Uma Aplicação de Redes Bayesianas no Auxílio à Tomada de Decisões Médicas

Luiz Gonzaga de Queiroz Silveira Júnior

Dissertação de Mestrado submetida à Coordenação dos Cursos de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Mestre no domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação - Comunicações

Prof. Francisco Marcos de Assis, Dr.
Orientador

Campina Grande, Paraíba, Brasil

S587a Silveira Junior, Luiz Gonzaga de Queiroz
Uma aplicacao de redes bayesianas no auxilio a tomada de decisoes medicas / Luiz Gonzaga de Queiroz Silveira Junior. - Campina Grande, 2003.
156 f.

Dissertacao (Mestrado em Engenharia Eletrica) - Universidade Federal de Campina Grande, Centro de Ciencias e Tecnologia.

1. Inteligencia Artificial 2. Redes Bayesianas 3. Inferencia Probabilistica 4. Diagnostico Medico 5. Dissertacao I. Assis, Francisco Marcos de II. Universidade Federal de Campina Grande - Campina Grande (PB) III. Título

CDU 004.89:621.39(043)

UMA APLICAÇÃO DE REDES BAYESIANAS NO AUXÍLIO À TOMADA DE DECISÕES MÉDICAS

LUIZ GONZAGA DE QUEIROZ SILVEIRA JÚNIOR

Dissertação Aprovada em 17.12.2003

FRANCISCO MARCOS DE ASSIS, Dr., UFCG
Orientador

ANTONIO MARCUS NOGUEIRA LIMA, Dr., UFCG
Componente da Banca

BENEMAR ALENCAR DE SOUZA, D.Sc., UFCG
Componente da Banca

JOÃO MARQUES DE CARVALHO, Ph.D., UFCG
Componente da Banca

JOVANY LUIS ALVES DE MEDEIROS, Dr., UFPB
Componente da Banca

CAMPINA GRANDE - PB
Dezembro - 2003

Dedicatória

Esta dissertação é dedicada aos meus pais pelo incondicional amor e total dedicação, fundamentais na minha vida.

Agradecimentos

- À Deus, por tudo;
- Aos meus pais Luiz Gonzaga e Maria de Jesus, pelo incondicional amor, pela infinita dedicação e compreensão em todos os momentos;
- Aos meus irmãos Fernando, Felipe e Cecília, que sempre me incentivaram;
- À Aninha, pelo incentivo e carinho nesta jornada;
- Ao professor Francisco Marcos de Assis, pela orientação, amizade, incentivo e ensinamentos;
- Ao professor Edmar Candeia Gurjão, pela orientação, paciência, confiança e amizade, inestimáveis para a realização deste trabalho;
- Ao professor Jovany Medeiros, pela orientação, amizade e valorosas contribuições e sugestões;
- Aos amigos da graduação e da pós-graduação Emmanuel, Murali, Natasha, Danilo, Darlan, Pierre, Emanoel, Netto, Suzete, Leonardo, Karina, Valnyr, Waslon, José Antônio, José Carlos, Rex, Edmar, Protásio e Denis, pela amizade, lealdade e momentos de descontração;
- Às acadêmicas de medicina Ligia Bezerra e Márcia Colares pela amizade, valorosas contribuições e incentivos;
- Aos professores Romulo Valle, Ubirajara, Benemar, José Ewerton Farias, Bruno Albert e Antonio Marcus, pelos ensinamentos e amizade;
- Aos amigos do LTI Pedro, Leandro e Marcos Morais, do LAPS Rinaldo e Daniel, do LAT Fabiano e Diana, da UNICAP Francisco Madeiro e da PUC-PR Luiz A. Neves, pela amizade, sugestões e contribuições ao longo do trabalho;
- Aos demais professores do DEE-UFCG;
- A todos os funcionários do DEE-UFCG, em especial a Ângela, Rosilda, Eleonora e Pedro, pela amizade;
- Ao CNPQ, pelo importante apoio financeiro.

Resumo

O Diagnóstico Médico se insere numa categoria ampla de problemas, onde a tomada de decisão é realizada considerando-se as evidências conhecidas e estas apresentando diferentes níveis de confiança. Além disso, não é rara a ocorrência de diferentes patologias com sintomas em comum. A particularidade deste cenário está presente na Neurologia, onde patologias raras com sintomas semelhantes tornam a emissão do diagnóstico diferencial difícil e até mesmo imprecisa.

Com o objetivo de reduzir o grau de incertezas envolvido, utiliza-se geralmente ferramentas de aquisição de novas evidências, como exames clínicos e neurológicos. Acontece que esta obtenção geralmente não reduz a complexidade da emissão de um diagnóstico diferencial. Para isto, é necessário dispor de ferramentas computacionais que auxiliem na tomada de decisão.

Neste trabalho é avaliado o desempenho da inferência probabilística em Redes Bayesianas no auxílio à tomada de decisões médicas. O desempenho dessa técnica é analisado sob bases de dados com diferentes cenários considerados. Além disso, são propostos novos algoritmos de redução da complexidade computacional da inferência probabilística, os quais se baseiam em conceitos da Teoria da Informação. Os resultados obtidos mostram que a inferência probabilística em Redes Bayesianas pode ser promissora no auxílio à emissão do diagnóstico médico.

Abstract

Medical Diagnosis belongs to a wide category of problems, where decision making is accomplished considering the known evidences with different trust levels. Moreover, it is not rare the occurrence of different pathologies with common symptoms. The particularity of this scenario is present in Neurology, where rare pathologies with similar symptoms make differential diagnosis difficult and even imprecise.

With the objective of reducing the degree of uncertainties involved, acquisition tools of new evidences are usually used, like clinical and neurological exams. However, these new evidences usually do not reduce the complexity of the emission of one differential diagnosis. Thus, it is necessary to make use of computational tools that help decision making.

In this work the performance of probabilistic inference in Bayesian Networks is evaluated as an aid to medical decisions. The performance of this technique is evaluated under databases with different scenarios. Moreover, new algorithms for reduction of the computational complexity of the probabilistic inference are considered, which use concepts of Information Theory. The results show that the probabilistic inference in Bayesian Networks can be promising as an aid to medical diagnosis.

Lista de Símbolos e Abreviaturas

DAG - Grafo Acíclico Orientado (*Directed Acyclic Graph*)

MDL - Mínima Descrição de Comprimento (*Minimum Description Length*)

FPM - Função Produto Marginalizada

X, Y, Z - Variável aleatória ou nó em uma rede bayesiana.

x, y, z - Valores assumidos por variáveis aleatórias.

$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ - Conjunto de Variáveis Aleatórias.

$\mathbf{x}, \mathbf{y}, \mathbf{z}$ - Conjunto de Valores assumidos por $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, respectivamente.

A, B, C - Eventos.

Θ - Parâmetro populacional estimado.

θ - Valor do parâmetro populacional estimado.

$P(A)$ - Probabilidade de um evento.

$P(A|B)$ - Probabilidade condicional.

$P(X = x)$,$p(x)$ - Probabilidade da variável aleatória X assumir o valor x.

$f(x)$ - Função Distribuição de Probabilidade de uma Variável Aleatória X.

$f(x; a_1, a_2, ..., a_n)$ - Função Distribuição de Probabilidade de uma Variável Aleatória X baseada no conjunto de parâmetros $a_1, ..., a_n$.

$f(x, y)$ - Distribuição conjunta de X e Y.

$f(x|y)$ - Distribuição marginal de X em relação a Y.

$E_{f(x)}[X]$ - Valor esperado de X em ralação à distribuição $f(x)$.

$\lambda(X)$ - Vetor λ do nó X.

$\lambda_Y(X)$ - Mensagem λ que o nó Y envia para X.

$\pi(X)$ - Vetor π do nó X.

$\pi_Y(X)$ - Mensagem π que o nó Y envia para X.

$\lambda_Y(X)$ - Mensagem λ que o nó Y envia para X.

Pa_i - Conjunto de pais do nó X_i.

pa_{ii} - Uma instância do conjunto de pais do nó X_i.

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

Nos últimos anos, vem crescendo o número de sistemas computacionais que utilizam redes bayesianas para auxiliar a tomada de decisões. Aplicações correntes incluem diagnóstico médico [5] e visão computacional [28]. Neste trabalho, as redes bayesianas são utilizadas no desenvolvimento de um sistema computacional de auxílio à emissão do diagnóstico médico de patologias neurológicas raras e que apresentam sintomas em comum. A complexidade presente neste cenário torna a emissão do diagnóstico diferenciado por parte dos não especialistas do domínio bastante difícil, não sendo rara a ocorrência de diagnósticos errados.

As principais contribuições deste trabalho são:

1. Desenvolvimento de um sistema computacional de auxílio à tomada de decisão médica, o qual reduz significativamente a complexidade do cenário envolvido na tomada de decisão, assegurando ao usuário respostas consistentes;

2. Desenvolvimento de métodos de supressão de arcos baseados em conceitos da Teoria da Informação, os quais apresentam resultados bastante satisfatórios e, portanto, podem ser usados como novas abordagens no desenvolvimento de aplicações práticas baseadas em redes bayesianas;

3. Desenvolvimento de métodos de ordenação de variáveis aleatórias baseados em medidas da Teoria da Informação, os quais proporcionaram um aumento de eficácia na realização de inferências em redes bayesianas;

4. Formação da única base de casos clínicos de Miastenia Gravis conhecida.

Na próxima seção são apresentados as abordagens empregadas no desenvolvimento de sistemas de auxílio à tomada de decisão. Contudo, caso o leitor deseje verificar logo as propostas e os testes dos métodos de supressão de arcos apresentados no Capítulo 2, poderá fazê-lo sem perda de continuidade ou de fundamentação.

1.1 Abordagens Usadas no Desenvolvimento de Sistemas de Auxílio à Tomada de Decisão

Na década de 60, os primeiros sistemas computacionais para dar suporte à tomada de decisão médica começaram a surgir, sendo esses sistemas baseados no tratamento da incerteza segundo a teoria probabilística de Bayes. Em especial os sistemas para diagnóstico requeriam que os conjuntos de possíveis doenças a diagnosticar fossem mutuamente exclusivos, com casos clínicos plenamente coletados. As evidências (sintomas observáveis) eram assumidas condicionalmente independentes entre si dada qualquer doença possível, podendo cada paciente possuir apenas uma única patologia [29]. Os sistemas resultantes eram desenvolvidos para domínios com pequeno número de hipóteses e de evidência limitada. O interesse no uso de probabilidades diminuiu em parte, devido à percepção da época de que esta era uma técnica intratável e inadequada para expressar a estrutura do conhecimento humano [22], e em parte, porque em domínios maiores as simplificações adotadas em geral produziam resultados matematicamente incorretos, além de não existirem mecanismos de explicação para os não especialistas do domínio.

Nos anos 70, surgem os sistemas especialistas. Eles utilizavam uma linguagem para representar o conhecimento do especialista em uma forma análoga aos predicados lógicos, isto é, coerente com a aplicação, empregando algum método de raciocínio heurístico. Para manipular incerteza utilizavam métodos *ad hoc* derivados da teoria de probabilidade, mas em desacordo com os axiomas da probabilidade. Eles podem manipular com êxito domínios de problemas maiores e mais complexos do que os anteriores, além de proverem facilidades de explicação, permitindo o seu uso por não especialistas. O MYCIN [21] é um exemplo de sistema especialista que obteve grande sucesso.

Contudo, no final dos anos 80, houve uma retomada no interesse por abordagens probabilísticas, sendo motivada pela descoberta de que se for considerado o relacionamento causal e a independência condicional entre variáveis do domínio, é necessário re-

presentar apenas probabilidades condicionais entre variáveis diretamente dependentes. Essa retomada está associada ao aparecimento de modelos baseados em representações gráficas de dependências probabilísticas conhecidas como redes probabilísticas. O uso dessas redes apresenta as seguintes vantagens com relação às abordagens anteriores: a) permite representar e manipular a incerteza com base em princípios matemáticos fundamentados e b) modela o conhecimento do especialista do domínio de uma forma intuitiva.

Um exemplo de rede probabilística é a rede bayesiana. A década de 90 marcou o início da maior parte das pesquisas realizadas a respeito das redes bayesianas. Desde essa época, essas redes vêm sendo utilizadas para a solução de vários tipos de problemas em diversas áreas, como por exemplo, diagnóstico médico [23] e visão computacional [28]. A Figura 1.1 ilustra um exemplo de rede bayesiana formada por cinco nós.

Uma rede bayesiana é uma grafo orientado acíclico, no qual os nós representam variáveis aleatórias e um arco unindo dois nós representa a dependência probabilística entre as variáveis associadas. Cada nó armazena a função de distribuição de probabilidades condicional dos valores que podem ser assumidos pela variável aleatória associado a ele, dado os valores de seus nós pais (isto é, àqueles diretamente ligados ao nó em questão). Por meio de um formalismo gráfico, uma rede bayesiana permite tanto representar a função distribuição de probabilidades conjunta (DPC), quanto dependências probabilísticas entre um conjunto de variáveis aleatórias. A DPC resume o conhecimento sobre um domínio específico, sendo a característica principal a possibilidade para reduzir o cálculo da DPC global a uma séria de cálculos locais, utilizando somente a incerteza local com base no conhecimento dos estados das variáveis adjacentes.

O raciocínio probabilístico em redes bayesianas, a ser utilizado como suporte a tomada de decisões, é baseado na realização de inferências probabilísticas, isto é, cálculos da probabilidade de um evento, dadas todas as evidências disponíveis. Esse cálculo é baseado em probabilidade condicional e no teorema de Bayes. A probabilidade condicional é vista como uma medida de crença no evento dada todas as evidências conhecidas. O raciocínio probabilístico em redes bayesianas é, então, entendido como um processo de atualização de crenças. As inferências podem ser do tipo causal, partindo das causas para relacionar os efeitos; de diagnósticos, dos efeitos parte-se para as causas; intercausal, quando diversas causas contribuem para um efeito comum; e misto, quando há uma combinação de dois ou mais tipos acima.

A independência condicional formaliza a noção qualitativa de irrelevância. A no-

tação **I**(A,B,C) representa que A é independente de C, dado B, ou seja, o conhecimento de C não afeta a crença sobre A, quando já se conhece B. A independência condicional e a dependência fornecem a base para expressar a noção direta e qualitativa de relevância na forma gráfica de rede bayesiana, antes de fazer-se qualquer atribuição numérica de probabilidade.

A representação gráfica de independência condicional proposta por Pearl é relevante, mas a idéia de expressar informação probabilística através de representações gráficas foi introduzida inicialmente pelo geneticista *Sewal Wright* (1921) que utilizou um modelo causal (grafo acíclico orientado, DAG) "como uma ajuda na análise biométrica de certas classes de dados", o qual deu origem às redes bayesianas. Devemos a Pearl a introdução do uso de redes bayesianas em sistemas inteligentes [33], através do artigo, de enorme repercussão, intitulado *Fusion, Propagation, and Structuring in Belief Networks* [39].

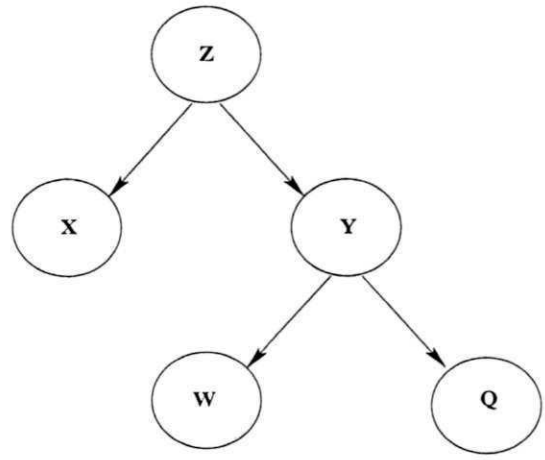


Figura 1.1: Rede bayesiana com 5 nós com topologia em árvore

Neste trabalho é seguida a abordagem adotada na década de 90 no desenvolvimento de sistemas de auxílio à tomada de decisão. Assim, as redes bayesianas são utilizadas na realização de inferências de base probabilística, as quais irão auxiliar à emissão do diagnóstico médico de patologias neuromusculares raras.

1.2 Organização do trabalho

Esta dissertação está organizada como segue:

No Capítulo 2 é apresentada uma caracterização das redes bayesianas. Apresenta-se os principais algoritmos de aprendizagem de redes bayesianas a partir de bases de dados. Posteriormente, estes algoritmos são avaliados e suas principais propriedades apresentadas. Ainda no Capítulo 2, alguns princípios da Teoria da Informação empregados em redes bayesianas são apresentados. Em seguida, são propostos algoritmos para a simplificação da rede, os quais se baseiam também em conceitos da Teoria da Informação. O final do Capítulo 2 é dedicado a uma análise das redes aprendidas a partir de diferentes bases de dados e que tiveram seus arcos removidos de acordo com os algoritmos propostos.

O Capítulo 3 apresenta os principais algoritmos utilizados na realização de inferências em redes probabilísticas. Este Capítulo também apresenta a Lei Distributiva Generalizada e ilustra que uma importante técnica de realização de inferências em Redes Bayesianas é um caso particular desta Lei. O final do capítulo 3 é dedicado a avaliação de desempenho da inferência probabilística nas redes que tiveram os arcos removidos de acordo com os algoritmos propostos para remoção de arcos e os efeitos da ordenação de variáveis, de acordo com medidas de informação, na eficácia do raciocínio probabilístico.

O Capítulo 4 descreve as patologias neurológicas sob foco e apresenta as características da base de casos clínicos formada para avaliar o desempenho da inferência probabilística em redes bayesianas no auxílio à emissão de diagnóstico médico.

No Capítulo 5 é apresentado o sistema desenvolvido para auxiliar à emissão do diagnóstico diferenciado.

O trabalho termina com o Capítulo 6, um espaço dedicado às considerações finais e às perspectivas para trabalhos futuros.

Capítulo 2

Redes Bayesianas

2.1 Introdução

Durante as últimas duas décadas, o interesse no uso de técnicas exatas de inferência probabilística para o desenvolvimento de sistemas de auxílio à tomada de decisão têm chamado bastante a atenção de pesquisadores em universidades e empresas. No entanto, o uso de tais técnicas de inferência pressupõe a disponibilidade de um modelo de conhecimento válido. Na ausência deste modelo, surge a necessidade de se desenvolver técnicas capazes de combinar o conhecimento especialista com os dados, obtidos a partir de uma base de dados, de forma a serem usados no desenvolvimento e, mais adiante, no refinamento desses modelos de conhecimento. Esta abordagem apresenta algumas dificuldades como, por exemplo, a impossibilidade, em algumas vezes, de se contar com a presença de um especialista do domínio do conhecimento. Sendo assim, a obtenção do modelo diretamente a partir da base de dados se faz necessária e permite uma aprendizagem automática, livre da supervisão de especialistas e, portanto, agrega um interesse maior.

Uma aproximação promissora a este problema é tentar construir automaticamente, ou em outras palavras aprender, uma rede probabilística que represente ou modele o conhecimento humano embutido nos dados. Especificada uma rede, é possível utilizar-se dela como um modelo probabilístico usual do domínio do problema.

Na tentativa de representar o conhecimento humano, a motivação para utilização de redes probabilísticas é que os seres humanos têm uma certa facilidade em moldar os fatos e fenômenos em forma de relacionamentos causais. Isto é, o que se busca é uma forma adequada de modelar o conhecimento humano, com respeito a um determinado

domínio do problema, no qual coloca-se a questão de realizar prognósticos a respeito de causas e a partir de observações de indicadores probabilísticos relacionados com estas causas. Neste sentido, as redes probabilísticas, ou bayesianas constituem um modelo do domínio do problema e não um modelo de raciocínio; como é comum quando se utiliza outros esquemas como, por exemplo, Redes Neurais. Devido a isto, a abordagem Bayesiana já tornou-se um importante paradigma para a representação e tratamento com incertezas.

2.2 Estrutura das Redes Bayesianas

Um ponto de partida adequado para se visualizar a montagem de uma representação eficiente é considerar com atenção o mecanismo de inferência humano, entendido como o meio pelo qual as pessoas integram dados de múltiplas fontes e geram uma interpretação coerente destes dados. A análise deste mecanismo indica que a maneira como o ser humano armazena e utiliza o seu conhecimento a respeito de um determinado domínio não fica bem caracterizada por uma distribuição conjunta mas sim através de distribuições marginais e distribuições condicionais envolvendo pequenos agrupamentos de variáveis [5].

Sendo assim, o conhecimento humano acerca de um problema de diagnóstico pode ser adequadamente modelado utilizando um grafo no qual os nós estão associados às variáveis aleatórias envolvidas e cada nó fica ligado a outros que julgamos serem diretamente relacionados.

Escolhendo uma ordenação arbitrária `ord` das variáveis $X_1, ..., X_n$, considera-se a fatoração de sua distribuição de probabilidade conjunta pela regra da cadeia, da forma $P(X_1, X_2, ..., X_n) = P(X_n \mid X_{n-1}, ..., X_1)...P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$, em que cada fator contém somente uma variável no lado esquerdo da barra de condicionamento.

Essa fatoração pode ser usada como uma prescrição para quantificar consistentemente as dependências entre os nós de um grafo arbitrário. Define-se Pa_i como o conjunto de variáveis que influenciam diretamente x_i (i.e. $P(X_i \mid Pa_i) = P(X_i \mid X_{i-1}, X_{i-2}, ..., X_1)$, as quais serão doravante denominadas **pais** da variável aleatória X_i.

Representando $P(X_i \mid Pa_i)$ por uma função $F_i(X_i, Pa_i)$ tal que

$$\sum_{X_i} F_i(X_i, Pa_i) = 1, \quad 0 \leq F_i(X_i, Pa_i) \leq 1,$$

então a fatoração acima pode ser reescrita como

$$P(X_1, X_2, ..., X_n) = \prod_{X_i} F_i(X_i, Pa_i).$$

A distribuição $P(X_1, X_2, ..., X_n)$, junto com a ordenação `ord` identifica univocamente um conjunto de nós pais para cada variável aleatória X_i, que pode ser associada a um grafo direcionado acíclico que representa as relações de dependência probabilística entre as variáveis $X_1, X_2, ..., X_n$. Este grafo e a especificação das probabilidades condicionais entre cada nó e seus pais constituem uma rede bayesiana (RB). Segue-se uma definição formal.

Definição 1 [3] *Seja U um conjunto de variáveis $\{X_1, ..., X_n\}$, $n \geq 1$. Cada variável $X_i \in U$ assume um valor do conjunto $\{x_{i_1}, ..., x_{i_r}\}$, $r \geq 1$, $i = 1, ..., n$. Uma* **rede bayesiana** *B sobre U é um par $B = (B_S, B_P)$ sendo B_S a estrutura da rede que consiste de um grafo direcionado e acíclico, com um nó para cada variável em U; B_P é um conjunto de probabilidades condicionais associadas a B_S.Para cada variável $X_i \in U$, B_P contém as probabilidades condicionadas $P(x_{i_j} \mid Pa_i)$ que consiste nas probabilidades do nó que representa a variável X_i estar no estado x_{i_j} dados valores das variáveis no* **conjunto de pais** *Pa_i de X_i.*

Deve-se notar que a especificação completa de uma rede bayesiana é determinada pela ordenação `ord` das variáveis $X_i \in U$ e pela função de distribuição de probabilidade conjunta $P(X_1, \ldots, X_n)$. A topologia de uma rede bayesiana é fortemente sensível à ordenação `ord`, sendo possível que um grafo em árvore se transforme em um grafo completo pela modificação da ordenação das variáveis. A sensibilidade à ordenação se deve à relação entre `causalidade` e a estrutura da rede, ou seja, modificar a ordenação significa modificar relações de causalidade, ainda que a medida global de probabilidade seja a mesma [41].

Em redes bayesianas, adota-se um critério de separabilidade mais elaborado do que o que o senso comum recomenda (ou seja, considerar separação apenas como ausência de ligação entre dois nós). Para facilitar o entendimento, considere os três tipos distintos de conexões entre nós ilustrados na Figura 2.1, denominados por Charniak [20] como linear, convergente e divergente. O conceito de separabilidade em redes bayesianas, proposto por Pearl [39] está intrinsicamente relacionado com as propriedades das conexões entre os nós e é parte fundamental na identificação das relações de independência estatística. Para estabelecer esse critério considere as definições:

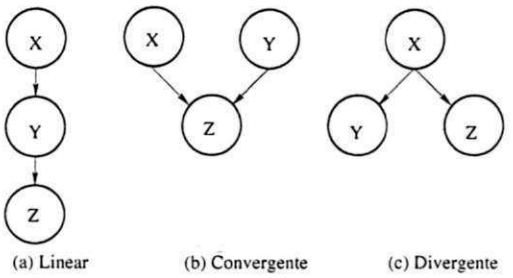


Figura 2.1: Tipos Diferentes de Conexões

Definição 2 (Evidência) [37]

- *Evidência em uma rede bayesiana é um conjunto de nós cujos valores são conhecidos*

Definição 3 (Caminho d-conectado) [37]

- *Dado uma evidência, **E**, o caminho entre dois nós X e Y é dito ser d-conectado (conexão de dependência) em relação a **E** se uma das seguintes condições se verificarem:*

  1. *O caminho entre X e Y é linear ou divergente e não possui nós em **E**.*
  2. *O caminho entre X e Y é convergente e os nós no interior do caminho ou um de seus descendentes está contido em **E**.*

Definição 4 (Nós d-separados) [37]

- *Dado uma evidência, dois nós são ditos ser d-separados (separação de dependência) se não existir nenhum caminho d-conectado que os una.*

Definição 5 (Independência entre nós) [37]

- *Dado uma evidência, **E**, diz-se que dois nós são independentes em relação a **E** se forem d-separados.*

Definição 6 (Grafo em Árvore) [10]

- *Um grafo em árvore é um grafo onde todos os nós, exceto aquele denominado de nó raiz, tem apenas um único pai.*

Definição 7 (Grafo em Árvore Múltipla) [39]

- *Um grafo em árvore múltipla é um grafo simplesmente conectado, isto é, um grafo onde não existe mais que um único caminho entre quaisquer dois nós.*

Os grafos cuja topologia é uma árvore múltipla (no inglês é utilizada a expressão *polytree*), podem ser vistos como uma coleção de várias árvores unidas. Sendo assim, podem existir nós com mais de um pai, ou diversos nós raízes. Contudo, para quaisquer dois nós pertencentes ao grafo, existe um único caminho que passa por eles.

Definição 8 (Grafo com Ciclo) [32]

- *Um grafo é dito ter ciclo se entre quaisquer dois nós existe mais de um caminho que os una.*

A Figura 2.2(a) ilustra um grafo em árvore. Na Figura 2.2(b) é mostrado um grafo em árvore múltipla, já a Figura 2.2(c) apresenta um grafo com ciclo.

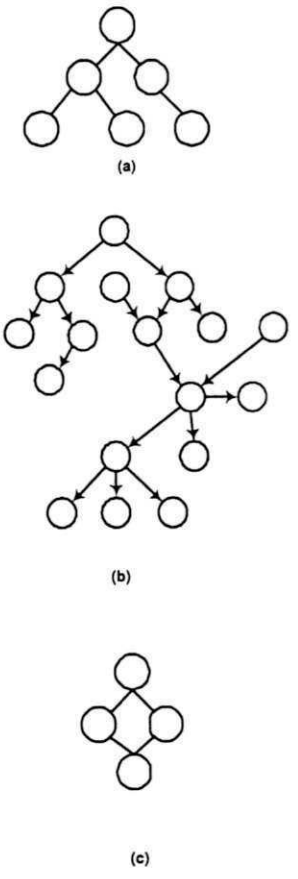


Figura 2.2: Grafos com topologias diferentes

Para verificar se uma rede é bayesiana, basta que uma única condição seja satisfeita: cada nó do grafo deve ser condicionalmente independente de todos os nós que não são seus descendentes, exceto seus pais [41]. Segundo Pearl [41], com base nesta condição, pode-se derivar um procedimento recursivo para a construção de uma rede bayesiana. Dada a distribuição de probabilidade conjunta $P(X_1, X_2, ..., X_n)$ e uma ordenação `ord` destas variáveis, inicia-se a construção do grafo escolhendo o nó raiz X_1 e atribuindo ele a probabilidade marginal $P(X_1)$. Em seguida, acrescenta-se mais um nó X_2 no grafo. Se X_2 for dependente de X_1 então traçamos um arco ligando ambas as variáveis com a seta apontando para X_2 e quantificando esta dependência ponderando o arco com a medida $P(X_2|X_1)$, caso contrário, mantém-se as variáveis desconectadas a atribui-se uma probabilidade à *priori* $P(X_2)$ a X_2. Repete-se este procedimento para as demais variáveis e, então, obtém-se uma rede bayesiana.

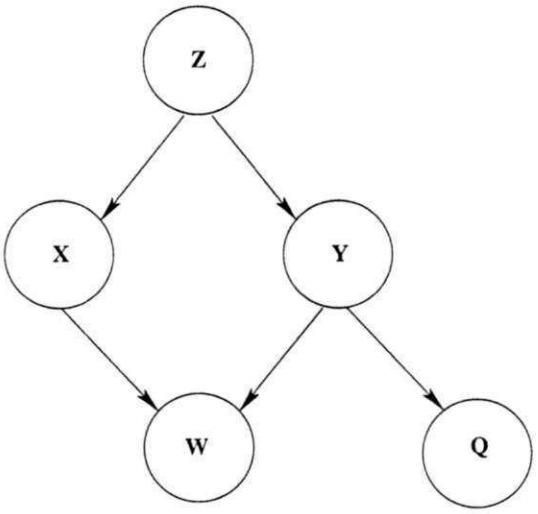


Figura 2.3: Exemplo de rede bayesiana mostrando um dos nós (W) com dois pais

A rede ilustrada pela Figura 2.3 representa a seguinte distribuição presente nos dados:

$$P(Z, X, Y, W, Q) = P(Q|Y)P(W|X, Y)P(Y|Z)P(X|Z)P(Z). \quad (2.1)$$

As relações de independência apresentadas pela rede bayesiana reduzem o esforço no cálculo da distribuição de probabilidade conjunta. A ordenação das variáveis pode

ser feita com base no conhecimento humano para identificar as relações de dependência entre as variáveis, estabelecendo para isto, relações de causalidade entre estas e, para aquelas que influenciam diretamente outra, é necessário definir pai e filho, respectivamente.

Nem sempre se conhece as relações de causalidade que estabelecem a estrutura da rede (B_S), nem o conjunto de probabilidades condicionais (B_P). Portanto, os principais problemas envolvendo redes bayesianas são:

1. Construir uma rede bayesiana a partir de uma base de dados. Isto é, aprender a estrutura, B_S, e identificar as probabilidades condicionais, B_P, existentes entre os elementos constituintes, a partir de uma base de dados com amostras das variáveis aleatórias do domínio da aplicação;

2. Construir um método de atualização de crenças (do inglês, *Belief*) sobre o estado das variáveis a partir de evidências apresentadas à rede.

A próxima seção avalia os aspectos envolvidos na aprendizagem automática de redes bayesianas a partir de bases de dados.

2.3 Aprendizado de Redes Bayesianas a Partir de Bases de Dados

O problema que se apresenta consiste em descobrir estruturas arbitrárias de rede a partir da base de dados sem a necessidade do conhecimento a *priori* da distribuição de probabilidade que modela esses dados. No caso de B_S, pode-se levar ao aprendizado de redes com múltiplas conexões que são mais expressivas do que redes de árvore, já que na prática fenômenos aleatórios podem ter mais de uma causa.

Apesar das redes de múltiplas conexões nos permitirem uma precisão maior nestes casos, as suas desvantagens são numerosas. A complexidade temporal relacionada com o aprendizado da estrutura dessas redes é alto. Em adição, essas redes também apresentam o problema da complexidade espacial, relacionada com o armazenamento dos parâmetros, que aumenta com a conectividade da rede [25]. Redes bayesianas com mais conexões entre seus nós requerem um armazenamento de mais parâmetros de probabilidade, com o número de parâmetros de probabilidade "armazenados" em cada nó aumentando exponencialmente com o número de arcos que chegam. Em adição

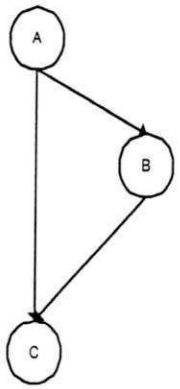
às vantagens computacionais, redes com baixa conectividade também possuem vantagens conceituais. A topologia de uma rede bayesiana expressa a informação sobre o entendimento causal e assegura relações probabilísticas do domínio. Redes com topologias muito simples são mais fáceis de compreender. Isto pode ser particularmente importante já que existe o desejo de explicar os resultados calculados utilizando a rede.

Portanto, é preciso haver algum tipo de compromisso entre precisão e usabilidade. Redes mais conectadas permitem obter modelos mais precisos, contudo, ao mesmo tempo, são computacionalmente e conceitualmente mais difíceis de usar. Redes mais simples, por sua vez, podem levar à perda de precisão no modelamento dos dados. Logo, deve-se adotar algum método que opte sempre por uma rede mais simples se esta rede é suficientemente precisa, mantendo a capacidade de escolher outra mais complexa se nenhuma rede mais simples possui a precisão desejada. O Exemplo 2.1 [50] ilustra este ponto.

Exemplo 2.1 [50] : Considere as duas redes ilustradas na Figura 2.4, onde todos os nós assumem valores binários. No grafo G1, o nó C tem dois pais, A e B; enquanto em G2, C tem somente um pai, B. G2 é a rede mais simples envolvendo os nós A,B e C para a dada relação de causalidade. Apesar disso, se forem examinados as probabilidades condicionais associadas a C, no grafo G1, descobre-se que o valor de C tem uma dependência maior com os valores de B e com pequena proporção do valor de A. Então, as relações de dependência da distribuição descrita por G1 são quase as mesmas daquelas obtidas pelo grafo G2. Portanto, estas duas redes bayesianas podem ser consideradas como estruturas aproximadamente equivalentes, muito embora tenham topologias diferentes.

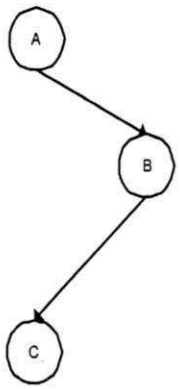
As redes bayesianas são comumente usadas em problemas envolvendo inferência probabilística, as quais gerenciam atualizações na distribuição de probabilidade a *posteriori* (DPP) quando alguns dos nós tornam-se instanciados à valores particulares. Considerando este aspecto, as duas redes podem ser consideradas como sendo aproximadamente equivalentes se estas exibem resultados semelhantes após a atualização na DPP. Admita agora que $p(a_1) = 0,3$. Após a execução da atualização da DPP, ambas as redes apresentam praticamente o mesmo resultado: $p(b_1) = 0,31$, em ambas, $p(c_1) = 0,356$ em G1, e $p(c_1) = 0,336$ em G2. Neste caso, há pouca perda de precisão no modelamento da distribuição básica utilizando a rede mais simples G2 ao invés de G1. $\Diamond$

A negociação envolvendo precisão e usabilidade pode ser feita utilizando um princípio



G_1

$p(a_1) = 0,5$ $p(a_0)=0,5$
$p(b_1 | a_1)= 0,8$ $p(b_0 | a_1) = 0,2$
$p(b_1 | a_0)= 0,1$ $p(b_0 | a_0) = 0,9$
$p(c_1 | a_1 , b_1) = 0,7$ $p(c_0 | a_1 , b_1) = 0,3$
$p(c_1 | a_0 , b_1) = 0,8$ $p(c_0 | a_0 , b_1) = 0,2$
$p(c_1 | a_1 , b_0) = 0,1$ $p(c_0 | a_1 , b_0) = 0,9$
$p(c_1 | a_0 , b_0) = 0,2$ $p(c_0 | a_0 , b_0) = 0,8$



G_2

$p(a_1) = 0,5$ $p(a_0)=0,5$
$p(b_1 | a_1)= 0,8$ $p(b_0 | a_1) = 0,2$
$p(b_1 | a_0)= 0,1$ $p(b_0 | a_0) = 0,9$

$p(c_1 | b_1) = 0,75$ $p(c_0 | b_1) = 0,25$
$p(c_1 | b_0) = 0,15$ $p(c_0 | b_0) = 0,85$

Figura 2.4: Redes Bayesianas Aproximadamente Equivalentes

bem conhecido, denominado de **Mínimo Comprimento de Descrição** (MDL, do inglês *Minimum Description Length*) de Rissanen [47]. O princípio MDL afirma que o melhor modelo de um conjunto de dados é aquele que minimiza a soma da codificação do modelo e da codificação dos dados dado o modelo. Contudo, tentar encontrar a rede (modelo) que minimiza a soma destas duas componentes é computacionalmente uma tarefa difícil, devido ao grande número de redes presentes no espaço de busca. Por exemplo, Robinson [45] utilizando o princípio de inclusão e exclusão mostrou que a quantidade de redes que pode ser formada por um conjunto de n nós, $G(n)$, é dada pela Equação (2.2).

$$G(n) = \left[\sum_{k=1}^{n}(-1)^{k+1}\begin{pmatrix} n \\ k \end{pmatrix}\right] G(n-1) \tag{2.2}$$

com $G(0) = 1$.

Assim, para um conjunto de 10 nós existe aproximadamente $4,2 \times 10^{18}$ diferentes estruturas possíveis. Devido ao grande número de combinações, o teste de todas as

possibilidades é impraticável e para tanto deve ser usado um algoritmo que elimine uma parte dessas estruturas.

A próxima seção descreve algumas das diferentes abordagens utilizadas no aprendizado de redes bayesianas.

2.3.1 Recuperação de Redes a Partir de Suposições Sobre a Distribuição dos Dados

O trabalho mais simples que pode ser visualizado como modelo de aprendizagem de rede foi o de Chow e Liu em 1968 [10]. Suas aproximações foram capazes de recuperar a DPP em redes apresentando topologias em árvore, a partir de registros em banco de dados. Se o banco de dados foi gerado por uma distribuição, cuja rede associada possui uma topologia em árvore, esta poderia ser recuperada com exatidão, desde que existam dados suficientes. Caso contrário, o método proposto garantia que a distribuição de probabilidades da rede aprendida pela estrutura em árvore seria a mais próxima de todas as redes em árvore para a distribuição dos dados. O critério adotado neste caso para a "proximidade" foi a medida de entropia cruzada de Kullback-Leibler [36]. A principal restrição a este trabalho era que o mesmo considerava no aprendizado apenas as redes com topologias em árvores. Logo, se os dados presentes na base foram obtidos segundo uma distribuição que não estivesse associada a uma estrutura em árvore, o aprendizado levaria a uma rede imprecisa. O trabalho subseqüente de Rebane e Pearl [43] permitiu extender os métodos de Chow e Liu para a recuperação de redes com a topologia de árvore múltipla. Se a distribuição básica tem uma rede associada com topologia de árvore múltipla, esta poderia ser recuperada com exatidão. Contudo, novamente, se os dados são oriundos de uma distribuição que não represente uma topologia de árvore múltipla, a estrutura aprendida seria bastante imprecisa.

Ambas as aproximações falham na recuperaração de uma classe de redes que apresentem múltiplas conexões, a qual é mais realística do ponto de vista esperado para as aplicações práticas. Isto acontece porque tais métodos compartilham a desvantagem de se realizar suposições sobre a distribuição básica. Acontece que na maioria das vezes não se conhece, a *priori*, a distribuição básica, a qual, conforme já salientado no começo deste capítulo, é possível obter a partir do conhecimento da estrutura da rede bayesiana aprendida, ou seja, o conhecimento da distribuição leva ao conhecimento da estrutura. Tais métodos poderiam então produzir modelos muito imprecisos se a distri-

buição básica não se encaixa nas categorias de distribuições com as quais tais métodos podem lidar. Apesar disto, estes trabalhos forneceram muita informação relevante ao aprendizado de redes bayesianas.

Uma aproximação alternativa bastante interessante que permite tratar com redes de múltiplas conexões é a que foi sugerida por Cooper e Herskovits em 1991 [8]. Esta aproximação tenta encontrar a rede mais provável utilizando uma *aproximação bayesiana*, ou seja, assume-se uma distribuição a *priori* sobre o espaço contendo todas as possíveis estruturas de rede. Desta forma, foi considerado que a distribuição a *priori* era uniforme. Lam e Baccus [50] sugerem que esta consideração não leva a escolha razoável, já que, neste caso, o método proposto por Cooper e Herskovits [8] sempre preferiria a rede com maior precisão, mesmo nas situações onde uma rede muito complexa é só ligeiramente mais precisa.

De forma a evitar que isto ocorra, pode-se adotar o princípio MDL. Uma forma de visualizar o princípio MDL é segundo uma aproximação bayesiana, na qual a distribuição a *priori* sobre os modelos é inversamente relacionada aos comprimentos de suas codificações, isto é, as suas complexidades. Por exemplo, seja $\mathcal{H} = \{H_1, H_2, ...\}$ um espaço de hipóteses a cerca da origem dos dados. Tendo em vista que não se conhece a lei de formação que originou os dados, D, a MDL é obtida por uma busca em $\mathcal{H}$, ou seja, procura-se obter de $\mathcal{H}$ o elemento mais provável dado o conjunto D, que matematicamente pode ser expresso maximizando $P(H|D)$, podendo ser expresso de acordo com o Teorema de Bayes, como:

$$P(H|D) = \frac{P(D,H)}{P(D)} = \frac{P(D|H)P(H)}{P(D)}. \tag{2.3}$$

Aplicando o negativo do logaritmo na base 2 em ambos os lados, tem-se:

$$-\log_2 P(H|D) = -\log_2 P(D|H) - \log_2 P(H) + \log_2 P(D). \tag{2.4}$$

sendo $P(D)$ constante quando se varia H, a maximização na Equação (2.3) pode ser feita minimizando-se apenas

$$-\log_2 P(D|H) - \log_2 P(H), \tag{2.5}$$

Em Teoria da Informação, a aplicação do negativo do logaritmo na base 2 é interpretada como sendo o tamanho médio da palavra-código, em bits, necessária para codificar o modelo, H, e os dados de posse do modelo,$D|H$.

Portanto, o princípio MDL guarda restrições quanto aos modelos de aprendizagens obtidos que são tão simples quanto o possível. Já que se H for muito simples esta poderia ser descrita ou codificada necessitando de poucos bits; contudo, não iria explicar satisfatoriamente os dados, levando a um maior erro na descrição dos mesmo. Por outro lado, sendo H complexa os dados seriam explicados de forma mais precisa, mas também necessitaria de um espaço maior para a codificação da hipótese ou modelo. A MDL, como será explicada a seguir é uma solução de compromisso entre precisão e complexidade.

Tanto o trabalho de Cooper e Herskovits [8], quanto o de Lam e Baccus [50] desenvolveram um método heurístico que busca uma restrição no conjunto de estruturas pertencentes ao espaço de busca. No primeiro, a busca termina selecionando a estrutura com a mais alta probabilidade a *posteriori* e, no segundo, a rede com o mínimo comprimento de descrição. Além disto, no primeiro é necessário uma ordenação prévia das variáveis, enquanto que no segundo tal ordenação não é necessária, o que constitui uma vantagem em situações onde não existe informação suficiente para promover tal ordenação. A ordenação constitui uma restrição no tamanho do espaço de busca, visto que, além de ser formado pelas variáveis que pertencem ao domínio do problema, este deverá guardar uma relação de causalidade entre estas, o que sem dúvida diminui o número de possíveis estruturas.

A próxima seção discute em detalhes o princípio do Mínimo Comprimento de Descrição de Rissanen (MDL).

2.3.2 O Princípio MDL

O mínimo comprimento de descrição (MDL, do inglês *Minimum Description Length*) é um formalismo bastante estudado em Teoria da Informação, sendo proposto na década de 70 por Rissanen do Centro de Pesquisas da IBM em Almaden [47]. Assim, a descrição de comprimento é uma medida, expressa na unidade de informação - bit , de um conjunto de dados. O princípio MDL é baseado na idéia que o melhor modelo de uma coleção de ítens é o que minimiza a soma do:

1. Comprimento da codificação do modelo e,

2. Comprimento da codificação dos dados dado o modelo.

Sendo ambos medidos em bits.

O próximo exemplo ilustra que o princípio MDL tenta encontrar um compromisso entre a codificação do modelo e a dos dados realizada utilizando-se o modelo.

Exemplo 2.2 (Polinômios) [50]: Considere que os ítens de dados consistem de n pontos no plano, cada ponto especificado por um par de coordenadas reais de precisão fixada, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Admita que é desejado encontrar uma função (modelo) que ajuste estes pontos. Neste sentido, um polinômio de grau n que passe precisamente através destes pontos necessitaria de $n+1$ números para especificar os coeficientes do polinômio (item 1). Contudo, para armazenar os dados dado o polinômio, seria necessário armazenar as coordenadas X, isto é, $x_1, x_2, ..., x_n$ (item 2). Como é possível observar, não é necessário armazenar as coordenadas Y, isto é, $y_1, y_2, ..., y_n$, já que cada y_i poderia ser precisamente calculado a partir do polinômio e da respectiva coordenada X, x_i. Logo, a soma dos comprimentos das descrições seria $2n + 1$ vezes o número de bits necessários para armazenar os números na precisão desejada.

Na tentativa de minimizar esta soma, considere que seja usado um polinômio de ordem inferior, dito de ordem k. A concordância com o item 1 exigiria apenas $k + 1$ números para armazenar as coordenadas. Já atendendo observação do item 2, novamente poderíamos armazenar os pontos de informação pela especificação das coordenadas X, o que implica em n números. Desta vez, contudo, não poderíamos garantir que o nosso polinômio se "ajustaria" precisamente aos dados devendo, portanto, existir algum erro δ_i entre o valor y avaliado em x_i e a coordenada Y do i-ésimo ponto y_i. Sendo assim, para codificar os pontos seria necessário armazenar estes fatores de erro conjuntamente com as coordenadas x. Entretanto, se o $\max(\delta_1, ..., \delta_n)$ é desprezível, então seria necessário menos bits para armazenar estes fatores de erro do que os números ordinários da codificação com o polinômio de ordem n. Portanto, poderá existir algum polinômio de grau $k < n$ que resulta na mínima descrição do comprimento.$\Diamond$

O princípio MDL pode ser explicado também de forma intuitiva. Para isto, suponha, por exemplo, que em um grande conjunto de dados observou-se uma lei de formação, o que leva a pressupor a existência de uma máquina de Turing que modelaria tal conjunto. Sendo assim, a codificação dos dados poderia ser feita através da codificação de um algoritmo (modelo) e de alguns elementos do conjunto, ditos elementos representativos (já que os demais podem ser obtidos a partir desta lei de formação e dos elementos

representativos deste conjunto). Por outro lado, admita a existência de um outro conjunto de dados com o mesmo número de elementos do anterior, mas que neste caso não foi observado qualquer padrão de regularidade, ou seja, não foi encontrado qualquer lei de formação que explique os dados. Desta forma, a ausência do algoritmo leva a uma complexidade da descrição, visto que esta implica na descrição de todos os elementos do conjunto. Naturalmente, a medida da descrição neste caso será maior do que aquela realizada de posse do algoritmo e dos elementos representativos.

A MDL associada ao problema de obtenção de uma rede bayesiana a partir dos dados, implica na codificação de duas componentes: a rede propriamente dita e a adequação dos dados à rede proposta. Estas duas partes serão discutidas separadamente nas próximas seções.

Codificando uma Rede Bayesiana

Uma rede bayesiana é formada de duas partes: a estrutura,B_S, e o conjunto de probabilidades, B_P. Portanto, para representar uma rede bayesiana particular, as seguintes informações são necessárias e suficientes:

- Uma lista dos pais de cada nó,
- Um conjunto de probabilidades condicionais associadas a cada um dos nós.

Ambas são requeridas para parametrizar a rede.

De forma a facilitar o entendimento, estas duas partes serão codificadas separadamente. Na codificação da estrutura, B_S, para cada nó, X_i, codifica-se a identificação dos elementos que formam Pa_i (em que Pa_i é o conjunto de pais de X_i em B_S), seguido da codificação das probabilidades condicionais envolvendo X_i e os nós em Pa_i. A codificação de cada um dos elementos de Pa_i é dada simplesmente por $k_i^{\{d\}}$, em que k_i corresponde à identificação do i-ésimo elemento de Pa_i e d ao número de bits usados na sua codificação.

Admita que existem n nós num domínio de um problema. Para um nó X_i com cardinalidade do conjunto de pais dado por $|Pa_i|$, são necessários $k_i^{\{\log_2 n\}}$ bits para listar cada um de seus pais. Naturalmente como $|Pa_i| < n$, d pode ser limitado superiormente por $\log_2 n$, implicando em:

$$k_i^{\{d\}} \leq k_i^{\{\log_2 n\}}. \tag{2.6}$$

Agora, avalia-se a codificação das probabilidades condicionais. Em uma rede bayesiana, uma probabilidade condicional é necessária para toda instanciação distinta dos nós pais e do próprio nó (exceto pelo fato que um das probabilidades condicionais pode ser calculada a partir de todas as outras, já que a soma tem que ser 1). Assim em uma rede bayesiana, as probabilidades condicionais formam um conjunto cuja quantidade de elementos dependem do número de instâncias de X_i, r_i, e de Pa_i, q_i. Assim, para cada nó X_i existe armazenada uma matriz de probabilidades condicionais ,$[p(x_{ir}|pa_{is})]$, a qual representa a probabilidade do nó X_i está no estado x_{ir} dado que seus pais são instanciados a pa_{is}.

$$[p(x_{ir}|pa_{is})] = \begin{bmatrix} p(x_{i1}|pa_{i1}), & p(x_{i1}|pa_{i2}), & \dots & p(x_{i1}|pa_{iq_i}) \\ p(x_{i2}|pa_{i1}), & p(x_{i2}|pa_{i2}), & \dots & p(x_{i2}|pa_{iq_i}) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_{ir_i}|pa_{i1}), & p(x_{ir_i}|pa_{i2}), & \dots & p(x_{ir_i}|pa_{iq_i}) \end{bmatrix}, \tag{2.7}$$

com $r = 1, 2, ..., r_i$ e $s = 1, 2, ..., q_i$

Para representar as probabilidades condicionais, o comprimento da codificação será o produto do número de bits requeridos para armazenar o valor numérico de cada probabilidade condicional pelo número total de probabilidades condicionais que são requeridas. Sendo assim, como cada nó X_l em Pa_i possui r_l diferentes instâncias, o número de elementos de Pa_i, q_i, é dado por $\prod_{X_l \in Pa_i} r_l$. Assim, em princípio seria necessário realizar a descrição de uma matriz com $r_i \times \prod_{X_l \in Pa_i} r_l$ elementos. Contudo, ocorre que como o somatório das probabilidades de $(x_{ik}|Pa_{ij})$ deve ser igual a unidade, então um elemento não precisa ser codificado a cada instância de Pa_i, já que poderia ser deduzido. Portanto, necessita-se codificar apenas:

$$d(r_i - 1) \prod_{X_l \in Pa_i} r_l, \tag{2.8}$$

em que d é o número de bits usados na codificação de $P(x_{ik}|Pa_{ij})$.

Por exemplo, se um nó pode assumir 5 valores distintos e tem 4 pais, cada um dos quais pode assumir 3 valores distintos, seriam necessários $3^4 \times (5 - 1)$ probabilidades condicionais.

Assim, a descrição mínima usada para codificar a rede será dada por:

$$\sum_{i=1}^{n}[k_i^{\{log_2(n)\}} + d(r_i - 1) \prod_{X_l \in Pa_i} r_l]. \tag{2.9}$$

Analisando a Equação (2.9), observa-se que redes altamente conectadas requerem codificações mais longas. Primeiro, para muitos nós a lista de seus pais tonar-se-á grande e, desta forma, a lista de probabilidades condicionais, a qual é necessária armazenar para cada um dos nós, também aumentará. Em adição, redes nas quais os nós possuem um grande número de valores e com os pais tendo um grande número de possíveis instâncias, requerem codificações ainda mais compridas. Desta forma, o princípio MDL tenderá a favorecer redes nas quais os nós tem o menor número de pais (isto é, redes que são pouco conectadas) e também redes nas quais os nós que assumem um grande número de valores não são pais de outros nós que possuam grande número de instâncias.

Codificando os dados utilizando o modelo

Neste momento, é preciso definir melhor a forma da base de dados. A tarefa consiste em aprender a distribuição conjunta de um conjunto de variáveis aleatórias $U = \{X_1, ..., X_n\}$. Cada variável X_i está associada a uma coleção de valores $\{x_{i1}, ..., x_{ir_i}\}$, os quais definem as instância que elas podem assumir. Toda escolha distinta de valores para todas as variáveis em U define um evento singular na distribuição conjunta, sendo atribuído a uma probabilidade particular por esta distribuição.

Por exemplo, admita que temos três variáveis aleatórias X_1, X_2 e X_3, com X_1 admitindo valores no conjunto $\{1, 2\}$, X_2 tendo como instâncias valores em $\{1, 2, 3\}$ e X_3, tendo $\{1, 2\}$ como valores possíveis. Logo, existem $2 \times 3 \times 2$ diferentes instanciações completas das variáveis. Cada uma destas é um evento singular na distribuição conjunta, tendo uma probabilidade particular de ocorrência. Por exemplo, o evento no qual $\{X_1 = 1, X_2 = 3, X_3 = 1\}$ é um destes eventos singulares.

Portanto, é possível admitir que as linhas na matriz que forma o banco de dados são eventos singulares. Isto é, cada linha especifica um valor para todas as variáveis em U. Além do mais, é possível admitir que as linhas são resultados de experimentos aleatórios independentes. Então, é possível esperar, via a lei dos grandes números, que cada instanciação particular das variáveis eventualmente aparece no banco de dados com uma freqüência relativa aproximadamente igual as suas probabilidades.

Dada uma matriz de N linhas, admita que é desejado codificar, ou armazenar, as linhas como cadeias binárias. Existem diversas formas de se fazer isto, mas aqui o interesse reside no comprimento da codificação como uma métrica, via item 2 do princípio MDL, para comparar o mérito das redes bayesianas candidatas a modelo da distribuição. Sendo assim, admita que para esta tarefa vamos utilizar códigos de caracteres. Com códigos de caracteres cada evento singular é atribuído a uma única cadeia binária. Cada uma das linha são convertidas para seus códigos de caracteres e as N linhas são representados pela cadeia formada da concatenação dos códigos de caracteres. Por exemplo, foi atribuído o código 0000 para o evento $e_{111} = \{X_1 = 1, X_2 = 1, X_3 = 1\}$, e o código 1011 ao evento $e_{232} = \{X_1 = 2, X_2 = 3, X_3 = 2\}$. Portanto, se o base de dados consiste de seqüências de eventos singulares, $e_{111}, e_{111}, e_{232}$ seria então codificado como a cadeia binária 000000001011 utilizando-se códigos de caracteres.

A Teoria da Informação apresenta diversas formas para minimizar o comprimento médio da cadeia concatenada, levando-se em consideração a freqüência de ocorrência de diferentes eventos singulares. De fato, existem diversos algoritmos para a codificação ótima, isto é, comprimento mínimo, para os códigos de caracteres. Exemplos destes são o algoritmo de codificação aritmético [14] e o algoritmo de Huffman para a geração do código de Huffman [6]. No caso do código de Huffman o que se faz é atribuir à eventos que ocorrem mais freqüentemente, palavras-códigos mais curtas, de tal forma que o comprimento médio da cadeia representante da informação seja menor. Por exemplo, admita que existam 1000 linhas no banco de dados e os 12 eventos singulares especificados acima; então, utilizando um código de comprimento fixo, este necessitaria de 4 bits para codificar cada linha e de 4000 bits para codificar toda a base de dados. Por outro lado, admita que o evento e_{111} ocorre 500 vezes, o evento e_{232}, 300 vezes, e todos os outros 10 eventos ocorrem 20 vezes cada. Logo, se for atribuído a palavra-código 0 a e_{111}, 10 a e_{232} e as palavras-código 11111, 11110, 11101, 11100, 110111, 110110, 110101, 110100, 11001, 11000, para os 10 eventos restantes, seriam necessários $500 + 300 \times 2 + 6 \times (20 \times 5) + 4 \times (20 \times 6) = 2180$ bits para codificar todo a base de dados.

O algoritmo de Huffman requer como entrada a freqüência de ocorrência de cada evento no banco de dados. De fato, sua operação depende somente das freqüências relativas de ocorrência. Isto é, os números 500, 300 e 20 usados acima poderiam ter sido substituídos por 1/2, 3/10 e 2/100, onde o número total de linhas acima foi

fatorado fora.

Sendo assim, admita que na distribuição cada evento singular e_i tem probabilidade p_i. Então, o algoritmo de Huffman atribuirá a cada e_i, uma palavra-código de comprimento aproximado $-\log_2(p_i)$. Admitindo N linhas, com N muito grande, pode-se esperar Np_i ocorrências do evento e_i. Então, o comprimento da cadeia codificada representante do banco de dados será aproximadamente:

$$-N\sum_i p_i \log_2 p_i, \tag{2.10}$$

considerando um banco de dados formados apenas de eventos singulares.

É evidente que estas probabilidades p_i não são conhecidas, já que seu conhecimento implica na construção de uma rede bayesiana diretamente a partir desta informação. Ao contrário, é possível construir uma rede bayesiana determinada pela base de dados, a partir de algum esquema de aprendizado. Esta rede bayesiana serve como modelo da distribuição, atribuindo uma probabilidade q_i a todo evento singular e_i. Em geral, q_i não será igual a p_i, já que o esquema de aprendizado não pode garantir que será construída uma rede perfeitamente precisa. Sendo assim, a meta é tornar q_i o mais próximo possível de p_i para obter um modelo suficientemente preciso.

A aprendizagem de redes bayesianas é admitida como a representação da melhor hipótese da distribuição que originou os dados. Portanto, dado que as probabilidades q_i determinadas pela rede são as melhores suposições dos verdadeiros valores, pode-se projetar o nosso código de Huffman utilizando estas probabilidades. Isto significa dizer que será atribuído a cada evento e_i uma palavra-código de comprimento $-\log_2(q_i)$, ao invés do valor ótimo $-\log_2 p_i$. A despeito do uso dos valores q_i na atribuição das palavras-código, a base de dados continuará a ser determinada pelas verdadeiras probabilidades p_i. Isto é, espera-se que para N grande, teremos Np_i ocorrências do evento e_i, sendo p_i a probabilidade do evento e_i ocorrer. Assim, quando é utilizado o aprendizado com redes bayesianas para codificar os dados, o comprimento da palavra-código representante do banco de dados será aproximadamente:

$$-N\sum_i p_i \log_2(q_i). \tag{2.11}$$

Neste ponto, surge a dúvida da relação entre este comprimento e àquele obtido do conhecimento das probabilidades verdadeiras, p_i. Para esclarecer isto, considere a

definição de Distância de Kullback-Leibler, também conhecida por entropia relativa [36].

Definição 9 (Distância de Kullback-Leibler) [36]

- *A entropia relativa ou Distância de Kullback-leibler entre duas funções massa de probabilidade $p(x)$ e $q(x)$ é definida como:*

$$d(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}, \tag{2.12}$$

onde na definição acima, foi usada a convenção $0 \log \frac{0}{q} = 0$ e $p \log \frac{p}{0} = \infty$. A distância de Kullback-Leibler é sempre não-negativa e somente é nula se e somente se $p = q$. A entropia relativa pode ser pensada como uma "distância entre distribuições".

Agora, pode-se verificar o custo adicional em bits/símbolo por não utilizar-se a verdadeira distribuição p_i.

$$-\sum p_i \log_2 q_i = -\sum p_i \log_2 \left(q_i \cdot \frac{p_i}{p_i} \right)$$

$$= -\sum p_i \log_2 p_i + \sum p_i \log_2 \frac{p_i}{q_i}$$

$$= H(p) + \sum p_i \log_2 \frac{p_i}{q_i}$$

$$= H(p) + d(p||q)$$

Assim, a discriminação pode ser interpretada como o custo adicional pelo uso de uma distribuição diferente da verdadeira distribuição para codificar o evento e_i. Em outras palavras, a codificação deste evento pode ser idealmente realizada usando $H(p)$ bits por símbolo. Contudo, se for utilizado uma distribuição $q \neq p$ para codificar este mesmo evento, o melhor resultado exigirá pelo menos $H(p) + d(p||q)$ bits por símbolo.

O princípio MDL afirma que deve-se escolher uma rede que minimize a soma das suas próprias codificações de comprimento (dependentes da complexidade da rede), com as codificações de comprimentos dos dados dado o modelo, os quais dependem da proximidade das probabilidades q_i, determinadas pela rede, em relação as probabilidades verdadeiras p_i, isto é, dependentes da precisão do modelo.

É possível usar a Equação (2.12) para avaliar o segundo item requerido pelo princípio MDL, a codificação dos dados utilizando o modelo. Contudo, existem dois problemas com a utilização direta desta equação. O primeiro deles é o desconhecimento dos valores p_i. Em alguns casos, apesar disto, este problema pode ser superado. Pela lei dos grandes números, espera-se que o evento e_i, aparecerá no banco de dados de N pontos aproximadamente Np_i vezes, no caso de N grande. Logo, pode-se usar o número total de ocorrências de e_i dividido pelo número de pontos de dados como uma estimação de p_i. O segundo problema é mais difícil de resolver. A Equação (2.12) envolve uma soma sobre todos os eventos singulares, os quais são exponenciais com o número de variáveis.

As limitações consideradas nesta Seção tornam o cálculo da MDL ideal não-realizável, já que o tamanho da base de dados pode não ser grande. Assim, Bouckaert [2] propôs uma solução heurística que proporciona uma aproximação para a MDL ideal. Esta se fundamenta no Teorema 2. Segue-se uma definição formal para o comprimento de descrição.

Definição 10 (Comprimento de Descrição) [2]

- *Seja U um conjunto de nós $\{X_1, X_2, .., X_n\}$, $n \geq 1$, em que cada X_i pode assumir valores em $\{x_{i1}, x_{i2}, ..., x_{ir_i}\}$, $r_i \geq 1$, $i = 1, 2, ..., n$. Seja D uma base de dados de casos em U e para cada variável X_i, seja Pa_i o conjunto de pais de X_i em B_S. Além disso, para cada conjunto Pa_i, seja pa_{ij} sua j-ésima instância em relação a D, $j = 1, 2, ..., q_i$, $q_i \geq 0$. Fazendo N_{ijk} ser o número de casos em D no qual a variável X_i possui o valor x_{ik} e Pa_i possui o valor pa_{ij} e sendo $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, então o comprimento de descrição, $L(B_s, D)$, de uma estrutura B_S dado uma base D é definida como:*

$$L(B_s, D) = -\log P(B_S) + NH(B_s, D) + \frac{\log N}{2} K. \tag{2.13}$$

Sendo

$$H(B_s, D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} -\frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}}, \tag{2.14}$$

e,

$$K = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} - \left(\prod_{X_j \in Pa_i} r_j \right) (r_i - 1). \tag{2.15}$$

No apêndice A encontra-se parcialmente reproduzido o trabalho de Bouckaert [2], no qual é apresentado uma aproximação para a MDL ideal. O leitor que desejar pode utilizar-se do apêndice para avaliar as considerações realizadas por Bouckaert na obtenção da aproximação para MDL ideal.

O termo $H(B_S, D)$ corresponde a *entropia condicional* da rede em relação aos dados e o termo K é uma constante em relação a N. K diz respeito à complexidade da rede, quanto maior for o número de conexões entre os nós, maior será seu valor. Pelo gráfico da Figura 2.5, pode-se perceber que que a MDL estimada pela Equação (2.13) procura selecionar uma solução de compromisso entre os termos $NH(B_S, D)$ e $\frac{\log N}{2} K$, considerando $-\log P(B_S)$ constante, ou seja, considerando que todas as redes tem igual probabilidade de ser selecionada. Na realidade, deseja-se a codificação mais simples, mas não uma que seja tão simples ao ponto de não explicar adequadamente o comportamento dos dados.

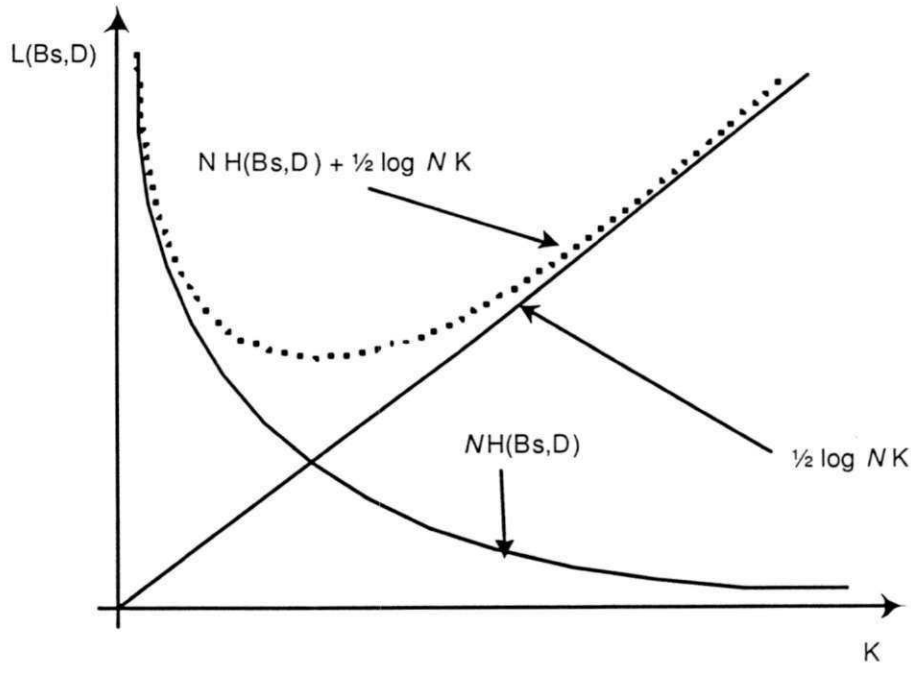


Figura 2.5: Medida de Descrição de Comprimento *versus* complexidade da rede

A próxima definição é útil na apresentação de uma propriedade importante da métrica de descrição de comprimento, formalizada no Teorema 1.

Definição 11 [2] *Seja U um conjunto de variáveis, P uma distribuição de probabilidade sobre U e B_S uma estrutura de uma rede probabilística definida sobre U. Sejam A, B e C conjuntos de variáveis aleatórias em U. Diz-se que A e B são* **condicionalmente independentes dado** *C, se*

$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

para todos os possíveis valores atribuídos às variáveis em A, B e C. Neste caso diz-se que A, B e C satisfazem uma **relação de independência** *em B_S. Um* **modelo de independência** *é o conjunto de todas as relações de independência em B_S.*

Uma propriedade importante da medida MDL, $L(B_S, D)$, é que todas as estruturas de rede que representam um mesmo conjunto de relações de independência entre as variáveis aleatórias têm a mesma medida, ou seja, a qualidade dessas estruturas, segundo o critério da medida MDL, é a mesma. Esta propriedade é formalmente estabelecida no seguinte teorema.

Teorema 1 [2]

Seja U um conjunto de variáveis, D uma base de dados sobre U e B_S uma estrutura de rede para U. Admita que a distribuição de probabilidade a priori *das estruturas de rede é uniforme. Então, para toda estrutura da rede $B_{S'}$ que representa o mesmo modelo de independência tem-se*

$$L(B_S, D) = L(B_{S'}, D)$$

A próxima seção descreve a métrica bayesiana, que é uma medida utilizada por Bouckaert [2] para estabelecer a adequação de um estrutura, B_S, na representação dos dados D.

2.3.3 Critério Bayesiano para Determinar B_S

O critério bayesiano para determinar B_S, também conhecido por medida bayesiana, tem o objetivo de encontrar a estrutura de rede mais provável dado um conjunto de dados D, isto é, maximizando a probabilidade $P(B_S|D)$. Para isto, Cooper e Herskovits

[8] derivaram uma fórmula para calcular $P(B_S, D)$ baseado na suposição que nenhum conjunto de funções de probabilidades condicionais B_P é preferido para uma estrutura antes da avaliação do banco de dados. Esta suposição implica que a distribuição de probabilidades a *priori* sobre os valores das funções de probabilidade condicional é uniforme [3].

O critério bayesiano é derivado da expressão:

$$P(B_S|D) = \frac{P(D|B_S)P(B_S)}{P(D)}. \tag{2.16}$$

Assim, se B_S é melhor representativo que $B_{S'}$ deve-se ter

$$\wedge(B_S, B_{S'}; D) = \frac{P(B_S|D)}{P(B_{S'}|D)} = \frac{P(D|B_S)P(B_S)}{P(D|B_{S'})P(B_{S'})} > 1. \tag{2.17}$$

Logo, a solução, B_S para o problema de encontrar a estrutura da rede a partir dos dados pode ser expressa como:

$$B_S^* = arg\max_{B_S} = \{P(D|B_S)P(B_S)\} \tag{2.18}$$

ou,

$$B_S^* = arg\max_{B_S} = \{P(B_S, D)\}. \tag{2.19}$$

Desta forma, a medida bayesiana pode ser utilizada para guiar a busca pela estrutura dentro do espaço de estruturas de redes bayesianas definido pelas variáveis do domínio do problema. De fato, Herskovits [25] desenvolveu uma expressão que fornece o valor de $P(B_S, D)$ e um algoritmo denominado K2, o qual é um procedimento para maximizá-la. O próximo Teorema formaliza a medida bayesiana.

Teorema 2 [25]

Seja U um conjunto de nós $\{X_1, X_2, .., X_n\}$, $n \geq 1$, em que cada X_i pode assumir valores em $\{x_{i1}, x_{i2}, ..., x_{ir_i}\}$, $r_i \geq 1$, $i = 1, 2, ..., n$. Seja D uma base de dados de casos em U e para cada variável X_i, seja Pa_i o conjunto de pais de X_i em B_S. Além disso, para cada conjunto Pa_i, seja pa_{ij} sua j-ésima instância em relação a D, $j = 1, 2, ..., q_i$,

$q_i \geq 0$. *Fazendo* N_{ijk} *ser o número de casos em* D *no qual a variável* X_i *possui o valor* x_{ik} *e* Pa_i *possui o valor* pa_{ij}. *Fazendo* $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, *então*

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2.20}$$

Prova

A Prova deste Teorema foi obtida de [25] e encontra-se reproduzida no Apêndice A caso o leitor deseje verificá-la.

■

A Equação (2.20) fornece uma medida para comparar a adequação de uma estrutura aos dados. O algoritmo K2, proposto por Herskovits [25] descreve um procedimento para a obtenção de $B_S^* = arg \max_{B_S} = \{P(D|B_S)P(B_S)\}$ baseado na métrica enunciada pelo Teorema 2. A idéia deste algoritmo será ilustrada na Seção 2.3.4.

O próximo Teorema, cuja prova foi omitida neste texto, relaciona a descrição do comprimento com a medida bayesiana.

Teorema 3 [2]

Seja U *um conjunto de variáveis aleatórias* $\{X_1, X_2, .., X_n\}$. *Seja* B_S *a estrutura de uma rede bayesiana e* D *uma base de dados completa com* N *casos. Seja* $P(B_S, D)$ *a medida bayesiana entre* B_S *e* D *e seja* $L(B_S, D)$ *uma medida de descrição de* B_S *e* D. *Então:*

$$L(B_S, D) = -\log P(B_S, D) + C \tag{2.21}$$

em que C *é uma constante que não depende de* N.

A próxima seção descreve as estratégias empregadas por alguns dos algoritmos de busca heurística que, utilizando algum método de avaliação da estrutura, procuram selecionar a estrutura mais "próxima" aos dados.

2.3.4 Estratégias de Aprendizado

Em princípio as relações entre os diferentes nós e suas correspondentes distribuições de probabilidades condicionais poderiam ser obtidas dos dados por ensaio e erro, testando

todas as possíveis combinações dos nós com relação a quem pode ser pai de quem. No entanto, devido ao grande número de combinações, o teste de todas as possibilidades é impraticável e para tanto deve ser usado um algoritmo que elimine uma parte desses nós. Em 92, Cooper e Herskovits demonstraram que o problema do aprendizado de redes bayesianas a partir de bases de dados é NP-completo [9].

Os algoritmos propostos com este objetivo podem ser separados em duas categorias: bayesianos e não bayesianos. Estes últimos usam testes estatísticos na base de dados para decidir a existência de arcos na rede a ser construída. Por outro lado, os algoritmos bayesianos admitem a existência de uma distribuição de probabilidades a *priori* sobre todas as possíveis redes, e calculam distribuições atualizadas usando a base de dados e escolhem a rede com a melhor distribuição atualizada. Uma questão central nestes métodos bayesianos é o estabelecimento de um critério de qualidade adequado para avaliação das distribuições atualizadas das estruturas de rede ("distribuições a *posteriori*")[2].

A abordagem bayesiana apresenta algumas vantagens bem conhecidas. Por exemplo, ela fornece critérios de parada naturais para os algoritmos, em lugar de comparações com limiares estabelecidos de forma arbitrária, usual em métodos não bayesianos. Além disto, a abordagem bayesiana permite incorporar de maneira fácil o conhecimento a *priori* sobre o domínio de interesse.

As próximas Seções descrevem algumas das diferentes abordagens utilizadas para o aprendizado de redes bayesianas. Maiores detalhes podem ser encontrados no trabalho de Cristiane Koehler [35].

Árvores de Chow-Liu

O método de Chow e Liu [10] foi a primeira estratégia desenvolvida para o aprendizado da estrutura de redes Bayesianas a partir de bases de dados. O método proposto considera a estrutura de árvore para n variáveis e apresenta complexidade de tempo $O(n^2)$ [24]. O algoritmo de Chow e Liu teve grande influência na área de aprendizado de redes bayesianas. Ele necessita de uma distribuição de probabilidade P como entrada e produz uma estrutura de árvore como saída. A idéia básica é comparar distribuições diferentes sobre duas variáveis no domínio que são estimadas a partir de bancos de dados. Na primeira distribuição, as duas variáveis são consideradas dependentes. Na segunda, elas são tomadas por independentes. Um grafo não direcionado é formado iniciando com um grafo sem arcos e adicionando um arco entre dois nós com máxima

entropia cruzada [11], desde que não crie um ciclo no grafo. Este processo é repetido até que todas as variáveis tenham sido consideradas. O passo final consiste em associar direções aos arcos de maneira a formar uma árvore. Como somente redes bayesianas com topologia de árvore são recuperadas por este método, a sua aplicação é restrita [3].

A principal contribuição do método de Chow e Liu não foi propriamente o aprendizado de redes bayesianas, mas sim o fato de aproximar uma distribuição de probabilidade conjunta sobre um conjunto de variáveis pelo produto de distribuições (condicionais) sobre duas variáveis. Esta idéia foi generalizada de forma a permitir distribuições sobre qualquer número de variáveis para aproximar a distribuição de probabilidade conjunta sobre o domínio [3].

Árvores Múltiplas de Rebane-Pearl

Este algoritmo é uma extensão direta do algoritmo de ChowLiu. Uma rede em árvore múltipla (também chamada de grafo simplesmente conectado) é uma estrutura que não contém ciclos, tal que exista no máximo um caminho entre dois nós quaisquer do grafo. Rebane e Pearl [43] provaram que quando a distribuição de probabilidade que gera os dados é originado de uma estrutura em árvore múltipla, o algoritmo proposto recupera esta com precisão. Eles também desenvolveram um método para encontrar a direcionalidade dos arcos no grafo, baseado no critério d - separação.

O algoritmo de Rebane e Pearl é baseado na avaliação das relações de dependência entre todos as triplas de variáveis presentes na árvore que é obtida pelo emprego do algoritmo de Chow e Liu. Assim, dada três variáveis, existem três casos possíveis: arcos formam uma estrutura serial, divergente ou convergente. Os dois primeiros casos são de difícil distinção, mas o terceiro é mais fácil, desde que as variáveis "pais" são marginalmente independente. Observe que em um primeiro momento não se utiliza relações de causalidade e o teste é estatístico.

O algoritmo proposto baseia-se nas seguintes etapas:

1. Gera-se o esboço da árvore através do algoritmo Chow e Liu;

2. Busca-se os nós internos da estrutura, saindo das camadas mais externas e caminhando em direção as camadas mais internas, até que um nó com mais de um pai seja encontrado (convergência);

3. Determina-se a direcionalidade de cada um dos arcos existentes no nó encontrado;

4. Para todo nó com pelo menos um arco de entrada, determinar a direcionalidade de todos os seus ramos adjacentes;

5. Repetir os passos de 2 a 4 até que nenhuma direcionalidade de arco possa mais ser determinada;

6. Caso ainda haja algum arco sem direção, utiliza-se algum conhecimento empírico de forma a direcionar o arco;

7. A partir da distribuição dos dados, calcula-se as probabilidades condicionais.

O algoritmo está restringido a árvores múltiplas e não garante obter todas as direções. Do ponto de vista prático, o uso de um conhecimento empírico pode levar a obtenção de uma estrutura de rede imprecisa.

Kutató

Em 90, Herskovits e Cooper desenvolveram um algoritmo que utiliza a medida de entropia condicional [11]. O algoritmo Kutató requer uma ordenação entre variáveis. Este aplica um algoritmo de busca que seleciona a rede de máxima verossimilhança. Segundo a abordagem adotada, a maximização da medida de verossimilhança é equivalente a minimização da entropia condicional do modelo dado os dados.

K2

O algoritmo K2 busca obter a estrutura B_S^* com complexidade de tempo polinomial, empregando para isto uma busca heurística, associada a uma ordenação prévia das variáveis, cuja idéia é bastante simples. Para cada nó, X_i, um dos possíveis predecessores, X_j, é acrescido ao conjunto Pa_i se $\{X_j\} \cup Pa_i$ maximiza a métrica bayesiana, isto é, se maximiza $P(B_S, D)$.

A maior desvantagem do algoritmo K2 é que a ordenação de variáveis influencia na complexidade da rede aprendida. Contudo, tal ordenação constitui uma restrição nas possíveis estruturas, reduzindo o espaço de busca, já que esta assegura que, dados dois nós X_i e X_j, se $i < j$, então X_i pode ser pai de X_j, mas X_j não pode ser pai de X_i. Por exemplo, considere o seguinte conjunto de nós $U = \{X_1, X_2, X_3\}$ dispostos

segundo esta ordenação. De acordo com este exemplo, os possíveis conjuntos de Pa_i são.

$$\begin{array}{lll} Pa_3 = \{X_1, X_2\} & Pa_2 = \{X_1\} & Pa_1 = \{\} \\ Pa_3 = \{X_1\} & Pa_2 = \{\} & \\ Pa_3 = \{X_2\} & & \\ Pa_3 = \{\} & & \end{array}$$

Baseado neste exemplo, pode-se perceber que se fosse possível realizar para cada nó, X_i, a busca do conjunto Pa_i que maximizasse localmente $P(B_S, D)$, seria necessário avaliar a Equação (2.20) um total de $4 + 2 + 1 = 2^2 + 2^1 + 2^0 = 2^3 - 1$ vezes. De um modo geral, pode-se perceber que para uma rede com n nós e uma ordenação dos mesmos, a obtenção exata de B_S^* requereria a avaliação da Equação (2.20) $2^n - 1$ vezes, demandando um tempo de complexidade exponencial. Logo, a complexidade de tempo polinomial é garantida evitando-se a avaliação de cada elemento do conjunto de estruturas individualmente de acordo com a medida bayesiana, onde, ao invés disso, busca-se a maximização desta métrica.

K3

Bouckaert [2] propôs um algoritmo denominado K3, o qual emprega a medida MDL em substituição a métrica bayesiana utilizada no algoritmo K2.

A idéia básica do algoritmo abaixo detalhado é buscar uma estrutura de rede que possibilite uma representação da base de dados com a menor quantidade de símbolos possível. Devido a sua simplicidade, esse algoritmo foi escolhido para ser implementado neste trabalho.

O procedimento parte de uma ordenação prévia das variáveis, feita por um especialista, e de uma rede chamada *rede atual*, que consiste de somente um nó representando a primeira variável. Em seguida é criada uma segunda rede chamada de *rede nova* na qual um novo nó é inserido e os nós atuais são avaliados um a um como candidatos a serem pai do novo nó: para cada possibilidade é avaliada uma medida associada ao critério MDL e a alternativa que forneça o menor valor é escolhida para compor a nova rede atual. O procedimento segue até que todas as variáveis sejam incluídas na estrutura de rede. Segue-se a apresentação de um algoritmo para a aprendizagem da rede bayesiana que utiliza a métrica MDL acima apresentada.

Algoritmo 1 $K3$[2]

Entrada Um Conjunto de Nós, uma ordenação e uma base de dados.

SAIDA: B_S, um conjunto de pais de cada nó.

1: $B_S \leftarrow \emptyset$ {Faz B_S uma estrutura com n nós e nenhum arco}
2: $\min MDL_x \leftarrow MDL(B_S, D)$ {calcula $MDL(B_S, D)$ utilizando a Equação (2.13) }
3: **para todo** nó X_i **faça**
4: **repita**
5: **para todo** nó X_i em Pa_i **faça**
6: **se** Arco $(X_l, X_i) \notin B_S$ **então**
7: $B_{S'} \leftarrow B_S \cup$ Arco(X_l, X_i)
8: $mdl \leftarrow MDL(B_{S'}, D)$ {Utilizando a Equação (2.13)}
9: **se** $mdl < \min MDL$ **então**
10: $\min MDL \leftarrow mdl$
11: $X_{lmin} \leftarrow X_l$
12: $X_{min} \leftarrow X_i$
13: **fim se**
14: **fim se**
15: **fim para**
16: **se** $\min MDL < MDL(B_S, D)$ **então**
17: $B_S \leftarrow B_S \cup$ Arco(X_l, X_i)
18: **fim se**
19: **até** $\min MDL \geq MDL(B_S, D)$
20: **fim para**

Após obtida a estrutura da rede faz-se necessário calcular as probabilidades condicionadas, definidas pelas relações de pais/filhos apresentada nessa estrutura. A próxima seção avalia este cálculo.

2.3.5 Obtenção das Probabilidades

Uma vez obtida a estrutura B_S é necessário voltar ao conjunto de dados para estimar as probabilidades condicionais dessa rede, ou seja o conjunto B_P. A aquisição de probabilidades consiste na determinação de todas as instâncias de um nó, $X_i = \{x_{i1}, x_{i2}, x_{i3}, ..., x_{ir_i}\}$ e de seus pais, avaliando a freqüência de ocorrência condicional de uma instância particular x_{ir}, dada a ocorrência da j-ésima instância de

Pa_i.

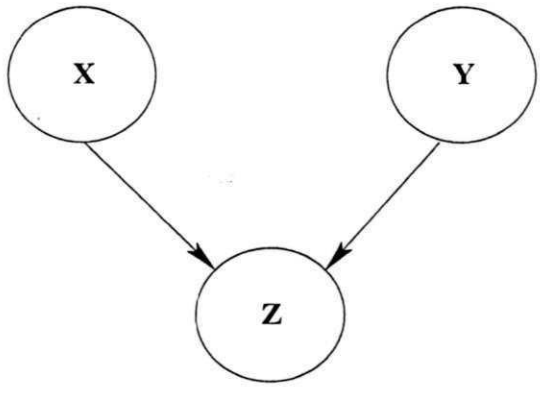


Figura 2.6: Instâncias avaliadas na aquisição de probabilidades

Considere inicialmente o desenvolvimento de uma solução baseada em um conjunto de dados, do qual foi obtida a rede esboçada na Figura 2.6, a qual ilustra uma rede bayesiana formada por três nós. Como X_1 e X_2 são nós raízes, o conjunto dos pais de X_1 e X_2 são vazios. Logo, as probabilidades condicionais se restringem ao nó X_3, isto é, aos valores $P(x_{31}|pa_{3j})$, $P(x_{32}|pa_{3j})$, ..., $P(x_{3r_3}|pa_{3j})$, em que pa_{3j} pode ser qualquer uma das $r_1 \times r_2$ possíveis combinações de instâncias dos nós X_1 e X_2, ou seja, $pa_{3j} \in \{x_{11}x_{21}, x_{11}x_{22}, ..., x_{1r_1}x_{1r_2}\}$.

Considere que a base de dados de observações das variáveis X_1, X_2 e X_3 seja formada por um conjunto de tuplas com os valores instanciados de cada uma destas três variáveis, isto é, $\mathcal{D} = \{(x^1_{1k_{11}}, x^1_{2k_{12}}, x^1_{3k_{13}}), ..., (x^N_{hk_{11}}, x^N_{hk_{12}}, x^N_{hk_{13}})\}$. Cada uma destas tuplas é denominada *caso* e a base de dados é dita ser completa se em cada caso existir observações (valores) sobre cada uma das n variáveis que compõem a rede.

Considere o caso geral em que o conjunto $\mathcal{D}$ é obtido a partir de uma rede com n variáveis. Uma vez que B_P é o elemento desconhecido e que se conhece B_S, suponha de agora em diante que a base de dados possa ser reagrupada conforme a matriz abaixo:

$$\mathcal{D} = \begin{bmatrix} N_{111}(x_{11}, pa_{11}), & N_{112}(x_{12}, pa_{11}), & \dots & N_{11r_i}(x_{1r_i}, pa_{11}) \\ N_{121}(x_{11}, pa_{12}), & N_{122}(x_{12}, pa_{12}), & \dots & N_{11r_i}(x_{1r_i}, pa_{12}) \\ \vdots & \vdots & \vdots & \vdots \\ N_{1q_i1}(x_{11}, pa_{1q_i}), & N_{1q_i2}(x_{12}, pa_{1q_1}), & \dots & N_{1q_1r_1}(x_{1r_1}, pa_{1q_1}) \\ \vdots & \vdots & \vdots & \vdots \\ N_{n11}(x_{n1}, pa_{n1}), & N_{n12}(x_{n2}, pa_{n1}), & \dots & N_{n1r_n}(x_{nr_n}, pa_{n1}) \\ \vdots & \vdots & \vdots & \vdots \\ N_{nq_n1}(x_{n1}, pa_{nq_n}), & N_{nq_n2}(x_{n2}, pa_{nq_n}), & \dots & N_{nq_nr_n}(x_{nr_n}, pa_{nq_n}) \end{bmatrix}, \quad (2.22)$$

com N_{ijk} denotando a quantidade de observações da tupla (x_{ik}, pa_{ij}). Utiliza-se o índice r_i para denotar a quantidade máxima de instâncias de um nó X_i. O índice q_i denota a quantidade máxima de instâncias do conjunto Pa_i. Ao longo de toda esta seção as letras i, j e k serão usadas com o seguinte sentido: a letra i denota o índice do nó, o qual pode variar de 1 até n; a letra j denota o índice dos pais de um nó e k, o índice de uma instância de um nó.

Os desenvolvimentos realizados nesta Seção são baseados nas seguintes suposições [25]:

1. A distribuição $f(x_{ik}|pa_{ij}) = P(x_{ik}|pa_{ij})$ é multinomial com parâmetros $\theta = \{\theta_{ij1}, ..., \theta_{ijr_i}\}$.

2. A distribuição a *priori* de θ é Dirichlet, com parâmetros $\nu_{ij1}, ..., \nu_{ijr_i}$.

3. Os dados amostrados são completos.

Naturalmente, a obtenção de θ_{ijk} consiste na estimação das proporções $\frac{N_{ijk}}{\sum_{k=1}^{r_i} N_{ijk}}$, que por hipótese possuem distribuição a *priori* Dirichlet. Utilizando o método de inferência bayesiana a estimação de θ_{ijk} segue a partir da aplicação do Teorema 10.

$$P(x_{ik}|pa_{ij}) = E_{h(\theta_{ij}|(x_{ik}, pa_{ij}))}[\theta_{ijk}] = \frac{\nu_{ijk} + N_{ijk}}{\nu_{ij} + N_{ij}} \quad (2.23)$$

em que $\nu_{ij} = \sum_{k=1}^{r_i - 1} \nu_{ijk}$ e $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

A Equação (2.23) pode ser empregada satisfatoriamente no cálculo de $E[\theta_{ijk}]$ se existir um conhecimento a *priori* sobre a distribuição de θ. Foi assumido que $h(\theta)$

é Dirichlet, mas nada foi sobre os valores dos parâmetros da distribuição. Sem o conhecimento a *priori* sobre estes parâmetros, Zabell[51] mostra que a Equação (2.23) pode ser aproximada por:

$$P(x_{ik}|pa_{ij}) = \frac{N_{ijk} + K}{N_{ij} + Kr_i}. \tag{2.24}$$

Em que K é uma constante tal que $N_{ij} + kr_i \neq 0$. Existem alguns valores para K e as escolhas mais conhecidas são $k = 1$, $k = \frac{1}{2}$ e $k = \frac{1}{r}$ [25].

Assim, a probabilidade condicionada é obtida pelas freqüências relativas das ocorrências dos nós e de seus pais. Para que essas probabilidades sejam confiáveis é necessário que a estatística fornecida pela freqüência relativa seja bem próxima da estatística real e, como é sabido, para que isso ocorra é necessário que a quantidade de dados seja suficientemente grande.

Dadas a estrutura da rede e suas relações de probabilidade condicionada, elas podem ser usadas para realizar as inferências. O algoritmo para isto é descrito no próximo Capítulo.

2.4 Redes Bayesianas Aproximadas

Cada vez mais ferramentas computacionais que auxiliam a tomada de decisão estão surgindo de forma a facilitar a solução de problemas complexos que tratam com incertezas. Neste sentido, aplicações baseadas em redes bayesianas vem constituindo uma solução promissora. Isto se deve ao fato de que estas permitem um tratamento adequado da incerteza, possibilitam a extração do conhecimento humano embutido na base de dados e permitem a inclusão do conhecimento especialista do domínio da aplicação. Aplicações correntes incluem diagnóstico médico [5] e visão computaconal [28].

Contudo, as aplicações práticas envolvem redes complexas, com topologias densas, devido ao grande número de nós (múltiplas causas) e o alto grau de conectividade (interação para um mesmo efeito), contribuindo para um cenário em que o raciocínio probabilístico com complexidade de tempo polinomial seja inviabilizado. É importante notar que redes complexas, com topologias apresentando um alto grau de conectividade possuem um conjunto de nós em que, ignorando-se os sentidos dos arcos, existe um caminho fechado formado por esses nós. Esse caminho fechado é denominado laço. Por

exemplo, a rede apresentada na Figura 2.7 consiste de um laço e o nó C é denominado o *nó extremo* desse laço.

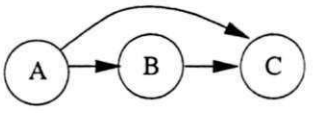


Figura 2.7: Exemplo de Rede Probabilística com Laço.

Assim, a questão abordada nesta seção pode ser sintetizada da seguinte forma: "após o aprendizado da rede tem-se o modelo mais razoável do problema. Contudo, normalmente este modelo é muito complexo e precisa ser aproximado". Ou seja, durante a etapa do aprendizado desejamos obter a rede mais próxima dos dados, mas após obtê-la a preocupação passa a ser reduzir a complexidade desta (o alto grau de conectividade), através de aproximações realizadas sobre o modelo, para efetivamente poder utilizá-la. Para isto, serão propostas técnicas que irão nortear a simplificação da rede pela mínima perda de informação proporcionada pela simplificação do modelo, através da remoção criteriosa de arcos com objetivo de eliminar laços [16] [17] [18].

2.4.1 Critérios para Remoção de Arcos

Os dois primeiros critérios usam para o cálculo da respectiva medida os parâmetros da rede como um todo e por isso proporciona uma otimização global. O terceiro é baseado em parâmetros de um conjunto de nós da rede, e por isso é dito ser local.

Basicamente os três critérios são baseados no algoritmo apresentado a seguir. O que diferencia um critério do outro será alterado no item 2.1 do algoritmo.

Algoritmo 2 Algoritmo para Remoção de Arcos

ENTRADA: Uma rede Probabilística

SAÍDA: Uma rede de árvore múltipla

ITERAÇÃO:

1. **Para** $i = 1...N_{nos} - 1$ **Faça**

2. Aplique o algoritmo com busca prioritária em profundidade para verificar se existe um laço.

2.1 Será modificado de acordo com o critério escolhido.

2.2 Aplique o algoritmo com busca prioritária em profundidade ao nó i para determinar se ainda existe laços

2.2.1 **Se** não existam mais laços que o nó i pertença, **então** i = i+1 e retorne para 2

2.2.2 **Se** $i = N_{nos}$ **FIM.**

2.2.3 **Se não** retorne para 2

3. **Fim do Algoritmo**

Critério 1: Distribuições Conjuntas

O procedimento proposto por Engelen em [19] consiste em, partindo da rede original, remover um dos arcos dessa rede e calcular a nova distribuição conjunta assim obtida, para em seguida calcular a divergência de informação de Kullback-Leibler [11] entre esta distribuição e a distribuição conjunta original. Os arcos cuja retirada proporcionam uma menor divergência são candidatos a serem removidos. Uma das desvantagens deste procedimento é que, para cada arco candidato à remoção, a distribuição de probabilidades da nova rede sem esse arco deve ser obtida. Sendo assim, o método de remoção de arcos proposto em [19] requer, necessariamente, várias construções de distribuições de probabilidades.

O primeiro critério aqui proposto é uma modificação do método de Engelen. Para isto, optamos por remover apenas um dos arcos presentes no laço, justamente aquele que minimiza a distância de Kullback-Leibler ou divergência informacional entre distribuições $p(x)$ e $q(x)$, doravante chamada simplesmente de divergência quando ficar

claro do contexto, cuja expressão encontra-se novamente reproduzida abaixo.

$$D(p||q) = \sum_{x} p(x) log \frac{p(x)}{q(x)}, \tag{2.25}$$

sendo $q(x)$ a distribuição conjunta da rede original e $p(x)$ a distribuição conjunta da rede após a remoção do arco.

A identificação de laços presentes na rede pode ser feita usando o algoritmo *depth-first search* [48]. Note que a remoção de um arco pode quebrar mais de um laço. Para se obter um procedimento eficiente (e sub-ótimo) de busca dos laços a serem quebrados, propõe-se o tratamento seqüencial e isolado de cada nó, no sentido de identificar e eliminar laços. Desta forma, após a retirada dos arcos necessários para quebrar todos os laços formados por um dado nó, o nó seguinte é testado quanto à existência de laços, e assim sucessivamente, até que o penúltimo nó seja testado.

O procedimento completo para remoção de arco que aqui se propõe usa o algoritmo proposto no início dessa seção com a alteração do ítem 2.1 pelo que se segue:

Algoritmo 3 Distribuições Conjuntas

[2.1] Se existir, vá ao nó extremo desse laço e remova um dos arcos presentes neste laço e obtenha a distribuição de probabilidades condicionais dessa nova rede, a partir da base de dados.De posse das probabilidades condicionais, obtenha a distribuição conjunta da rede corrente. Em seguida calcule a divergência entre essa distribuição e a rede original e armazene o valor. A seguir, proceda da mesma forma para os outros arcos presentes no laço. Remova o arco que resulte na menor divergência e obtenha a rede corrente.

É possível observar que esse método requer uma série de avaliações de distribuições conjuntas e que, para cada novo laço encontrado, as distribuições têm que ser re-calculadas.

Critério 2: Descrições de Comprimento Mínimo

A escolha de determinada rede bayesiana para modelar uma base de dados normalmente é feita pela maximização de algum critério de "proximidade" entre essa rede e a base de dados. Por exemplo, no algoritmo K3 [2], o parâmetro do comprimento de descrição é usado. Nesse caso, a rede escolhida é a que apresenta a menor métrica do comprimento de descrição (CD). Portanto, parece razoável utilizar este mesmo parâmetro para medir a "proximidade" entre as redes obtidas pela remoção seqüenciada dos arcos presentes

num laço. O arco escolhido para ser removido é aquele que altere menos essa medida de proximidade.

Esse raciocínio pode ser usado para obter um algoritmo de eliminação de laços que dispense as construções de novas distribuições de probabilidades e, assim, produza uma redução significativa de complexidade em relação ao procedimento proposto em [19] e adaptado neste trabalho, resultando no critério 1 acima descrito. Com base nesta idéia e na métrica CD propõe-se um segundo critério e um outro algoritmo para a eliminação de laços.

O objetivo do algoritmo é buscar a rede sem laços derivada da rede original por retirada de arcos cuja métrica CD mais se aproxima daquela calculada utilizando a rede original, considerada ótima. Isto é, aplicando localmente este raciocínio a cada laço a ser quebrado, a escolha de um arco para ser removido com o objetivo de eliminar este laço deveria recair sobre o arco candidato cuja retirada produzisse a menor alteração na medida CD da rede. A seguir, temos o ítem 2.1 correspondente a este critério, que deve ser substituído no algoritmo proposto no início da seção.

Algoritmo 4 Descrições de Comprimento Mínimo

[2.1] Se existir, vá ao nó extremo desse laço e remova um dos arcos presentes neste laço e calcule o CD da rede obtida pela remoção do arco. Armazene o valor da DL corrente e, na rede atual, insira o arco removido. A seguir, proceda da mesma forma para cada um dos arcos presentes no laço. Remova o arco que resulte no CD mais próxima da rede original, obtendo a rede corrente.

Critério 3: Entropia Condicionada

A entropia condicionada $H(X \mid Y)$ entre duas variáveis discretas X e Y é dada pela Equação (2.26) abaixo e pode ser vista como uma medida de incerteza sobre a primeira variável quando se tem algum conhecimento da segunda [11]. Assim, quanto maior $H(X \mid Y)$ maior a incerteza sobre X dado que se conhece Y.

$$H(X \mid Y) = -\sum_{x}\sum_{y} P(X,Y)\log(P(X \mid Y)) \tag{2.26}$$

Dados dois arcos que chegam em um nó Na provenientes dos nós Nb e Nc, propõe-se usar para escolha do arco a ser eliminado o critério de maior entropia condicionada. Isso se justifica pelo fato de que, se $H(Na \mid Nb) > H(Na \mid Nc)$, o conhecimento de Nb

mantém a incerteza sobre Na em um nível maior do que quando se tem conhecimento de Nc. Portanto, se for necessário eliminar um dos arcos que chegam a Na deve ser eliminado o que liga os nós Na e Nb. As distribuições de probabilidades necessárias para o cálculo das entropias condicionadas podem ser obtidas dos dados.

Para esse critério, tem-se o item 2.1 do algoritmo inicial dado por

Algoritmo 5 Entropia Condicionada

[2.1] Se existir, vá ao nó extremo desse laço a encontre as entropias condicionadas entre esse nó e os seus antecessores que pertencem ao laço. Retire o arco que liga os nós de maior entropia condicionada e volte para 2.

Aplicações

Para testar os critérios propostos foram usadas duas redes obtidas de bases de dados obtidas na Internet em (*http://www.cs.auc.dk/ fvj/BNIDpage/datamine.htm*) a primeira base está relacionada a ocorrência de gravidez e a segunda a possibilidade do paciente ter angina, cada base de dados contém 10.000 casos. Para comparar os critérios foi escolhido o valor da medida comprimento de descrição (CD) da rede final, ou seja, a rede final que apresentar o menor CD será considerada a mais adequada para representar a base de dados.

A primeira base de dados refere-se a um conjunto de quatro variáveis binárias. Aplicando o algoritmo K3 [2] obtivemos a rede apresentada na Figura 2.8, onde pode ser observada a ocorrência de laços.

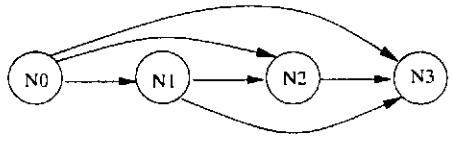


Figura 2.8: Rede que contém laços aprendida da base de casos de gravidez.

Aplicando os critérios propostos para a remoção de arcos à rede apresentada na Figura 2.8, obtém-se no primeiro e no segundo critérios a rede apresentada na Figura 2.9 e para o terceiro critério a rede apresentada na Figura 2.10.

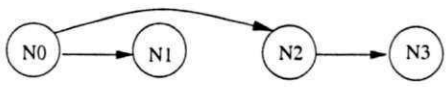


Figura 2.9: Rede obtida após a quebra dos laços da rede apresentada na Figura 2.8 usando os Critérios 1 e 2.

Figura 2.10: Rede obtida após a quebra dos laços da rede apresentada na Figura 2.8 usando o Critério 3.

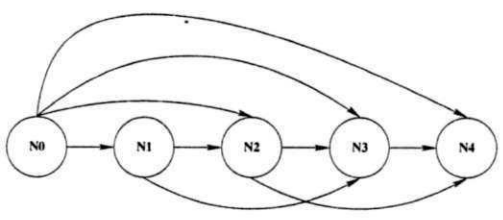


Figura 2.11: Rede que contém laços aprendida da base de casos de angina.

Os valores do CD para a rede original e para cada das redes obtidas são os seguintes: a rede original tem $CD = 25.741.332$ bits, a rede obtida pelo primeiro e segundo critérios tem $CD = 27.402.540$ bits e pelo terceiro $CD = 28.315,210$ bits.

Como era de se esperar, a aplicação dos 3 algoritmos de eliminação de laços aumenta a medida CD, refletindo o fato de que uma retirada de arcos da rede original implica uma descrição menos precisa dos dados.

Em termos de comparação entre os diferentes critérios de eliminação de laços, observa-se que as redes obtidas aplicando os critérios 1 e 2 apresentam métrica CD menor, sendo portanto melhores sob este aspecto.

Repetindo o procedimento acima para segunda base de dados, que refere-se a um conjunto de cinco variáveis, foi obtida a rede apresentada na Figura 2.11 pelo algoritmo K3, e as redes apresentadas nas Figuras 2.12, Figuras 2.13 e Figuras 2.14 para o primeiro, segundo e terceiro critérios respectivamente.

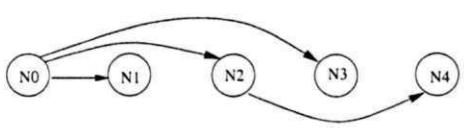


Figura 2.12: Rede obtida após a quebra dos laços da rede apresentada na Figura 2.11 usando o Critério 1.

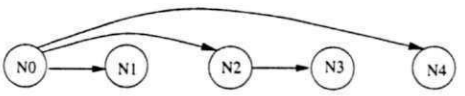


Figura 2.13: Rede obtida após a quebra dos laços da rede apresentada na Figura 2.11 usando o Critério 2.

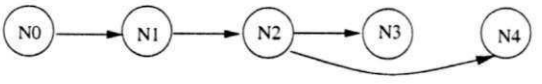


Figura 2.14: Rede obtida após a quebra dos laços da rede apresentada na Figura 2.11 usando o Critério 3.

A rede original tem $CD = 12.605,085$ bits, a rede obtida pelo primeiro critério tem $CD = 13.023.3517$ bits, pelo segundo tem $CD = 13.135,25$ bits e pelo terceiro $CD = 13,047,221$ bits.

Observou-se mais uma vez que todos os critérios aumentaram a medida CD, como esperado. Dos três critérios, o primeiro forneceu uma rede mais "próxima" da rede original. O terceiro critério apesar de ser aplicado localmente, forneceu uma rede melhor que o segundo que é global.

Com relação ao primeiro critério, que apresentou melhores resultados nos dois experimentos realizados, cabe lembrar que ele requer um esforço computacional bem maior que os demais, devido aos vários cálculos de distribuições condicionais e conjuntas necessários para sua aplicação.

Na obtenção da rede a partir dos dados foi usado o algoritmo K3, que é baseado na busca de uma rede que minimize a medida CD com relação aos dados. Portanto, é natural questionar o por quê de não se incluir o critério da remoção de arcos no algoritmo de construção da rede, ou seja logo na construção da rede ao detectar um laço usar o critério de menor CD para quebrá-lo. O algoritmo K3 faz a busca a partir do segundo nó até o último de quais nós podem ser seus pais. Caso o critério de remoção de laços fosse incluído na obtenção da rede, ao detectar o laço em um determinado nó, os nós seguintes a ele ainda não tem sua relação de parentesco estabelecida, e assim sua influência na rede ainda não está estabelecida; sendo possível que ao estabelecer a relação de parentesco do nós seguintes, ao que foi detectado o laço, os valores do CD para os arcos que compõem o laço possam mudar. Portanto, para que a influência de todos os nós possa ser levada em conta, deve-se construir a rede e, em seguida, usar o critério desejado.

2.5 Conclusões

Este Capítulo apresentou uma breve descrição das diferentes abordagens adotadas no desenvolvimento de sistemas de auxílio à tomada de decisão, mostrando que nas últimas décadas vêm crescendo o interesse no desenvolvimento de aplicações baseadas em redes bayesianas. Isto é devido a características como, por exemplo, o fato de que estas permitem um tratamento adequado da incerteza, com base em princípio matemáticos bem fundamentados (Teoria das Probabilidades); possibilitam a extração do conhecimento humano embutido na base de dados (obtenção automática do modelo) e permitem a inclusão do conhecimento especialista do domínio da aplicação.

Muitos dos princípios utilizados na aprendizagem bayesiana são baseados em conceitos da Teoria da Informação, os quais também são empregados na obtenção de redes aproximadas por meio da remoção criteriosa de arcos.

No próximo Capítulo, é descrito o mecanismo de inferência em redes bayesianas, sendo este avaliado nas redes que tiveram seus arcos removidos de acordo com os critérios propostos nesta Seção.

Capítulo 3

Inferência em Redes Bayesianas

3.1 Introdução

O termo "inteligência artificial" (IA) surgiu em 1956 durante o encontro de Marvin Minsky e John McCarthy no *Dartmouth College* em *New Hampshire*, EUA. Para John McCarthy, criador do termo, o conceito de IA deve ser compreendido como a ciência e a engenharia aplicadas à elaboração de máquinas inteligentes, em especial, programas de computadores inteligentes.

Atualmente, estudos de IA podem ser divididos em duas grandes áreas: o desenvolvimento de sistemas que facilitem o trabalho físico humano (robôs) e o desenvolvimento de sistemas que auxiliam a tomada de decisões. Neste último, duas abordagens são utilizadas: o raciocínio lógico e o raciocínio probabilístico. O raciocínio lógico pondera sobre o conhecimento prévio a respeito do problema e de acordo com este, obtém conclusões. Esta abordagem pode não ser útil em situações onde o conhecimento prévio possui diferentes níveis de confiabilidade. Para estes casos, o raciocínio probabilístico surge como uma solução promissora. Assim, ao invés de tomar decisões através da classificação de evidências utilizando valores booleanos como verdadeiro (V) ou falso (**F**), é possível obter, através do raciocínio probabilístico, conclusões mesmo em situações onde as informações envolvidas não podem ser completamente classificadas como puramente verdadeiras, isto é, V; ou puramente falsas, isto é, F.

Na realidade, a grande dificuldade do processo de "tomar uma decisão" é devido à constante presença de incertezas no domínio do problema. Assim, aplicações viáveis não devem procurar eliminar as incertezas envolvidas, mas tratá-las de forma rigorosa, procurando minimizar os riscos da tomada de decisões erradas.

3.2 Inferência Probabilística

A inferência pode ser definida como a operação mental pela qual extraímos uma proposição nova (conclusão) de uma ou mais proposições já conhecidas (premissas). Existem várias formas de inferência, das quais é possível destacar, por exemplo, a dedução e a indução. É freqüente que alguns destes tipos de inferência apareçam associadas a uma dada forma de representação do conhecimento. Por exemplo, é habitual usar dedução, como forma de inferência, quando trata-se conhecimento em lógica.

A motivação para utilização de inferência de base probabilística é que os seres humanos têm certa facilidade em moldar fatos e fenômenos na forma de relacionamentos causais. Assim, um ponto de partida adequado para a avaliação de uma estratégia comum de IA é considerar com atenção o mecanismo de inferência humano, entendido como o meio pelo qual as pessoas integram dados de múltiplas fontes e geram uma interpretação coerente destes dados. A análise deste mecanismo indica que a maneira como o ser humano armazena e utiliza o seu conhecimento a respeito de um determinado domínio, não fica bem caracterizada por uma distribuição conjunta, mas através de distribuições marginais e condicionais envolvendo pequenos agrupamentos de variáveis. Neste sentido, por exemplo, poderíamos analisar a abordagem que é utilizada em escolas de medicina para o desenvolvimento de uma aplicação de IA: apresentação de casos clínicos aos estudantes e enumeração de hipóteses mais prováveis, ou seja, através de relacionamentos causais do tipo "dado que o paciente apresenta os sintomas "α", "β", "γ" e "δ", é mais provável ele ser portador da patologia θ". Evidentemente, poderíamos expressar estas dependências através de relações de dependência e independência condicional tal como definimos na abordagem bayesiana e os níveis de confiabilidade (presença de incertezas) por meio de probabilidades. Com isto, poderíamos desenvolver um sistema de auxílio ao diagnóstico com resposta probabilística, ou seja, a partir de entradas correspondendo a sintomas, este faria uma atualização das probabilidades a *posteriori* em função dos valores observados, retornando as hipóteses mais prováveis ordenadas na ordem decrescente de probabilidades, auxiliando médicos a tomarem uma decisão razoável. Evidentemente, precisaríamos de um modelo do domínio do problema baseado em relacionamentos causais, o qual permitiria a realização de inferências.

3.2.1 Mecanismo de Inferência

Admitindo que uma rede probabilística está especificada, ela pode ser usada para realizar interpretações a respeito de dados de entrada específicos. Isto é um processo que compreende instanciar um conjunto de variáveis, correspondentes aos dados de entrada, e calcular seu impacto nas probabilidades a "*posteriori*" do conjunto de variáveis de saída, selecionando as hipóteses mais prováveis.

A distribuição de probabilidade a *posteriori* associada a cada nó ao longo do processamento iniciado para estabilizar a rede, isto é, após a apresentação de cada uma das evidências à rede, será doravante denominada "distribuição induzida," "crença" ou "convicção" (em inglês é usado o termo *belief*). O Exemplo 3.1 ilustra o processo de tomada de decisão baseada na inferência probabilística, enquanto que o Exemplo 3.2 apresenta a idéia explorada ao longo de todo este Capítulo: a atualização das probabilidades a *posteriori* por meio de mensagens enviadas aos nós adjacentes.

Exemplo 3.1 [25]: Considere o modelo simplificado de representação de problemas visuais, ilustrado pela Figura 3.1 e pela Tabela 3.1, onde o nó A é a variável aleatória associada a idade do paciente $\geq$ 75 anos, G é a variável aleatória associada a necessidade do paciente de óculos, C modela a presença de catarata, V é a variável aleatória associada a melhora da visão do paciente, enquanto que S modela a presença da visão fraca. Por fim, R é a variável aleatória que acusa a presença pelo paciente do reflexo retinal detectável.

Tabela 3.1: Probabilidades da rede bayesiana do Exemplo 3.1. Nesta Tabela, "T" significa verdadeiro e "F" falso.

$P(A = T) = 0.10$	$P(G = T \mid A = T) = 0.75$
$P(C = T \mid A = T) = 0.40$	$P(V = T \mid G = T) = 0.80$
$P(S = T \mid G = T, C = T) = 0.95$	$P(S = T \mid G = F, C = T) = 0,40$
$P(R = T \mid C = T) = 0.25$	$P(G = T \mid A = F) = 0.15$
$P(C = T \mid A = F) = 0.01$	$P(V = T \mid G = F) = 0.05$
$P(S = T \mid G = T, C = F) = 0.75$	$P(S = T \mid G = F, C = F) = 0.05$
$P(R = T \mid C = F) = 0.95$	

Baseado neste modelo, pode-se realizar facilmente algumas inferências simples, como obter a probabilidade de um indivíduo com mais de 75 anos não usar óculos, que é

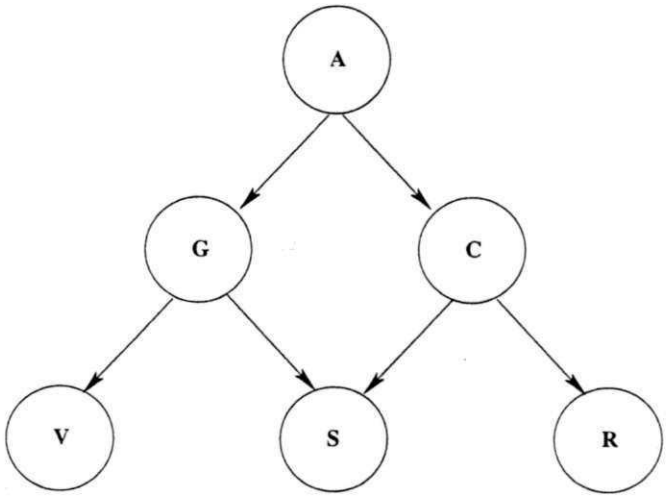


Figura 3.1: Rede bayesiana associada ao modelo de representação de problemas visuais.

dada por $P(G = F|A = T) = 1 - P(G = T|A = T) = 0.25$, ou obter a probabilidade de um indivíduo com mais de 75 anos possuir catarata e precisar usar óculos, que é dada pelo produto $P(G = T|A = T)P(C = T|A = T) = 0.30$. De um modo geral, conhecido o estado de um nó, a probabilidade conjunta de seus descendentes imediatos é dada pelo produto de suas probabilidades individuais, isto é, sendo V e S filhos de G, então V e S são independentes dado o estado de V. O que é expresso matematicamente como:

$$P(V, S|G) = P(V|G)P(S|G) \tag{3.1}$$

O sentido físico expresso nesta equação pode ser ilustrado com os dados da figura 3.1 como "*dado que um indivíduo possui mais de 75 anos e usa óculos, nada se pode dizer sobre o fato dele possuir catarata*". Isto significa que "*precisar de óculos*" (G) e "*possuir catarata*" (C) são **independentes** em relação à idade do paciente (A). Entretanto, uma relação de dependência entre G e C se verifica quando se conhece o estado da variável S. Uma vez que ambos G e C têm influência sobre S, isto é, tanto a "*necessidade de usar óculos*" como "*ter catarata*" contribuem para que o paciente acuse "*visão fraca*", então G e C **não são independentes** em relação à S. Na prática, sabendo-se que o paciente acusa visão fraca e sabendo-se que ele não usa óculos, poderia-se inferir, baseado nos dados do modelo, que ele possui catarata.◊

Exemplo 3.2 [44]: Considere a rede bayesiana mostrada na Figura 3.2(a). As

probabilidades a *priori* de todas as variáveis podem ser calculadas da seguinte forma:

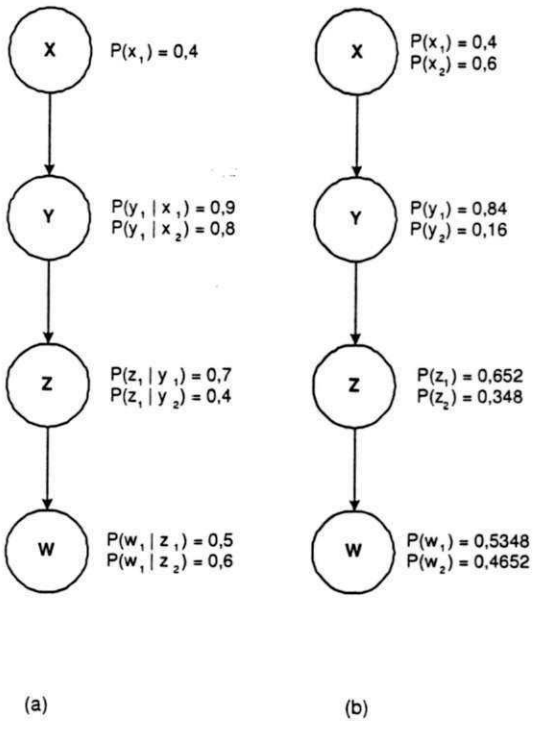


Figura 3.2: Uma rede bayesiana é mostrada em (a) e as probabilidades a *priori* nesta rede são mostradas em (b). Cada variável tem somente dois valores, um dos quais é apresentado em (a).

$$p(y_1) = p(y_1|x_1)p(x_1) + p(y_1|x_2)p(x_2) = 0,9 \times 0,4 + 0,8 \times 0,6 = 0,84$$

$$p(z_1) = p(z_1|y_1)p(y_1) + p(z_1|y_2)p(y_2) = 0,7 \times 0,84 + 0,4 \times 0,16 = 0,652$$

$$p(w_1) = p(w_1|z_1)p(z_1) + p(w_1|z_2)p(z_2) = 0,5 \times 0,652 + 0,6 \times 0,348 = 0,5348$$

Estas probabilidades são mostradas na Figura 3.2(b). Observe que o cálculo para cada variável requer informação (probabilidade a *priori*) disponibilizada pelo seu pai. Admita que esta informação possa ser passada através de "mensagens" enviadas ao nó sobre foco sempre que este desejar. Desta forma, podemos considerar este método de

cálculo através de um algoritmo de passagem de mensagens. Neste, cada nó envia ao seu nó descendente, denominado de nó filho, uma mensagem necessária para calcular as probabilidades a *priori* do seu filho.

Admita agora que o nó X é instanciado para x_1. Desde que numa rede bayesiana as relações de dependência condicional existem apenas entre nós adjacentes diretos, é possível calcular as probabilidades a *posteriori* das variáveis remanescentes através do envio de mensagens que se propagam abaixo do nó X, isto é, "descendo na rede".

$$p(y_1|x_1) = 0,9$$

$$\begin{aligned} p(z_1|x_1) &= p(z_1|y_1, x_1)p(y_1|x_1) + p(z_1|y_2, x_1)p(y_2|x_1) \\ &= p(z_1|y_1)p(y_1|x_1) + p(z_1|y_2)p(y_2|x_1) \\ &= 0,7 \times 0,9 + 0,4 \times 0,1 = 0,67 \end{aligned}$$

$$\begin{aligned} p(w_1|x_1) &= p(w_1|z_1, x_1)p(z_1|x_1) + p(w_1|z_2, x_1)p(z_2|x_1) \\ &= p(w_1|z_1)p(z_1|x_1) + p(w_1|z_2)p(z_2|x_1) \\ &= 0,5 \times 0,67 + 0,6 \times 0,33 = 0,533 \end{aligned}$$

A Instanciação precedente mostra como é possível utilizar a propagação de mensagens para baixo para calcular as probabilidades condicionais de variáveis abaixo da variável instanciada.

Admita agora que W foi instanciado para w_1 (e nenhuma outra variável é instanciada). É possível utilizar a propagação de mensagens para acima de forma a calcular as probabilidades condicionais das variáveis remanescentes como segue. Primeiro utiliza-se do Teorema de Bayes (Apêndice A) para calcular $p(z_1|w_1)$:

$$p(z_1|w_1) = \frac{p(w_1|z_1)p(z_1)}{p(w_1)} = \frac{0,5 \times 0,652}{0,5348} = 0,6096$$

Então, para calcular $P(y_1|w_1)$, novamente aplica-se o Teorema de Bayes como segue:

$$p(y_1|w_1) = \frac{p(w_1|y_1)p(y_1)}{p(w_1)}$$

Contudo, desta vez não podemos completar este cálculo porque $p(w_1|y_1)$ não é conhecido. Entretanto, é possível obter o seu valor da maneira mostrada quando apresentamos a discussão da propagação de mensagens abaixo da variável instanciada (caso anterior). Ou seja,

$$p(w_1|y_1) = p(w_1|z_1)p(z_1|y_1) + p(w_1|z_2)p(z_2|y_1)$$
$$= 0,50 \times 0,7 + 0,60 \times 0,3 = 0,53$$

Após concluir este cálculo, obtemos também $p(w_1|y_2)$ (porque X irá precisar deste valor posteriormente), então determinamos $p(y_1|w_1)$ e passamos as mensagens $p(w_1|y_1)$ e $p(w_1|y_2)$ para X. Então, na seqüência calculamos $p(w_1|x_1)$ e $p(x_1|w_1)$. Isto é,

$$p(w_1|y_2) = p(w_1|z_1)p(z_1|y_2) + p(w_1|z_2)p(z_2|y_2)$$
$$= 0,50 \times 0,4 + 0,60 \times 0,6 = 0,56$$

$$p(y_1|w_1) = \frac{p(w_1|y_1)p(y_1)}{p(w_1)} = \frac{0,53 \times 0,84}{0,5348} = 0,8325$$

$$p(w_1|x_1) = p(w_1|y_1)p(y_1|x_1) + p(w_1|y_2)p(y_2|x_1)$$
$$= 0,53 \times 0,90 + 0,56 \times 0,1 = 0,533$$

$$p(x_1|w_1) = \frac{p(w_1|x_1)p(x_1)}{p(w_1)} = \frac{0,533 \times 0,4}{0,5348} = 0,3986$$

É evidente que este esquema de propagação de mensagens se aplica a qualquer número de variáveis e não apenas as 4 consideradas neste exemplo. $\Diamond$

A partir de considerações sobre o modo humano de processamento de informações Pearl [41] sugere que os arcos entre os nós da rede devem ser tratados como os únicos caminhos e "centros de ativações" que comandam e propagam o fluxo de dados no processo de solicitação e atualização das crenças. Por isso, interpreta-se cada nó como sendo um processador em separado, que ao mesmo tempo em que mantém os parâmetros da distribuição para a variável a ele associada, gerencia as ligações aferentes (que saem) e eferentes (que chegam) de nós vizinhos representantes de variáveis aleatórias conceitualmente relacionadas.

Cada processador pode interrogar a qualquer tempo a respeito da crença associada com os parâmetros dos seus vizinhos e compará-la em relação aos seus próprios parâmetros. Se as quantidades comparadas satisfazem algumas restrições locais (as marginais somam 1), nenhuma atividade é iniciada, caso contrário o nó responsável é ativado para resolver a diferença. Com isso, revisões similares nos nós vizinhos serão ativadas e um processo de propagação multi-dimensional será iniciado até que um ponto de equilíbrio (uma distribuição consistente) seja alcançado. Neste ponto, todos os nós da rede promoveram uma atualização de suas crenças.

3.3 Estado da Arte

Como já foi dito, o raciocínio probabilístico em redes bayesianas é entendido como um processo de atualização de crenças, isto é, "diz respeito à fusão e propagação do impacto de nova evidência e crenças, através da rede bayesiana, de forma tal que a cada proposição seja atribuída uma medida de certeza compatível com os axiomas da teoria da probabilidade" [41]. Doravante, o termo atualização de crenças é considerado sinônimo de inferência probabilística. As redes bayesianas são utilizadas, em sistemas de apoio a decisão, com o objetivo de fornecer estimativas atuais da crença $P(x|\mathbf{E})$, face a observação da evidência $\mathbf{E}$, para eventos que não são diretamente observáveis ou observáveis a custo não aceitável [34]. A evidência $\mathbf{E}$ pode ser específica (valor observado que afeta a crença nas outras variáveis) ou virtual (julgamento baseado em observações externas, mas que influenciam as variáveis na rede, por exemplo, um laudo de um laboratório contendo os resultados possíveis e as probabilidades associadas a cada um deles). Em ambos os casos a evidência representa variáveis com valores conhecidos.

As inferências no raciocínio probabilístico são do tipo diagnóstico, causal, intercausal (em ingês é conhecido por *explaining away*) ou misto (combinação dos tipos acima). Todas podem ser realizadas em redes bayesianas e correspondem a calcular $P(x_i|\mathbf{E})$, a partir da DPC ou utilizando um mecanismo que leva em conta as relações de independência condicional no DAG, obtendo uma redução da complexidade do problema, função da topologia do DAG. Um DAG pode conter ciclos (*loops*) se não se levar em conta à orientação dos arcos. Ao se propagar às evidências através dos ciclos é necessário que o mecanismo de inferência permita a realização de inferências bidirecionais e garanta a prevenção à realimentação e raciocínio circular do tipo: uma leve evidência em favor de A influencia a crença em favor de B, que por sua vez influencia

a crença em favor de A e assim por diante.

Em 1987, Gregory Cooper [7] demonstrou que o problema da inferência probabilística exata em redes bayesianas é da classe NP completo para as redes com topologia multiconectada, isto é, aquelas que contêm um par ou mais de variáveis conectadas por mais de um caminho não orientado. Paul Dagum e Michael Luby [12] demonstraram que a inferência probabilística aproximada em redes bayesianas também é da classe NP completo. Para redes simplesmente conectadas existem algoritmos com tempo polinomial, como é o caso do método de passagem de mensagens [41].

A topologia de uma rede bayesiana representa um modelo probabilístico completo de um domínio, com representação das informações qualitativas, relações de dependências e quantitativas, função de distribuição de probabilidades. Além disto, a topologia da rede determina o tipo de método utilizado na atualização das crenças. Assim, não adianta escolher um método específico como, por exemplo, o algoritmo de Pearl para realizar inferências, se o conhecimento envolvido no domínio do problema não pode ser representado na topologia especificada pelo método escolhido para atualização das probabilidades a *posteriori*. Em muitos problemas reais o domínio requer uma representação com redes multiconectadas. Ao se obter uma evidência, é preciso considerar se existe mais de um caminho entre o nó, com a nova evidência, e aquele cuja probabilidade deve ser atualizada. Em redes multiconectadas, uma simples propagação local não é aplicável porque tais algoritmos não prevêem a possibilidade de um nó, ao receber evidência de dois dos seus vizinhos, detectar se essa evidência não se origina na mesma fonte e, assim, evitar de considerar esta duas vezes.

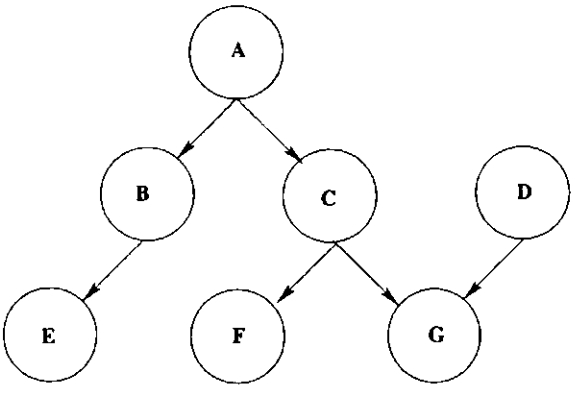


Figura 3.3: Rede bayesiana com topologia em árvore múltipla

O método Pearl de passagem de mensagens não é aplicável porque foi desenvolvido

para a topologia específica de árvore múltipla. A Figura 3.3 ilustra um exemplo de rede bayesiana com esta topologia. Os métodos de agregado e condicionantes, propostos por Pearl para a avaliação de redes multiconectadas, apresentam algumas dificuldades na transformação desses DAGs em grafos na topologia de árvores múltiplas. No primeiro caso, é necessário selecionar as variáveis que, ao serem substituídas por uma variável composta, permitem quebrar as conexões múltiplas do DAG. Quanto maior for esse grupo, maior será a dimensão da sua tabela de distribuição de probabilidades condicionais e maior a dificuldade para explicar a variação de crença devida a cada uma das variáveis. O método de condicionantes está baseado na habilidade de mudar a conectividade de uma rede, de forma a tornar esta simplesmente conectada, selecionando um conjunto adequado de variáveis a serem instanciadas. Nesses métodos, se a rede for muito conectada, pode ocorrer um grande número de combinações, o qual é função do número de nós e variáveis, a serem analisadas com o objetivo de quebrar os laços.

Lauritzen e Spiegelhalter [38] propuseram um método aplicável às redes bayesianas. O método tira partido da estrutura da rede original para propagar evidências, calculando probabilidades locais (envolvendo um pequeno número de variáveis) e evitando expressões globais (com grande número de variáveis). O método realiza um agrupamento de forma a transformar a estrutura da rede original em uma estrutura com topologia de árvore múltipla, mediante agrupamento de nós utilizando para isso a teoria de grafos.

Para isto, partindo do grafo original, segue-se os seguintes passos:

1. Realiza-se a triangularização do grafo (introdução de arcos em ciclos com mais de três nós. Os arcos são denominados cordas e o grafo resultante, grafo cordal ou triangular. Cada corda decompõe um ciclo em dois ciclos menores), adicionando-se nós em caso de necessidade,

2. Identifica-se todos os conjuntos de nós totalmente conectados (cliques);

3. Ordenam-se os cliques de forma que todos os nós comuns estejam em um clique anterior (seu pai)

4. É construído um novo grafo de forma que cada clique é um novo nó numa árvore múltipla de cliques

A realização de inferência é feita nesta árvore múltipla de cliques, obtendo a probabilidade conjunta de cada clique. Desta, é possível obter a probabilidade individual de cada variável presente no clique. Devido a complexidade envolvida, esse algoritmo é considerado ineficiente por Jensen et al. [29] [30].

A Tabela 3.1 apresenta as principais abordagens, com soluções aproximadas (A) ou exatas (E), para o problema da atualização de crenças. Esta Tabela foi obtida do Laboratório de IA do Instituto de Tecnologia da Força Aérea dos EUA. [1], sendo reproduzida aqui com o objetivo de chamar a atenção do leitor às diferentes abordagens utilizadas para realizar inferências, além da que é empregada neste texto, a qual será descrita na próxima Seção.

Tabela 3.2: Resumo das abordagens para o problema da atualização de crenças.

Referência	**Abordagens**	**E/A**
Becker 94	*Cut set*	E
Bouckaert 94	Simulação estocástica	A
Castilo 95	Enumeração	A
Cousins 91	Simualação estocástica	A
D'Ambrosio 94	Inferência simbólica	A
Darwiche 95	Condicionante	E/A
Delcher 95	Mensagens	E
Diez 96	Condicionante	E
Draper 95	Agregado	E
Hulme 95	Amostragem de Markov	A
Horvitz 89	Condicionante com limitante	E
Jensen 95	Árvore de Junção	A
Kanazawa 95	Simulação estocástica	A
Kjaerulff 95	Árvore de Junção	A
Lauritzen 88	Agregado	E
Li 94	Fatoração	E
Pearl 88	Passagem de mensagens	E
Santos 94	Enumeração	E
Shachter 89	Simulação estocástica	A
Santos 94	*Cut set*	E

3.4 Fusão de Influências e Propagação de Mensagens numa Rede Bayesiana

O mecanismo utilizado no raciocínio probabilístico proposto por Pearl [39], ficou conhecido como fusão de mensagens. No método proposto, o cálculo de $P(x_l|\mathbf{E})$ para um nó, X, e uma evidência $\mathbf{E}$, deve ser expresso como uma função envolvendo separadamente as probabilidades de X assumir um valor x_l, dado o estado de seus ascendentes e a probabilidade de seus descendentes terem assumido os valores apresentados em $\mathbf{E}$ dado x_l. Considerando a organização dos nós apresentados na Figura 3.4, deseja-se obter $p(\dot{e}|abcd)$ como função de $p(e|ab)$ e $p(cd|e)$. Esta forma precisa de expressar a crença de um nó é a origem do mecanismo de troca de mensagens.

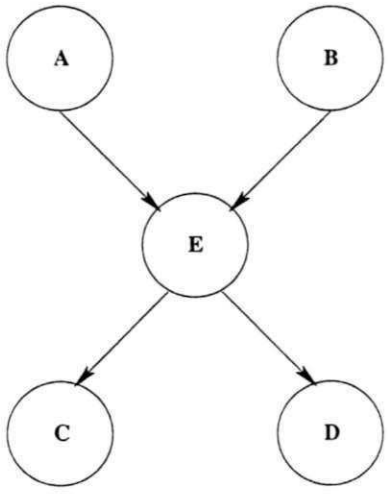


Figura 3.4: Rede bayesiana com 5 nós

Tem-se pelo Teorema de Bayes que:

$$p(e|abcd) = \frac{p(abcd|e)p(e)}{p(abcd)}$$

que pela regra da cadeia pode ser reescrito como

$$p(e|abcd) = \frac{p(e)p(ab|e)p(cd|abe)}{p(abcd)}$$

Novamente, pela regra da cadeia, segue-se que

$$p(e|abcd) = \frac{p(e)p(ab|e)p(cd|abe)}{p(ab)p(cd|ab)}$$

Em função da estrutura da rede, tem-se que $p(cd|abe) = p(cd|e)$. Assim,

$$p(e|abcd) = \frac{p(e)p(ab|e)}{p(ab)}p(cd|e)\frac{1}{p(cd|ab)}$$

Novamente, pelo Teorema de Bayes, tem-se que

$$p(e|abcd) = p(e|ab)p(cd|e)\frac{1}{p(cd|ab)}$$

Considerando que as probabilidades condicionais $p(e|ab)$ e $p(cd|e)$ possam ser atualizadas através de mensagens enviadas pelos nós ascendentes e descendentes sempre que ocorre uma evidência, pode-se dizer que a crença de um nó é obtida pela **fusão** destas mensagens e posterior produto das probabilidades condicionais por uma constante de normalização, $\frac{1}{p(ab|cd)}$, que assegura a geração de probabilidades consistentes. Fica então mais claro como se processa o mecanismo de atualização de crenças proposto. Ocorrido uma evidência, propaga-se para os **nós ascendentes uma revisão das probabilidades condicionais envolvendo este nós e aquele que recebeu a evidência, esta probabilidade é chamada de mensagem λ. Para os nós descendentes propaga-se uma revisão das probabilidades condicionais denominada mensagem π.** Os demais nós por sua vez, ao receber uma mensagem executam um procedimento de "fusão" a fim de atualizar seu vetor de crenças e propagar novas mensagens λ e π aos seus ascendentes e descendentes, respectivamente.

3.4.1 Algoritmo de Passagem de Mensagens de Pearl

Através da exploração de relações de independências entre variáveis, Pearl [39] [41], desenvolveu um algoritmo de passagem de mensagens para inferência em redes bayesianas. Dado um conjunto de valores, $\mathbf{a}$, de um subconjunto de variáveis instanciadas, $\mathcal{A}$, o algoritmo determina $P(x|\mathbf{a})$ para todos os valores x de cada uma das variáveis não instanciadas da rede. Este algoritmo começa enviando mensagens iniciadas em cada uma das variáveis instanciadas com direção aos seus vizinhos. Estes vizinhos, por

sua vez, passam mensagens aos seus próprios vizinhos. A atualização não depende da ordem na qual é iniciado a instanciação das variáveis, o qual significa que as variáveis podem ser instanciadas a medida que são obtidas as evidências.

Admita inicialmente que a rede bayesiana possui topologia de árvore múltipla, como mostrado na Figura 3.5, isto é, admita que a rede possui um único nó, denominado de nó raiz o qual não possui um nó pai e qualquer outro nó tem precisamente um único pai, sendo descendente do nó raiz. Normalmente, as direções dos arcos designam X como o conjunto de hipóteses e Y como o conjunto de conseqüências ou manifestações da hipóteses.

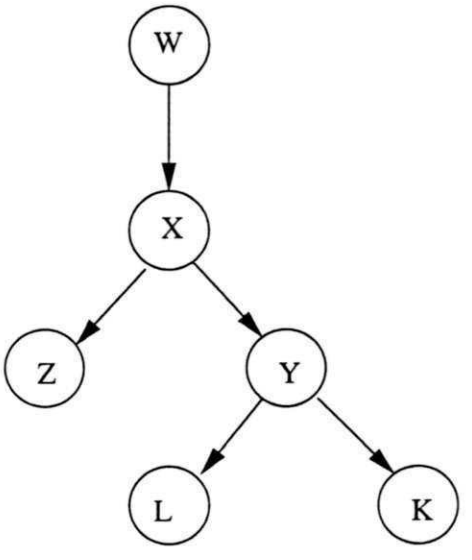


Figura 3.5: Rede bayesiana com um nó raiz

O algoritmo é baseado no Teorema 4.

Teorema 4 [44]

Seja $B = B_S + B_P$ uma rede bayesiana com topologia de árvore múltipla e seja $\mathbf{A}$ um subconjunto de variáveis aleatórias evidenciadas por valores pertencentes ao conjunto $\mathbf{a}$. Para cada variável X, defina as mensagens λ, os valores λ, as mensagens π e os valores π como segue:

1. *Definição das mensagens λ:*

   Para cada filho Y de X, para todos os valores x,

$$\lambda_Y(x) \equiv \sum_y p(y|x)\lambda(y)$$

2. *Definição dos valores* λ*:*

   Se $X \in \mathbf{A}$ *e o valor de* X *é* $\widehat{x}$*,*

$$\begin{aligned} \lambda(\widehat{x}) &\equiv 1 \\ \lambda(x) &\equiv 0, \end{aligned} \tag{3.2}$$

   para todo $x \neq \widehat{x}$

   Se $X \notin \mathbf{A}$ *e* X *é uma folha, para todos os valores* x*,*

$$\lambda(x) \equiv 1 \tag{3.3}$$

   Se $X \notin \mathbf{A}$ *e* X *não é uma folha, para todos os valores* x*,*

$$\lambda(x) \equiv \prod_{U \in CH_X} \lambda_U(x), \tag{3.4}$$

   onde CH_x *denota o conjunto de filhos de* X*.*

3. *Definição das mensagens* π*:*

   Se Z *é o pai de* X*, então para todos os valores* z*,*

$$\pi_X(z) \equiv \pi(z) \prod_{U \in CH_Z - \{X\}} \lambda_U(z). \tag{3.5}$$

4. *Definição dos valores* π*:*

   Se $X \in \mathbf{A}$ *e o valor de* X *é* $\widehat{x}$*,*

$$\begin{aligned} \pi(\widehat{x}) &\equiv 1 \\ \pi(x) &\equiv 0, \end{aligned} \tag{3.6}$$

para todo $x \neq \widehat{x}$

Se $X \notin \mathtt{A}$ *e* X *é a raiz, para todos os valores* x,

$$\pi(x) \equiv P(x). \tag{3.7}$$

Se $X \notin \mathtt{A}$, X *não é a raiz e* Z *é pai de* X, *para todos os valores* x,

$$\pi(x) \equiv \sum_z P(x|z)\pi_X(z). \tag{3.8}$$

5. *Dada as definições precedentes, para cada variável* X, *tem-se para todos os valores de* x,

$$P(x|\mathtt{a}) = \alpha\lambda(x)\pi(x), \tag{3.9}$$

onde α *é uma constante de normalização.*

Prova:

A prova deste Teorema foi obtida de [44] e se encontra reproduzida no Apêndice A caso o leitor deseje verificá-la.

■

A seguir, são apresentados os algoritmos baseados neste Teorema. Claramente, os algoritmos podem ser implementados em um programa orientado a objetos, no qual cada nó é um objeto que se comunica com outros nós passando mensagens λ e π. Contudo, ao invés de discutir detalhes de implementação, o melhor é mostrar os passos envolvidos nos algoritmos. Estes algoritmos são apresentados utilizando um projeto na arquitetura *top-down*.

Antes de apresentar os algoritmos, é importante observar como as rotinas pertencentes a este são chamadas. Por exemplo, a rotina `iniciar-arvore` é primeiro chamada como segue:

`iniciar-arvore`$((B_S,B_P),\mathtt{A},\mathtt{a},P(x|\mathtt{a}))$

Após esta chamada, A e a são ambos vazios e, para toda variáveis X, para todo valor x, $P(x|\mathtt{a})$ é a probabilidade condicional de x dado a e, sendo $\mathtt{a} = \emptyset$, é a probabilidade a priori de x. A cada vez, uma variável V é instanciada para o valor $\widehat{v}$ e a rotina `atualizar-arvore` é chamada como segue:

`atualizar-arvore`$((B_S,B_P),\mathtt{A},\mathtt{a},P(x|\mathtt{a}))$

Após esta chamada, V foi adicionado a A, $\widehat{v}$ foi adicionado a a e para toda variável X, $P(x|\mathtt{a})$ foi atualizado como sendo a probabilidade condicional de x dado um novo valor pertencente ao conjunto a. Segue o algoritmo.

Algoritmo 6 Inferência em Redes Bayesianas: Inicializar Rede[44]

Entrada Uma rede bayesiana em árvore múltipla, um subconjunto de variáveis a serem instanciadas, A, pelos valores pertencentes ao conjunto a.

SAIDA: Uma rede bayesiana atualizada.

void iniciar-arvore(rede bayesiana (B_S,B_P), um subconjunto de variáveis a serem instanciadas A, conjunto contendo as instâncias das variáveis a)

1: $\mathtt{A} = \emptyset; \mathtt{a} = \emptyset;$
2: **para** (cada $X \in B_S$) **faça**
3: **para** (cada valor x de X) **faça**
4: $\lambda(x) = 1$; // calcula os valores λ
5: **fim para**
6: **para** (o pai Z de X) **faça**
7: **para** (cada valor z de Z) **faça**
8: $\lambda_X(z) = 1$; // calcula as mensagens λ
9: **fim para**
10: **fim para**
11: **fim para**
12: **para** (cada valor r da raiz R) **faça**
13: $P(r|\mathtt{a}) = P(r)$; // calcula $P(r|\mathtt{a})$
14: $\pi(r) = P(r)$; // calcula os valores π de R
15: **fim para**
16: **para** (cada valor filho X de R) **faça**
17: $Envie - Msg - \pi(R, X)$;
18: **fim para**

Algoritmo 7 Inferência em Redes Bayesianas: Atualizar Rede[44]

Entrada Uma rede bayesiana em árvore múltipla, um subconjunto de variáveis a serem instanciadas, $\mathtt{A}$, pelos valores pertencentes ao conjunto $\mathtt{a}$, a variável $V \in \mathtt{A}$ e o seu valor $\widehat{v}$.

void atualizar-arvore(rede bayesiana (B_S,B_P), um subconjunto de variáveis a serem instanciadas $\mathtt{A}$, conjunto contendo as instâncias das variáveis $\mathtt{a}$, variável V, valor da variável $\widehat{v}$)

1: $\mathtt{A} = \mathtt{A} \cup \{V\}$; // adiciona V ao conjunto $\mathtt{A}$
2: $\mathtt{a} = \mathtt{a} \cup \{\widehat{v}\}$;
3: $\lambda(\widehat{v}) = 1$; $\pi(\widehat{v}) = 1$; $P(\widehat{v}|\mathtt{a}) = 1$;
// Instancia V ao valor de $\widehat{v}$.
4: **para** (cada valor de $v \neq \widehat{v}$) **faça**
5: $\lambda(v) = 0$; $\pi(v) = 0$;$P(v|\mathtt{a}) = 0$;
6: **fim para**
7: **se** ((V não é raiz) E (o pai Z de $V \notin \mathtt{A}$)) **então**
8: $Envie - Msg - \lambda(V, Z)$;
9: **fim se**
10: **para** (cada filho X de V tal que $X \notin \mathtt{A}$) **faça**
11: $Envie - Msg - \pi(V, X)$;
12: **fim para**

Algoritmo 8 Inferência em Redes Bayesianas: $Envie - Msg - \lambda$(nó Y, nó X) [44]

Entrada Uma rede bayesiana em árvore múltipla, um subconjunto de variáveis a serem instanciadas, $\mathtt{A}$, pelos valores pertencentes ao conjunto $\mathtt{a}$, a variável $V \in \mathtt{A}$ e o seu valor $\widehat{v}$.

void Envie-Msg-λ(nó Y, nó X)

1: **para** (cada valor de x) **faça**
2: $\lambda_Y(x) = \sum_y P(y|x)\lambda(y)$; // Y envia a X uma mensagem λ
3: $\lambda(x) = \prod_{U \in CH_x} \lambda_U(x)$; // calcula os valores λ de X
4: $P(x|\mathtt{a}) = \alpha\lambda(x)\pi(x)$; // calcula os valores $P(x|\mathtt{a})$.
5: **fim para**
6: *normalize* $P(x|\mathtt{a})$
7: **se** ((X não é raiz) E (o pai Z de $X \notin \mathtt{A}$)) **então**
8: $Envie - Msg - \lambda(X, Z)$;
9: **fim se**
10: **para** (cada filho W de X tal que $W \neq Y$ E $W \notin \mathtt{A}$) **faça**
11: $Envie - Msg - \pi(X, W)$;
12: **fim para**

Algoritmo 9 Inferência em Redes Bayesianas: $Envie - Msg - \pi$(nó Z, nó X) [44]

Entrada Uma rede bayesiana em árvore múltipla, um subconjunto de variáveis a serem instanciadas, $\mathtt{A}$, pelos valores pertencentes ao conjunto $\mathtt{a}$, a variável $V \in \mathtt{A}$ e o seu valor $\widehat{v}$.

void Envie-Msg-π(nó Z, nó X)

1: **para** (cada valor de z) **faça**
2: $\quad \pi_X(z) = \pi(z) \prod_{Y \in CH_Z - \{X\}} \lambda_Y(z)$; // Z envia a X uma mensagem π
3: **fim para**
4: **para** (cada valor de x) **faça**
5: $\quad \pi(x) = \sum_z P(x|z)\pi_X(z)$; // calcula os valores π de X
6: $\quad P(x|\mathtt{a}) = \alpha\lambda(x)\pi(x)$; // calcula os valores $P(x|\mathtt{a})$.
7: **fim para**
8: *normalize* $P(x|\mathtt{a})$
9: **para** (cada filho Y de X tal que $Y \notin \mathtt{A}$) **faça**
10: $\quad Envie - Msg - \pi(X, Y)$;
11: **fim para**

3.4.2 Aplicações do Algoritmo de Passagem de Mensagens de Pearl

O objetivo desta seção é promover um melhor entendimento do algoritmo de atualização de crenças adotado no desenvolvimento da ferramenta de apoio a decisão médica que será descrita no Capítulo 5.

Exemplo 3.3 [44]: Considere a rede bayesiana mostrada na Figura 3.6(a). A seguir, será mostrado os passos utilizados na inicialização da rede.

A chamada

`iniciar-arvore`$((B_S,B_P),\mathtt{A},\mathtt{a},P(x|\mathtt{a}))$

Resulta nos seguintes passos:

$$\mathtt{A} = \emptyset;$$

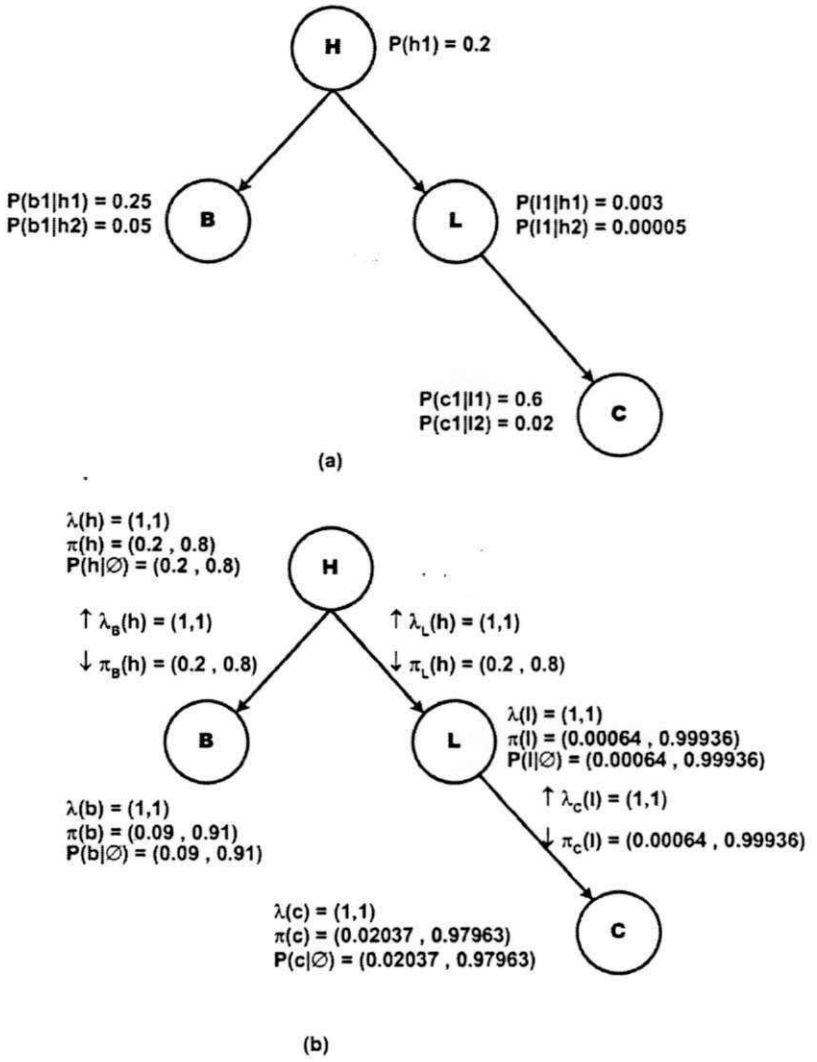


Figura 3.6: Em (b) é ilustrada a rede bayesiana de (a) inicializada.

$$a = \emptyset;$$

// calcula os valores λ

$$\lambda(h_1) = 1; \lambda(h_2) = 1;$$

$$\lambda(b_1) = 1; \lambda(b_2) = 1;$$

$$\lambda(l_1) = 1; \lambda(l_2) = 1;$$

$$\lambda(c_1) = 1; \lambda(c_2) = 1;$$

// calcula as mensagens λ

$$\lambda_B(h_1) = 1; \lambda_B(h_2) = 1;$$

$$\lambda_L(h_1) = 1; \lambda_L(h_2) = 1;$$

$$\lambda_C(l_1) = 1; \lambda_C(l_2) = 1;$$

// calcula os valores $p(h|\emptyset)$

$$p(h_1|\emptyset) = p(h_1) = 0,2;$$

$$p(h_2|\emptyset) = p(h_2) = 0,8;$$

// calcula os valores π de H

$$\pi(h_1) = p(h_1) = 0,2;$$

$$\pi(h_2) = p(h_2) = 0,8;$$

$$Envie - Msg - \pi(H, B);$$

$$Envie - Msg - \pi(H, L);$$

A chamada

`Envie-Msg-`$\pi(H, B)$;

Resulta nos seguintes passos:
// H envia a B uma mensagem π

$$\pi_B(h_1) = \pi(h_1)\lambda_L(h_1) = (0,2)(1) = 0,2$$

$$\pi_B(h_2) = \pi(h_2)\lambda_L(h_2) = (0,8)(1) = 0,8$$

// calcula os valores π de B

$$\pi(b_1) = p(b_1|h_1)\pi_B(h_1) + p(b_1|h_2)\pi_B(h_2)$$
$$= (0,25)(0,2) + (0,05)(0,8) = 0,09;$$

$$\pi(b_2) = p(b_2|h_1)\pi_B(h_1) + p(b_2|h_2)\pi_B(h_2)$$
$$= (0,75)(0,2) + (0,95)(0,8) = 0,91;$$

// calcula os valores $p(b|\emptyset)$

$$p(b_1|\emptyset) = \alpha\lambda(b_1)\pi(b_1) = \alpha(1)(0,09) = 0,09\alpha;$$

$$p(b_2|\emptyset) = \alpha\lambda(b_2)\pi(b_2) = \alpha(1)(0,91) = 0,91\alpha;$$

$$p(b_1|\emptyset) = \frac{0,09\alpha}{0,09\alpha + 0,91\alpha} = 0,09;$$

$$p(b_2|\emptyset) = \frac{0,91\alpha}{0,09\alpha + 0,91\alpha} = 0,91;$$

A chamada

`Envie-Msg-`$\pi(H, L)$;

Resulta nos seguintes passos:
// H envia a L uma mensagem π

$$\pi_L(h_1) = \pi(h_1)\lambda_B(h_1) = (0,2)(1) = 0,2$$

$$\pi_L(h_2) = \pi(h_2)\lambda_B(h_2) = (0,8)(1) = 0,8$$

// calcula os valores π de L

$$\pi(l_1) = p(l_1|h_1)\pi_L(h_1) + p(l_1|h_2)\pi_L(h_2)$$
$$= (0,003)(0,2) + (0,00005)(0,8) = 0,00064;$$

$$\pi(l_2) = p(l_2|h_1)\pi_L(h_1) + p(l_2|h_2)\pi_L(h_2)$$
$$= (0,997)(0,2) + (0,99995)(0,8) = 0,99936;$$

// calcula os valores $p(l|\emptyset)$

$$p(l_1|\emptyset) = \alpha\lambda(l_1)\pi(l_1) = \alpha(1)(0,00064) = 0,00064\alpha;$$

$$p(l_2|\emptyset) = \alpha\lambda(l_2)\pi(l_2) = \alpha(1)(0,99936) = 0,99936\alpha;$$

$$p(l_1|\emptyset) = \frac{0,00064\alpha}{0,00064\alpha + 0,99936\alpha} = 0,00064;$$

$$p(l_2|\emptyset) = \frac{0,99936\alpha}{0,00064\alpha + 0,99936\alpha} = 0,99936;$$

$$Envie - Msg - \pi(L, C);$$

A chamada

`Envie-Msg-`$\pi(L, C)$;

Resulta nos seguintes passos:
// L envia a C uma mensagem π

$$\pi_C(l_1) = \pi(l_1) = 0,00064;$$

$$\pi_C(l_2) = \pi(l_2) = 0,99936;$$

// calcula os valores π de C

$$\pi(c_1) = p(c_1|l_1)\pi_C(l_1) + p(c_1|l_2)\pi_C(l_2)$$
$$= (0,6)(0,00064) + (0,02)(0,99936) = 0,02037;$$

$$\pi(c_2) = p(c_2|l_1)\pi_C(l_1) + p(c_2|l_2)\pi_C(l_2)$$
$$= (0,40)(0,00064) + (0,98)(0,99936) = 0,97963;$$

// calcula os valores $p(c|\emptyset)$

$$p(c_1|\emptyset) = \alpha\lambda(c_1)\pi(c_1) = \alpha(1)(0,02037) = 0,02037\alpha;$$

$$p(c_2|\emptyset) = \alpha\lambda(c_2)\pi(c_2) = \alpha(1)(0,97963) = 0,97963\alpha;$$

$$p(c_1|\emptyset) = \frac{0,02037\alpha}{0,02037\alpha + 0,97963\alpha} = 0,02037;$$

$$p(l_2|\emptyset) = \frac{0,97963\alpha}{0,02037\alpha + 0,97963\alpha} = 0,97963;$$

A inicialização está completa. A rede inicializada é mostrada na Figura 3.6(b).$\Diamond$

O Exemplo 3.4 analisa os passos envolvidos na atualização de crenças quando uma variável da rede é instanciada a um valor particular.

Exemplo 3.4 [44]: Considere novamente a rede bayesiana mostrada na Figura 3.6(a). Admita que B é instanciado a b_1. A seguir, serão mostrados os passos no algoritmo quando os valores da rede são atualizados de acordo com esta instanciação.

A chamada

`atualizar-arvore`$((B_S,B_P),\mathtt{A},\mathtt{a},P(x|\mathtt{a}))$

Resulta nos seguintes passos:

$$\mathtt{A} = \emptyset \cup \{B\} = \{B\};$$
$$\mathtt{a} = \emptyset \cup \{b_1\} = \{b_1\};$$

$$\lambda(b_1) = 1; \pi(b_1) = 1; p(b_1|\{b_1\}) = 1;$$
$$\lambda(b_2) = 0; \pi(b_2) = 0; p(b_2|\{b_1\}) = 0;$$

$$Envie - Msg - \lambda(B,H);$$

A chamada

`Envie-Msg-`$\lambda(B,H)$;

Resulta nos seguintes passos:

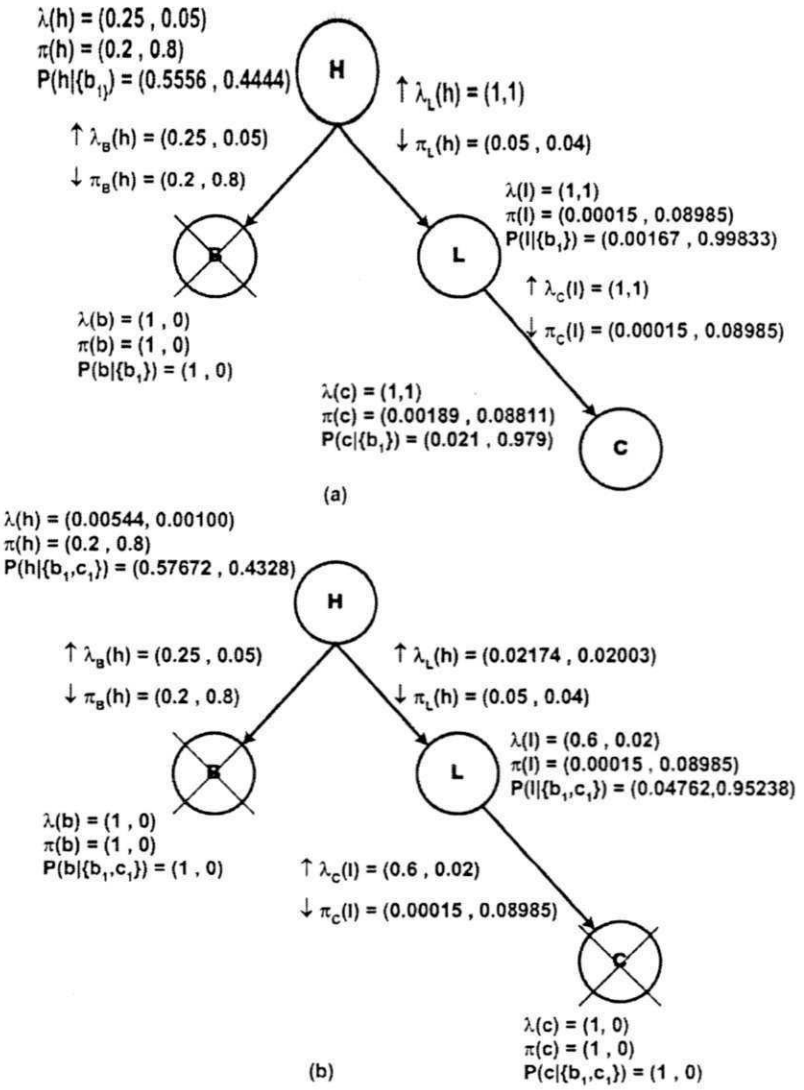


Figura 3.7: Em (b) é ilustrada a rede bayesiana de (a) atualizada.

// B envia a H a mensagem λ

$$\lambda_B(h_1) = p(b_1|h_1)\lambda(b_1) + p(b_2|h_1)\lambda(b_2)$$
$$= (0,25)(1) + (0,75)(0) = 0,25$$

$$\lambda_B(h_2) = p(b_1|h_2)\lambda(b_1) + p(b_2|h_2)\lambda(b_2)$$
$$= (0,05)(1) + (0,95)(0) = 0,05$$

// calcula os valores λ de H

$$\lambda(h_1) = \lambda_B(h_1)\lambda_L(h_1) = (0,25)(1) = 0,25;$$

$$\lambda(h_2) = \lambda_B(h_2)\lambda_L(h_2) = (0,05)(1) = 0,05;$$

// calcula os valores $P(h|\{b_1\})$

$$p(h_1|\{b_1\}) = \alpha\lambda(h_1)\pi(h_1) = \alpha(0,25)(0,20) = 0,05\alpha;$$

$$p(h_2|\{b_1\}) = \alpha\lambda(h_2)\pi(h_2) = \alpha(0,05)(0,8) = 0,04\alpha;$$

$$p(h_1|\{b_1\}) = \frac{0,05\alpha}{0,05\alpha + 0,04\alpha} = 0,5556;$$

$$p(h_2|\{b_1\}) = \frac{0,04\alpha}{0,05\alpha + 0,04\alpha} = 0,4444;$$

$$Envie - Msg - \pi(H, L);$$

A chamada

`Envie-Msg-`$\pi(H, L)$;

Resulta nos seguintes passos:
// H envia a L uma mensagem π

$$\pi_L(h_1) = \pi(h_1)\lambda_B(h_1) = (0,2)(0,25) = 0,05;$$

$$\pi_L(h_2) = \pi(h_2)\lambda_B(h_2) = (0,8)(0,05) = 0,04;$$

// calcula os valores π de L

$$\pi(l_1) = p(l_1|h_1)\pi_L(h_1) + p(l_1|h_2)\pi_L(h_2)$$
$$= (0,003)(0,05) + (0,00005)(0,04) = 0,00015;$$

$$\pi(l_2) = p(l_2|h_1)\pi_L(h_1) + p(l_2|h_2)\pi_L(h_2)$$
$$= (0,997)(0,05) + (0,99995)(0,04) = 0,08985;$$

// calcula os valores $p(l|\{b_1\})$

$$p(l_1|\{b_1\}) = \alpha\lambda(l_1)\pi(l_1) = \alpha(1)(0,00015) = 0,00015\alpha;$$

$$p(l_2|\{b_1\}) = \alpha\lambda(l_2)\pi(l_2) = \alpha(1)(0,08985) = 0,08985\alpha;$$

$$p(l_1|\{b_1\}) = \frac{0,00015\alpha}{0,00015\alpha + 0,08985\alpha} = 0,00167;$$

$$p(l_2|\{b_1\}) = \frac{0,08985\alpha}{0,00015\alpha + 0,08985\alpha} = 0,99936;$$

$$Envie - Msg - \pi(L, C);$$

A chamada

`Envie-Msg-`$\pi(L, C)$;

Resulta nos seguintes passos:
// L envia a C uma mensagem π

$$\pi_C(l_1) = \pi(l_1) = 0,00015;$$

$$\pi_C(l_2) = \pi(l_2) = 0,08985;$$

// calcula os valores π de C

$$\pi(c_1) = p(c_1|l_1)\pi_C(l_1) + p(c_1|l_2)\pi_C(l_2)$$
$$= (0,6)(0,00015) + (0,02)(0,08985) = 0,00189;$$

$$\pi(c_2) = p(c_2|l_1)\pi_C(l_1) + p(c_2|l_2)\pi_C(l_2)$$
$$= (0,40)(0,00015) + (0,98)(0,08985) = 0,08811;$$

// calcula os valores $p(c|\{b_1\})$

$$p(c_1|\{b_1\}) = \alpha\lambda(c_1)\pi(c_1) = \alpha(1)(0,00189) = 0,00189\alpha;$$

$$p(c_2|\{b_1\}) = \alpha\lambda(c_2)\pi(c_2) = \alpha(1)(0,08811) = 0,08811\alpha;$$

$$p(c_1|\{b_1\}) = \frac{0,00189\alpha}{0,00189\alpha + 0,08811\alpha} = 0,021;$$

$$p(c_2|\{b_1\}) = \frac{0,08811\alpha}{0,00189\alpha + 0,08811\alpha} = 0,979;$$

A rede foi atualizada na Figura 3.7(a). Observe que a agora podemos realizar inferências de posse dos valores atualizados das probabilidades a *posteriori* calculadas pelo algoritmo de passagem de mensagens de Pearl. O próximo exemplo ilustra uma nova instanciação sobre uma outra variável da rede e seu "impacto" nas probabilidades a *posteriori*. ◊

Exemplo 3.5 [44]: Considere novamente a rede bayesiana mostrada na Figura 3.6(a). Admita que B já foi instanciado a b_1, e agora C instanciado para c_1. A seguir, serão mostrados os passos no algoritmo quando os valores da rede são atualizados de acordo com a instanciação.

A chamada

`atualizar-arvore`$((B_S,B_P)$,`A`,`a`,C, $c_1)$;

Resulta nos seguintes passos:

$$\mathtt{A} = \{B\} \cup \{C\} = \{B, C\};$$
$$\mathtt{a} = \{b_1\} \cup \{c_1\} = \{b_1, c_1\};$$

$$\lambda(c_1) = 1; \pi(c_1) = 1; p(c_1|\{b_1, c_1\}) = 1;$$
$$\lambda(c_2) = 0; \pi(c_2) = 0; p(c_2|\{b_1, c_1\}) = 0;$$

$$Envie - Msg - \lambda(C, L);$$

A chamada

`Envie-Msg-`$\lambda(C, L)$;

Resulta nos seguintes passos:
// C envia a L a mensagem λ

$$\lambda_C(l_1) = p(c_1|l_1)\lambda(c_1) + p(c_2|l_1)\lambda(c_2)$$
$$= (0,6)(1) + (0,4)(0) = 0,60$$

$$\lambda_C(l_2) = p(c_1|l_2)\lambda(c_1) + p(c_2|l_2)\lambda(c_2)$$
$$= (0,02)(1) + (0,98)(0) = 0,02$$

// calcula os valores λ de L

$$\lambda(l_1) = \lambda_C(l_1) = 0,60;$$
$$\lambda(l_2) = \lambda_C(l_2) = 0,02;$$

// calcula os valores $p(l|\{b_1, c_1\})$

$$p(l_1|\{b_1, c_1\}) = \alpha\lambda(l_1)\pi(l_1) = \alpha(0,60)(0,00015) = 0,00009\alpha;$$

$$p(l_2|\{b_1, c_1\}) = \alpha\lambda(l_2)\pi(l_2) = \alpha(0,02)(0,08985) = 0,00180\alpha;$$

$$p(l_1|\{b_1, c_1\}) = \frac{0,00009\alpha}{0,00009\alpha + 0,00180\alpha} = 0,04762;$$

$$p(l_2|\{b_1, c_1\}) = \frac{0,00180\alpha}{0,00009\alpha + 0,00180\alpha} = 0,95238;$$

$$Envie - Msg - \lambda(L, H);$$

A chamada

`Envie-Msg-`$\lambda(L, H)$;

Resulta nos seguintes passos:
// L envia a H uma mensagem λ

$$\lambda_L(h_1) = p(l_1|h_1)\lambda(l_1) + p(l_2|h_1)\lambda(l_2)$$

$$= (0,003)(0,6) + (0,997)(0,02) = 0,02174$$

$$\lambda_L(h_2) = p(l_1|h_2)\lambda(l_1) + p(l_2|h_2)\lambda(l_2)$$

$$= (0,00005)(0,6) + (0,99995)(0,02) = 0,02003;$$

// calcula os valores λ de H

$$\lambda(h_1) = \lambda_B(h_1)\lambda_L(h_1) = (0,25)(0,02174) = 0,00544;$$

$$\lambda(h_2) = \lambda_B(h_2)\lambda_L(h_2) = (0,05)(0,02003) = 0,00100;$$

// calcula os valores $p(h|\{b_1, c_1\})$

$$p(h_1|\{b_1, c_1\}) = \alpha\lambda(h_1)\pi(h_1) = \alpha(0,00544)(0,2) = 0,00109\alpha;$$

$$p(h_2|\{b_1, c_1\}) = \alpha\lambda(h_2)\pi(h_2) = \alpha(0,00100)(0,8) = 0,00080\alpha;$$

$$p(h_1|\{b_1, c_1\}) = \frac{0,00109\alpha}{0,00080\alpha + 0,00109\alpha} = 0,57672;$$

$$p(h_2|\{b_1, c_1\}) = \frac{0,00080\alpha}{0,00080\alpha + 0,00109\alpha} = 0,42328;$$

A rede foi atualizada na Figura 3.7(b). $\Diamond$

O objetivo da próxima Seção é localizar o algoritmo de propagação de mensagens numa classe mais geral de algoritmos denominada classe dos algoritmos da Lei Distributiva Generalizada. A justificativa para esta discussão é melhorar a compreensão do algoritmo de Pearl.

3.5 Lei Distributiva Generalizada

A Lei Distributiva primitiva, na sua forma mais simples, afirma que $ab + ac = a(b + c)$. O lado esquerdo desta equação envolve três operações aritméticas (uma adição e duas multiplicações); contudo, o lado direito necessita de somente duas operações. Então, a lei distributiva fornece-nos um "algoritmo rápido" para calcular $ab + ac$. O objetivo

desta seção é demonstrar que a lei distributiva pode ser vastamente generalizada e que esta generalização nos leva a uma ampla família de algoritmos rápidos, incluindo o algoritmo de Viterbi e a transformada rápida de Fourier, além do algoritmo de "propagação de probabilidade" de Pearl. Para fornecer uma idéia mais clara do potencial da Lei Distributiva Generalizada e introduzir a perspectiva que será adotada nesta seção, considere o seguinte exemplo [49].

Exemplo 3.6 [49]: Seja $f(x, y, w)$ e $g(x, z)$ funções reais, onde x, y, z e w são variáveis assumindo valores num conjunto finito A com q elementos. Admita a necessidade da construção de tabelas de valores $\alpha(x, w)$ $\beta(y)$, definidas como:

$$\alpha(x, w) \doteq \sum_{y,z \in A} f(x, y, w) g(x, z) \tag{3.10}$$

$$\beta(y) \doteq \sum_{x,z,w \in A} f(x, y, w) g(x, z) \tag{3.11}$$

Então, pode ser dito que $\alpha(x, w)$ é obtido na Equação (3.10) "marginalizando" as variáveis y e z a partir do produto $f(x, y, w) \cdot g(x, z)$. De forma semelhante, $\beta(y)$ é obtido marginalizando x,z e w a partir da mesma função. Uma opção pelo caminho óbvio exigiria $2q^4$ operações para o cálculo de $\alpha(x, w)$, pois para cada um dos q^2 valores de (x, w) existem q^2 termos na soma definindo $\alpha(x, w)$, com cada um deles requerendo uma adição e uma multiplicação. Também o cálculo de $\beta(y)$ requer $2q^4$ operações, totalizando $4q^4$ operações aritméticas.

Optando pelo uso da Lei Distributiva, a soma definida na Equação (3.10) pode ser fatorada de acordo com:

$$\alpha(x, w) = \left(\sum_{y \in A} f(x, y, w) \right) \left(\sum_{z \in A} g(x, z) \right). \tag{3.12}$$

Utilizando este fato, é possível simplificar o cálculo de $\alpha(x, w)$. Para isto, é necessário a construção de tabelas de valores das funções $\alpha_1(x, w)$ e $\alpha_2(x)$, definidas como:

$$\alpha_1(x, w) \doteq \sum_{y \in A} f(x, y, w) \tag{3.13}$$

$$\alpha_2(x) \doteq \sum_{z \in A} g(x,z), \tag{3.14}$$

as quais requerem $q^3 + q^2$ adições. Logo, é possível calcular os q^2 valores de $\alpha(x,w)$ utilizando a fórmula:

$$\alpha(x,w) = \alpha_1(x,w) \cdot \alpha_2(x), \tag{3.15}$$

a qual requer q^2 multiplicações. Portanto, explorando a Lei Distributiva, foi possível reduzir o número total das operações requeridas para o cálculo de $\alpha(x,w)$ de $2q^4$ para $q^3 + 2q^2$. A Equação (3.11) pode ser escrita como:

$$\beta(y) = \sum_{x,w \in A} f(x,y,w) \left(\sum_{z \in A} g(x,z) \right) = \sum_{x,w \in A} f(x,y,w)\alpha_2(x). \tag{3.16}$$

A vantagem de realizar primeiro a construção da tabela de valores de $\alpha_2(x)$ (q^2 operações), para depois utilizarmos (3.16) ($2q^3$ operações adicionais) permite reduzir o número de operações necessárias para $\beta(y)$ de $2q^4$ operações iniciais para $2q^3 + q^2$ exigidas quando se aplica a Lei Distributiva. Adicionalmente, partindo-se de (3.15) e de (3.16) tanto a obtenção de $\alpha(x,w)$ quanto de $\beta(y)$ exige a construção da tabela de valores de $\alpha_2(x)$ uma única vez, reduzindo o número total de operações necessárias a obtenção de $\alpha(x,w)$ e de $\beta(y)$ para $3q^3 + 2q^2$, o qual representa uma economia quando é comparado ao número de operações requeridas na opção do caminho óbvio,$4q^4$. $\Diamond$

A simplificação no Exemplo 3.6 era fácil de acompanhar e os ganhos obtidos a partir das simplificações eram relativamente modestos. Contudo, em casos mais complicados, pode ser muito difícil visualizar a melhor forma de organizar os cálculos, mas a economia computacional pode ser alta. O objetivo desta seção é mostrar que os problemas do tipo descrito no Exemplo 3.6 tem uma faixa larga de aplicabilidade, além de descrever um procedimento geral, denominado de Lei Distributiva Generalizada (LDG), para solucionar estes problemas eficientemente. A LDG executa estes objetivos através da *passagem de mensagens em redes de comunicação cujo grafo básico é uma árvore.* Para facilitar o entendimento da LDG, esta seção será dividida em duas subseções. Assim, na Seção 3.5.1, será posicionado o problema de cálculo geral denominado de problema "Função Produto marginalizada" (FPM), o qual inclui a inferência probabilística em

redes bayesianas. Na seção 3.5.2, será fornecido uma algoritmo preciso para solucionar o problema FPM, denominado de Lei Distributiva Generalizada que, em geral, oferece uma solução eficiente.

3.5.1 Função Produto Marginalizada

A LDG pode reduzir em muito o número de adições e multiplicações requeridas por uma certa classe de problemas computacionais. É relevante ressaltar que muito do potencial da LDG é devido ao fato que ela é aplicável a situações nas quais as noções de adição e multiplicação são, por elas mesmas, generalizadas. O arcabouço do trabalho para esta generalização é a estrutura algébrica denominada semi-anel comutativo.

Definição 12 (Semi-anel Comutativo)

- *Um semi-anel comutativo é um conjunto $\mathcal{K}$, associado a duas operações binárias chamadas "+" e "·", as quais satisfazem os três axiomas seguintes:*

  1. *A operação "+" é associativa e comutativa e existe o elemento de identidade aditivo, chamado "0", tal que $k+0=k$, para todo $k \in \mathcal{K}$. Este axioma torna (k , +) um monóide comutativo.*
  2. *A operação "·" também é associativa e comutativa e existe o elemento de identidade multiplicativo, chamado "1", tal que $k \cdot 1 = k$, para todo $k \in \mathcal{K}$ Este axioma torna (k, ·) um monóide comutativo.*
  3. *A Lei Distributiva está assegurada, ou seja, $(a \cdot b) + (a \cdot c) = a \cdot (b+c)$, para todas as triplas (a,b,c) a partir de $\mathcal{K}$.*

A diferença entre semi-anel e anel é que em um semi-anel, o inverso aditivo necessariamente não existe, isto é, (k , +) é apenas um monóide, não um grupo. Portanto, todo anel comutativo é também um semi-anel comutativo. Sendo assim, por exemplo, o conjunto dos reais ou números complexos, com adição e multiplicação ordinária, forma um semi-anel comutativo. Da mesma forma, o conjunto dos polinômios em uma ou mais variáveis sobre qualquer anel comutativo forma um semi-anel comutativo. Contudo, há diversos outros semi-anéis, como os que estão resumidos na Tabela 3.3, onde foi assumido que nos semi-anéis 4-8, o conjunto $\mathcal{K}$ é um intervalo de números reais com a adição possível de $\pm\infty$.

Tabela 3.3: Alguns semi-anéis comutativos.

	$\mathcal{K}$	"(+,0)"	"(·,1)"	nome abreviado
1.	$\mathcal{A}$	(+,0)	(·,1)	
2.	$\mathcal{A}$[x]	(+,0)	(·,1)	
3.	$\mathcal{A}$[x,y,...]	(+,0)	(·,1)	
4.	[0,∞)	(+,0)	**(·,1)**	soma-produto
5.	(0,∞]	(min,∞)	(·,1)	min-produto
6.	[0,∞)	(max,0)	(·,1)	max-produto
7.	(−∞,∞]	(min,∞)	(+,0)	min-soma
8.	[−∞,∞)	(max,−∞)	(+,0)	max-soma
9.	{0,1}	(OR,0)	(AND,1)	booleano

Aqui $\mathcal{A}$ denota um anel comutativo arbitrário; enquanto $\mathcal{A}$[x] e $\mathcal{A}$[x,y,...] são conjuntos de polinômios nas variáveis x e $x, y, ...$, com coeficientes em $\mathcal{A}$.

Por exemplo, considere o semi-anel min-soma na Tabela 3.3. Neste caso, $\mathcal{K}$ é o conjunto de números reais, mais o símbolo especial "∞". A operação "+" é definida como a operação de *tomar o mínimo*, com o símbolo ∞ correspondendo ao elemento identidade, ísto é, $\min(k, \infty) = k$ para todo $k \in \mathcal{K}$. A operação "·" é definida como sendo *adição ordinária*, sendo 0 o elemento identidade, mas $k + \infty = \infty$, para todo k. Ocasionalmente, esta combinação forma um semi-anel, por causa da Lei Distributiva que é equivalente a $min(a + b, a + c) = a + min(b, c)$.

Tendo brevemente discutido os semi-anéis comutativos, será agora descrita uma função produto marginalizada, que é o problema geral resolvido pela LDG. Até o fim desta seção, serão fornecidos diversos exemplos de problema FPM, os quais demonstram como este pode ocorrer numa variedade larga de cálculos.

Seja $x_1, x_2, ..., x_n$ variáveis assumindo valores nos conjuntos finitos $A_1, A_2, ..., A_n$, com $|A_i| = q_i$, para $i = 1, 2, ..., n$. Seja $S = \{i_1, i_2, ..., i_r\}$ um subconjunto de $\{1, 2, ..., n\}$, denote o produto $A_{i_1} \times A_{i_1} \times ... \times A_{i_r}$ por A_S, a lista de variáveis $(x_{i_1}, ..., x_{i_r})$ por x_S e a cardinalidade de A_S, isto é, $|A_S|$, por q_S. Denote o produto $A_{\{1,...,n\}}$ simplesmente por $\mathbf{A}$ e a lista de variáveis $\{x_1, ..., x_n\}$ simplesmente por $\mathbf{x}$.

Seja agora $S = \{S_1, ..., S_M\}$ denotar M subconjuntos de $\{1, ..., n\}$. Admita que para cada um dos $i = 1, 2, ..., M$, existe uma função $\alpha_i : \mathbf{A}_{S_i} \rightarrow \mathbf{R}$, onde $\mathbf{R}$ é um anel comutativo. As listas de variáveis x_{S_i} são denominadas de "domínios locais" e a lista de funções α_i são chamadas de "kernels locais". Um kernel global $\beta : \mathbf{A} \rightarrow \mathbf{R}$ é

definido como segue:

$$\beta(x_1, x_2, ..., x_n) = \prod_{i=1}^{M} \alpha_i(x_{S_i}). \tag{3.17}$$

Com este *setup*, o problema FPM é este: Para cada um ou mais índices $i = 1, ..., M$, calcule uma tabela de valores da marginalização-S_i do kernel global β, o qual é uma função $\beta_i : \mathbf{A}_{S_i} \rightarrow \mathbf{R}$, definida por:

$$\beta_i(x_{S_i}) = \sum_{x^c_{S_i} \in A^c_{S_i}} \beta(x). \tag{3.18}$$

Na Equação (3.18), S_i^c denota o complemento do conjunto S_i relativo ao "universo" $\{1, ..., n\}$. Por exemplo, se n=4 e se $S_i = \{1, 4\}$, então:

$$\beta_i(x_1, x_4) = \sum_{x_2 \in A_2, x_3 \in A_3} \beta(x_1, x_2, x_3, x_4). \tag{3.19}$$

A função $\beta_i(x_{S_i})$ definida na Equação (3.17) é denominada *i-ésima* função objetiva ou função objetiva em S_i. É possível observar que o cálculo da *i-ésima* função objetiva na forma usual requer $q_1 q_1 ... q_n$ adições e $(M-1)q_1 q_1 ... q_n$ multiplicações, totalizando $M \cdot q_1 q_1 ... q_n$ operações aritméticas, sendo q_i o tamanho do conjunto A_i. Na seção 3.5.2, será ilustrado que o algoritmo denominado "Lei Distributiva Generalizada" pode reduzir a complexidade deste cenário. O próximo Exemplo, avalia o problema FPM presente na inferência em redes bayesianas.

Exemplo 3.7 (Rede bayesiana vista como uma aplicação da LDG) [49]: Considera a rede bayesiana esboçada na Figura 3.8. Numa rede bayesiana, os "pais" de um nó V são aqueles vértices (se existirem), os quais estão posicionados imediatamente "acima" de V. Logo, na Figura 3.8 $\mathbf{Pa}_A = \{B, E\}$ e $\mathbf{Pa}_B = \emptyset$. Lembrando que existe uma variável aleatória associada com cada um dos vértices, e assumindo que cada uma das variáveis aleatórias depende somente de seus "pais", temos a fatoração da função densidade de probabilidade conjunta de acordo com:

$$Pr\{B = b, E = e, A = a, R = r, W = w\} = Pr\{B = b\} Pr\{E = e\} \cdot$$

$$\cdot Pr\{A = a | B = b, E = e\} Pr\{R = r | E = e\} Pr\{W = w | A = a\}$$

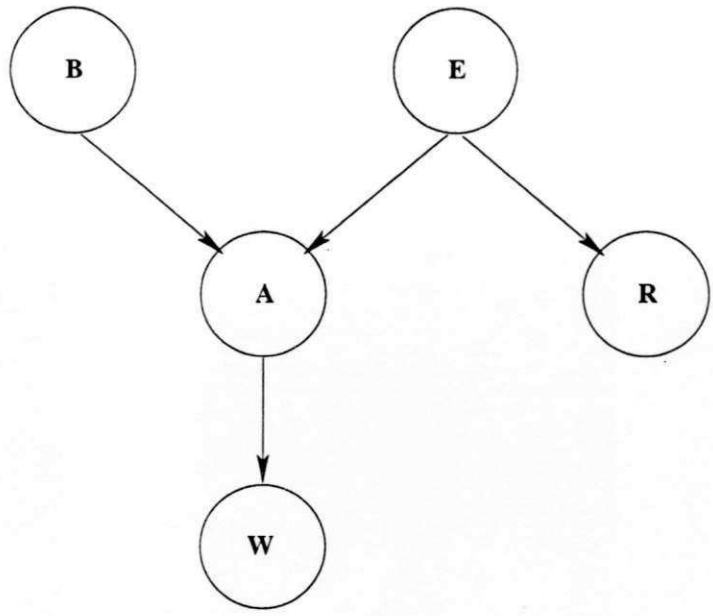


Figura 3.8: Rede Bayesiana no Exemplo 3.6

Ou,

$$p(b, e, a, r, w) = p(b)p(e)p(a|b, e)p(r|e)p(w|a). \tag{3.20}$$

Admita que as duas variáveis aleatórias **W** e **R** são observadas assumindo os valores w_0 e r_0, respectivamente. O problema da inferência probabilística, neste caso, consiste em calcular as probabilidades condicionais de uma ou mais variáveis aleatórias restantes, ou seja, **B**, **E** e **A**, estando o condicionamento associado as evidências $\{R = r_0, W = w_0\}$. Agora, considere o problema FPM tendo domínios e *kernels* locais dados pela Tabela 3.4

Então, da Equação (3.20) (utilizando o semi-anel 4 da Tabela 3.3, isto é, o conjunto dos números reais não-negativos com adição e multiplicação ordinárias) o kernel global $F(b, e, a)$ é somente a função $p(b, e, a, r, w)$, de forma que, por exemplo, a função objetiva no domínio local 1 é:

$$F_1(b) = \sum_{e,a} p(b, e, a, r_0, w_0) = p(b, r_0, w_0). \tag{3.21}$$

Aplicando a regra de Bayes:

Tabela 3.4: domínios e *kernels* locais do Exemplo 2.4

	domínio local	*kernel* local
1.	$\{b\}$	$p(b)$
2.	$\{e\}$	$p(e)$
3.	$\{a, b, e\}$	$p(a\|b, e)$
4.	$\{a\}$	$p(w_0\|a)$
5.	$\{e\}$	$p(r_0\|e)$

$$p(b|r_0, w_0) = \frac{p(b, r_0, w_0)}{p(w_0, r_0)}. \tag{3.22}$$

Sendo assim, a probabilidade de B, dada a "evidência" (r_0, w_0), é:

$$Pr\{B = b|R = r_0, W = w_0\} = p(b|r_0, w_0) = \alpha F_1(b), \tag{3.23}$$

onde a constante de proporcionalidade α é dada por:

$$\alpha = \left(\sum_b F_1(b)\right)^{-1}. \tag{3.24}$$

Da mesma forma, o cálculo das probabilidades condicionais de **A** e **E** podem ser avaliado via análise das funções objetiva nos domínios locais 4 e 5, respectivamente. Desta forma, o problema da inferência probabilística em redes bayesianas é um caso especial do problema FPM. Na seção 3.5.2 será mostrado que a LDG quando aplicada a problemas deste tipo, resulta num algoritmo equivalente ao algoritmo de "propagação de probabilidade" de Pearl.◊

3.5.2 Solução do Problema FPM: LDG

Se os elementos de S guardam entre si um relacionamento especial, então um algoritmo para solução do problema FPM pode ser baseado na noção de "passagem de mensagens". O relacionamento requerido é que os domínios locais possam ser organizados dentro de uma árvore de junção [34]. Isto significa que os elementos de S podem ser fixados como "etiquetas" nos vértices de um grafo teórico em árvore T, de tal forma

que para qualquer dois vértices, V_i e V_j, a interseção das etiquetas correspondentes, $S_i \cap S_j$, é um subconjunto da etiqueta posicionada sobre cada um dos vértices existentes num caminho único direcionado de V_i para V_j. Alternativamente, o subgrafo de T consistindo daqueles vértices cujas etiquetas incluem o elemento i junto com os arcos conectando estes vértices, é conectado para $i = 1, ..., n$.

Por exemplo, considere os seguintes domínios locais:

Tabela 3.5: domínios locais organizados numa árvore de junção

	domínio local
1.	$\{x_1\}$
2.	$\{x_1, x_2\}$
3.	$\{x_1, x_3\}$
4.	$\{x_2\}$
5.	$\{x_3, x_4\}$

Os domínios locais mostrados na Tabela 3.5 podem ser organizados numa árvore de junção, como ilustrado na Figura 3.9. Por exemplo, o único caminho ligando o vértice 2 ao vértice 3 é $v_2 \rightarrow v_1 \rightarrow v_3$. e $S_2 \cap S_3 \subseteq S_1$, como requisitado da condição de árvore de junção.

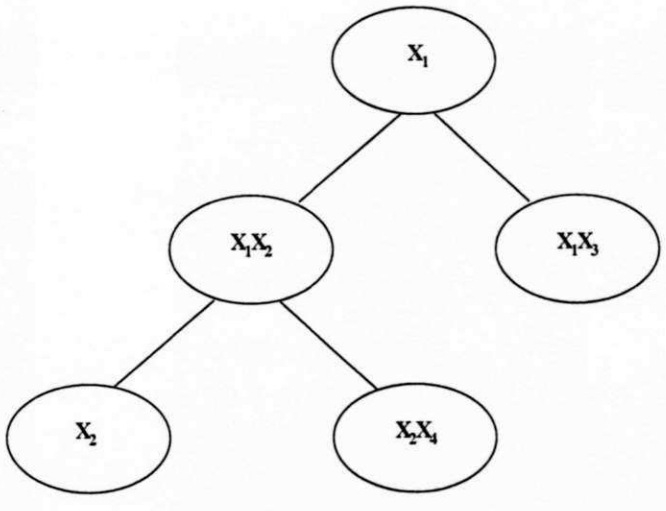


Figura 3.9: Exemplo de árvore de junção

Por outro lado, o conjunto de domínios locais mostrados na Tabela 3.6 não pode ser organizado em uma árvore de junção, como pode ser facilmente verificado.

Tabela 3.6: domínios locais que não podem ser organizados numa árvore de junção

	domínio local
1.	$\{x_1, x_2\}$
2.	$\{x_2, x_3\}$
3.	$\{x_3, x_4\}$
4.	$\{x_4, x_5\}$

Entretanto, adicionando os seguintes domínios redundantes $\{x_1, x_2, x_4\}$ e $\{x_2, x_3, x_4\}$ é possível projetar uma árvore de junção, como mostrado na Figura 3.10.

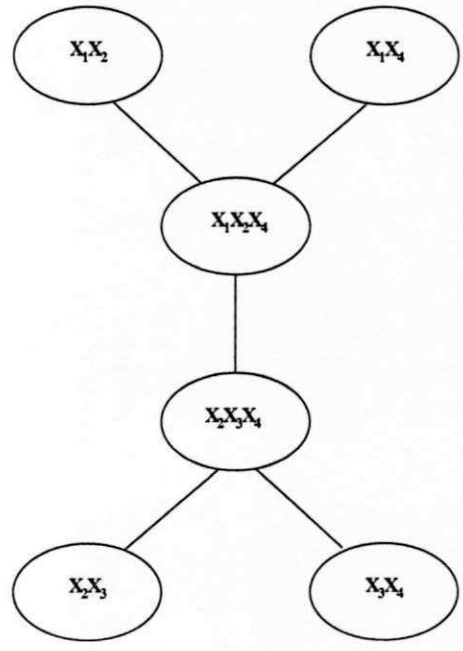


Figura 3.10: Árvore de Junção formada pela adição de domínios redundantes

Num algoritmo "árvore de junção", o qual é denominado de Lei Distributiva Generalizada - LDG, se V_i e V_j estão conectados por um arco (indicado pela notação V_i **adj** V_j), a "mensagem" direcionada de V_i para V_j é uma tabela contendo os valores de uma função $\mu_{i,j}$: $A_{S_i \cap S_j} \rightarrow R$. Inicialmente, todas estas funções são definidas como identicamente iguais a 1 (identidade multiplicativa do semi-anel), e quando uma mensagem particular $\mu_{i,j}$ é atualizada, a seguinte regra é usada [49]:

$$\mu_{i,j}(x_{S_i \cap S_j}) = \sum_{x_{S_i \setminus S_j} \in A_{S_i \setminus S_j}} \alpha_i(x_{S_i}) \prod_{v_k \mathbf{adj} v_i, k \neq j} \mu_{k,i}(x_{S_k \cap S_i}) \tag{3.25}$$

Uma boa forma de lembrar a Equação (3.25) é pensar que a árvore de junção é uma rede de comunicação, na qual o caminho ligando V_i a V_j é uma "linha de transmissão" que "filtra fora" as dependências em todas as variáveis, menos naquelas comuns a V_i e V_j. (A filtragem é realizada através da marginalização). Quando o vértice V_i deseja enviar uma mensagem à V_j ele forma o produto dos seus Kernels locais com todas as mensagens que ele recebeu dos outros vizinhos dele, além de V_j, e transmite o produto para V_j sobre a linha de transmissão (V_i, V_j).

De forma semelhante o "estado" de um vértice V_i é definido como sendo uma tabela contendo os valores de uma função $\sigma_i : A_{S_i} \rightarrow R$. Inicialmente, σ_i é definido para ser o kernel local $\alpha_i(x_{S_i})$, mas quando σ_i é atualizado, a seguinte regra é usada [49]:

$$\sigma_i(x_{S_i}) = \alpha_i(x_{S_i}) \prod_{v_k \mathbf{adj} v_i} \mu_{k,i}(x_{S_k \cap S_i}) \tag{3.26}$$

Em outras palavras, o estado do vértice V_i é o produto dos seus Kernels locais com cada uma das mensagens que ele recebeu de seus vizinhos. A idéia básica é que após uma quantidade suficiente de mensagens terem sido passadas, $\sigma_i(x_{S_i})$ será a função objetiva em S_i, como definido na Equação (3.18).

A questão restante se refere so percurso das mensagens sendo passadas e ao cálculo do estado. Aqui, consideraremos somente dois casos especiais, o problema de um "único vértice", no qual o objetivo é calcular a função objetiva em apenas um vértice v_0, e o problema de "todos os vértices", onde a meta é calcular a função objetiva em todos os vértices. Para o problema do vértice único, o natural roteiro da aplicação da LDG começa pelo direcionamento de cada um dos arcos com respeito ao vértice "sob alvo", v_0, e cada uma das mensagens direcionadas é enviada apenas uma vez. Um vértice envia uma mensagem a seu vizinho quando, pela primeira vez, ele recebeu mensagens dos seus outros vizinhos. O vértice alvo v_0 calcula o estado dele quando ele recebeu mensagens de cada um de seus vizinhos. Com este roteiro, as mensagens começam nas extremidades e procedem na direção de v_0. Quando v_0 receber as mensagens de todos os seus vizinhos, este calcula seu estado e o algoritmo termina.

Por exemplo, se nós desejamos resolver o problema de um "vértice único " para a árvore de junção de Figura 3.9, e o vértice alvo é v_1, então os arcos deveriam ser todos direcionados com respeito a v_1, como mostrado em Figura 3.11. Logo, uma possível seqüência de mensagens e cálculos de estados é mostrada na Tabela 3.7.

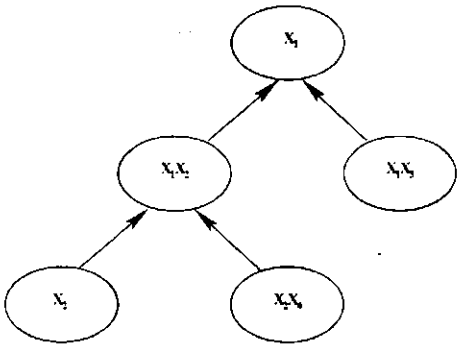


Figura 3.11: Árvore de Junção formada pela adição de domínios redundantes

Tabela 3.7: Roteiro para o cálculo da LDG de um único vértice para a árvore de junção da Figura 3.9 com vértice alvo v_1.

Etapa	Mensagem ou cálculo do estado
1.	$\mu_{3,1}(x_1) = \sum_{x_3} \alpha_3(x_1, x_3)$
2.	$\mu_{4,2}(x_2) = \alpha_4(x_2)$
3.	$\mu_{5,2}(x_2) = \sum_{x_4} \alpha_5(x_2, x_4)$
4.	$\mu_{2,1}(x_1) = \sum_{x_2} \alpha_2(x_1, x_2) \cdot \mu_{4,2}(x_2) \cdot \mu_{5,2}(x_2)$
5.	$\sigma_1(x_1) = \alpha_1(x_1) \cdot \mu_{2,1}(x_1) \cdot \mu_{3,1}(x_1)$

Para problema de todos os vértices, a LDG pode ser conduzida de diversas formas. Estas são apresentadas em [49]. Para finalizar esta Seção, considere a árvore de junção mostrada na Figura 3.12. Em [49] é mostrado que quando o algoritmo da LDG é aplicado a árvore de junção da Figura 3.12, o resultado obtido é equivalente ao algoritmo de propagação de crenças de Pearl [39] em redes bayesianas.

3.6 Inferência em Redes Bayesianas Aproximadas

Os critérios de remoção de arcos apresentados na Seção 2.4.1 tiveram sua importância ressaltada na redução da complexidade computacional de redes aprendidas automaticamente de bases de dados de aplicações práticas. Contudo, sua maior relevância

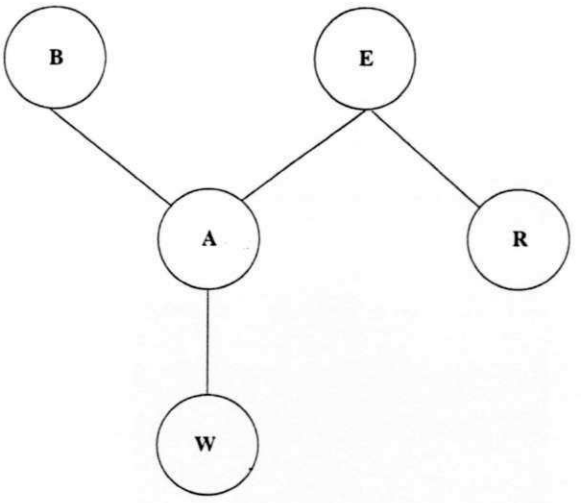


Figura 3.12: Árvore de Junção formada por cinco domínios locais

vai muito além disto. Conforme discutido nas Seções 3.4.1, 3.4.2 e 3.5, o algoritmo de complexidade de tempo polinomial de Pearl [39] é específico para redes que apresentam topologia em árvore múltipla. Evidentemente, na maior parte dos casos, isto constitui uma forte restrição de uso. Para avaliar os efeitos da presença de laços na propagação de mensagens, considere e rede esboçada na Figura 3.13 abaixo.

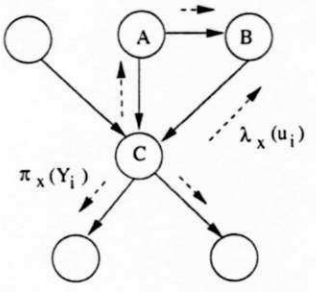


Figura 3.13: Laço com três nós numa rede bayesiana.

O algoritmo de fusão de influências proposto por Pearl trabalha com a suposição que a rede não possui laços, pois devido ao mecanismo de propagação de mensagens (ao receber uma mensagem um nó envia outras para os seus pais e para seus filhos), é possível que uma mensagem enviada por um nó que faz parte de um laço ao seu pai que está dentro do laço e assim sucessivamente para os demais nós do laço circule e retorne ao nó inicial, fazendo com que o processo de envio de mensagens reinicie, e se repita indefinidamente tornando a rede instável. O mesmo vale para a mensagem

enviada para os filhos.

Outros algoritmos têm sido propostos para avaliar a fusão de influências em redes com laços [41]. Em geral, esses algoritmos se caracterizam essencialmente por fazer alterações na estrutura da rede, seja através da fusão de nós ou do desdobramento de um nó em dois ou mais, objetivando a propagação de crenças mesmo com ciclos.

Contudo, a abordagem adotada nesta Seção é analisar a inferência probabilística na rede Angina, a qual teve seus arcos removidos de acordo com os critérios anteriormente apresentados com o objetivo de quebrar os ciclos. Para tanto, a base de casos com 10.000 casos foi dividida em duas bases, sendo uma de treinamento e outra de testes. A base de treinamento é constituída por 9.000 casos e a base de testes por 1.000 casos. A rede aprendida a partir da base de dados formada por 9.000 casos é mostrada na Figura 3.14. A medida do comprimento de descrição da rede original é de 11.345, 13 bits. A Tabela 3.8 relaciona os nós às variáveis aleatórias e suas instâncias. Na Tabela 3.9 são mostrados os padrões de entradas e suas freqüências de ocorrência na base de testes.

Tabela 3.8: Rede Angina: Variáveis aleatórias e suas Instâncias

Nó	**Variável Aleatória**	**Instâncias**		
	(sintoma)	**0**	**1**	**2**
N0	Febre	Ausente	Baixa	Alta
N1	Manchas	Ausente	Presente	
N2	**Dor**	Ausente	Presente	
N3	Angina	Ausente	Moderada	Severa
N4	**Frio**	Ausente	Presente	

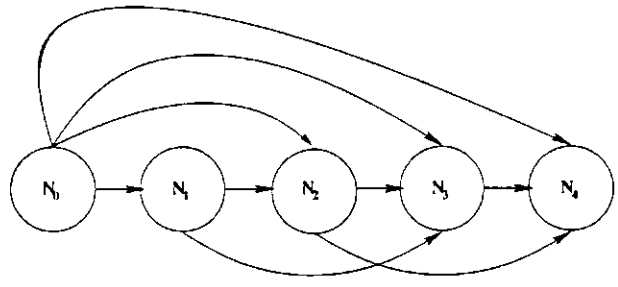


Figura 3.14: Rede Angina aprendida da base de dados formada por 9.000 casos.

Após a aplicação de cada um dos três critérios de remoção de arcos apresentados na Seção 2.4.1, foram obtidas as seguintes redes apresentadas nas Figuras 3.15, 3.16 e

Tabela 3.9: Base de Testes: Padrões Observados e suas Quantidades

Padrão					Freqüência de Ocorrência
Febre	**Manchas**	**Dor**	**Angina**	**Frio**	
0	0	0	0	0	750
1	0	0	0	0	170
0	0	0	0	1	12
0	0	1	0	0	36
1	0	0	0	1	9
0	0	1	0	1	6
1	0	1	1	0	2
1	0	1	0	0	6
1	0	1	0	1	3
1	1	1	2	0	1
2	0	1	1	0	1
2	0	0	0	0	1
2	0	1	0	1	2
2	0	0	0	1	1

3.17, as quais apresentaram, respectivamente, as medidas do comprimento de descrição de 11.747, 96 ,11.720, 16 e 11.726, 14 bits.

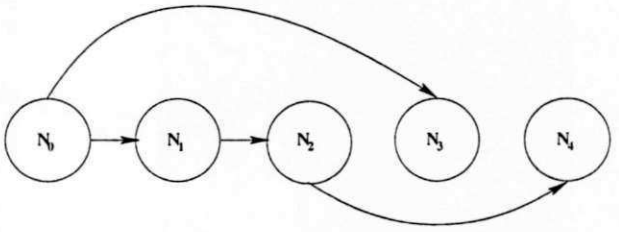


Figura 3.15: Rede Angina obtida após a aplicação do critério 1: Distribuições Conjuntas.

A eficácia da inferência probabilística nas redes bayesianas que tiveram seus arcos removidos através dos critérios propostos na Seção 2.4.1 pode ser analisada através da Tabela 3.10. Apesar do alto desempenho apresentado, o diagnóstico da ocorrência mais rara (presente em 0,17% dos casos), isto é angina severa, somente foi diagnosticado corretamente pela rede que teve seus arcos removidos através do critério baseado no comprimento de descrição.

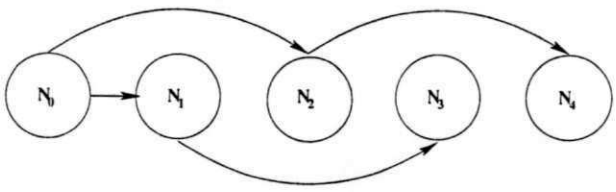


Figura 3.16: Rede Angina obtida após a aplicação do critério 2: Descrições de Comprimento.

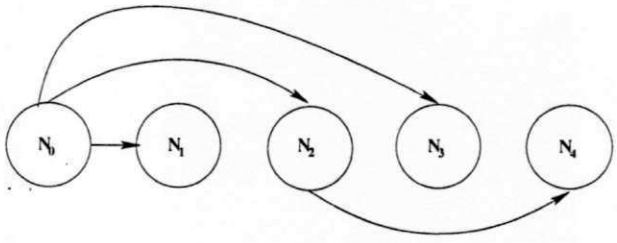


Figura 3.17: Rede Angina obtida após a aplicação do critério 3: Entropia Condicionada.

3.6.1 Efeito da Ordenação das Variáveis

As redes bayesianas constituem um modelo do domínio do problema e expressam relações causais entre as variáveis aleatórias presentes na base de dados. Conforme apresentado na Seção 2.3.4, o algoritmo K3 necessita de uma prévia ordenação entre as variáveis envolvidas no domínio do problema. Geralmente esta ordenação traz dois grandes benefícios:

1. Permite incorporar o conhecimento do especialista, traduzindo-se num modelo mais próximo do real, o que certamente facilita as explicações a cerca dos resultados obtidos e,

2. Elimina do espaço de busca as redes que não atendem a esta prévia ordenação, o que permite um aprendizado mais rápido.

Tabela 3.10: A eficácia da Inferência Probabilística em Redes Aproximadas

Critério	Taxa de Acerto	Observações
Distribuições Conjuntas	99,60%	
Descrição do Comprimento	99,70%	Diagnóstico da ocorrência mais rara
Entropia Condicionada	99,60%	

isto, a termometria foi incorporada na atividade médica, embora o termômetro fosse conhecido desde o século XVII, seu emprego como instrumento para medir a temperatura corporal teve inínio em 1852, quando Traube e, a seguir, Wunderlich, na Alemanha, introduziram o gráfico de temperatura ou curva térmica, que permitiu a caracterização dos vários tipos de febre. A medida indireta da pressão arterial só se tornou possível a partir de 1880, quando von Basch, na Alemanha, idealizou o primeiro aparelho, que nada mais era que uma bolsa de borracha cheia de água e ligada a uma coluna de mercúrio ou a um manômetro. Comprimindo a bolsa de borracha sobre a artéria até o