P.; LAURITZEN, S. L.; SPIEGELHALTER, D. J. (1999). **Probabilistic Networks and Expert Systems**. Springer-Verlag.

DASARATHY, B. (1980). Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** v.2, n.1, p.67-71.

DASGUPTA, A.; RAFTERY, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. **American Statistical Association**, v.93, p.294-302.

DAVISON (2006). Department of Biology, Davidson College. Disponível em: <http://www.bio.davidson.edu/COURSES/genomics/method/RFLP.html>. Acesso em: 5 nov. 2006

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. **Journal of the Royal Statistical Society,** v.7, n.39, p.1–38.

DUDA, R. O.; HART, P. E. (1973). **Pattern Classification and Scene Analysis**. John Wiley and Sons.

DUDA, R. O.; HART, P. E.; STORK, D. G. (2001). **Pattern Classification**. John Wiley, New York, 2 edition, 2001

ERUHIMOV, E.; MURPHY, K.; BRADSKI, G. (2003). Intel's open-source probabilistic networks library, Disponível em: < http://www.intel.com/research/mrl/pnl/pnl.pdf>. Acesso em: 15 jul. 2007.

EVERITT, B. S.; LANDAU, S., LEESE, M. (2001). **Cluster Analysis**. Oxford University Press Inc., New York, 4 edition, 2001.

FRALEY C.; RAFTERY A. E. (1998). How many clusters? Which clustering method? answers via model-based cluster analysis. **The Computer Journal**, v.41,n.8,p.578–588.

FRIEDMAN, N.; GOLDSZMIDT, M. (1996). Building classifiers using Bayesian Networks. **Proceedings of the Thirteenth National Conference on Artificial Intelligence**, p.1277-1284. AAAI Press.

GARRITY, G. M.; HOLT, J. G. (2001). The road map to the Manual. In: GARRITY, G. M.; BOONE, D. R.; CASTENHOLZ, R. W. eds. **Bergeys's manual of systematic bacteriology**. v.1, 2ed, New York, The Williams & Wilkins, 2001. p.119-154

GEIGER, D.; HECKERMAN, D. (1994). **Learning Gaussian Networks**. Technical Report MSR-TR-94-10, Microsoft Research.

GERMANO M. G.; MENNA, P.; MOSTASSO, F. L.; HUNGRIA, M. (2006). RFLP analysis of the RNA operon of a Brazilian collection of Bradyrhizobial strains from 33

legume species. **International Journal of Systematic and Evolutionary Microbiology**, v.56, n.1, pp.217 – 229

HELLER, K.A.; GHAHRAMANI, Z. (2005). Bayesian Hierarchical Clustering. **Proceedings of the 22nd International Conference on Machine Learning**, Bonn, Germany, p.297--304.

HUNGRIA, M.; VARGAS, M. A. T.; ARAÚJO, R. S. (1997). Fixação biológica do nitrogênio em feijoeiro. In: Vargas, M. A. T.; Hungria, M., eds. **Biologia dos Solos dos Cerrados**. Planaltina, EMBRAPA-CPAC, 1997. p.189-225.

IWAYAMA, M.; TOKUNAGA, T. (1995). Hierarchical Bayesian Clustering for automatic text classification. **Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence IJCAI-95**, v.2, p.1322-7.

JAIN, A. K.; DUBES, R. C. (1988). **Algorithms for Clustering Data**. Prentice-Hall Advanced Reference Series. Prentice-Hall, Inc., Upper Saddle River, NJ.

JENSEN, F. V.; LAURITZEN, S. L.; OLESEN, K. G. (1990a). Bayesian updating in Causal Probabilistic Networks by Local Computations. **Computational Statistics Quarterly**, v.5, n.4, p.269-282.

JOHNSON, R. A.; WICHERN, D. W. (1992). **Applied Multivariate Statistical Analysis**. 4th ed. New Jersey: Prentice Hall.

KAUFMAN, L.; ROUSSEEUW, P. J. (1990). **Finding Groups in Data: an Introduction to Cluster Analysis**. Wiley, New York.

KJÆRUL. U. (1990) **Triangulation of Graphs - Algorithms Giving Small Total State Space**. Technical Report R-90-09. Department of Mathematics and Computer Science. Aalborg University, Denmark, 1990.

KRIEG, N. R.; HOLT, J. G. (1984). **Bergeys's Manual of Systematic Bacteriology**. New York: The Williamas & Wilkins. p.235-244

KRUSKAL, J.; WISH, M. (1978). **Multidimensional Scaling.** Beverly Hills, CA: Sage Publications, 1978.

LAURITZEN, S. L.; WERMUTH, N. (1989). Graphical Models for Associations between Variables, Some of which are Qualitative and some Quantitative. **Annals of Statistics**, v.17, p.31-57.

LAURITZEN, S. L. (1992). Propagation of probabilities, means and variances in Mixed Graphical Association Models. **Journal of the American Statistical Association**, v.87, n.420, p.1098-1108.

LAURITZEN, S. L. (1995). The EM algorithm for graphical association models with missing data. **Computational Statistics and Data Analysis**, v.19, p.191–201.

LAURITZEN, S. L. (1996). **Graphical Models**. Clarendon Press.

MCMICHAEL, D.; LIN L.; PAN, H. (1999). Estimating the parameters of mixed Bayesian networks from incomplete data. **Proceedings of Information Decision and Control 99**, Adelaide, Australia, February 1999.

MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C. (1994). **Machine Learning, Neural and Statistical Classification**. Ellis Horwood Publishers.

MURPHY, K. (2001). The Bayes Net Toolbox for Matlab. **Computing Science and Statistics**, v.33.

MURTAGH, F.; RAFTERY, A. E. (1984). Fitting Straight Lines to Point Patterns. **Pattern Recognition**, v.17, n.5, p.479-483.

OOYEN, A. V. (2001). Theorical Aspects of Patterns Analysis. In: L. Dijkshoorn, K.J. Tower. **New Aproaches for the Generation and Analysis of Microbial Fingerprint**, Amsterdam, Elsevier, p.31-45.

PASKIN, M. A. (2003). A Short Course on Graphical Models, Disponível em: <http://ai.stanford.edu/~paskin/gm-short-course/>. Acesso em: 5 may. 2007.

PAZZANI, M. J. (1996a). Constructive Induction of Cartesian Product Attributes. **Proceedings of the Conference ISIS'96: Information, Statistics and Induction in Science**, p.66-77.

PAZZANI, M. J. (1996b). Searching for Dependencies in Bayesian Classifiers. **Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics**, p.239-248.

PEARL, J. (1988). **Probabilistic Reasoning in Intelligent Systems**. Morgan Kaufmann Publishers.

PEÑA, J. M. (2001). **On Unsupervised Learning of Bayesian Networks and Conditional Gaussian Networks**. PhD Thesis. Department of Computer Science. University of the Basque Country. Donostia – San Sebastián, July 2001.

RISSANEN, J. (1989). **Stochastic Complexity in Statistical Inquiry**. World Scientific Publishing Company, New Jersey.

SCHWARZ, G. (1978). Estimating the Dimension of a Model. **Annals of Statistics**, v.6, n.1, p.461–464.

VILLANUEVA, E. R.; CHRIST, R. E.; MACIEL, C. D. (2007). Classificador Bayesiano Hierárquico Utilizando Modelos Gaussianos. **Anais do VIII Simpósio Brasileiro de Automação Inteligente - SBAI 2007**. Florianópolis, 2007. No prelo.

WANG, E. T.; VAN BERKUM, P.; SUI, X. H.; BEYENE, D.; CHEN, W. X.; MARTINEZ-ROMERO, E. (1999). Diversity of Rhizobia associated with Amorpha Fructicosa from Chinese soils and description of Mesorhizobium Amorphae sp,. **International Journal of Systematic Bacteriology**, Washington, v.49, p.51-65.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo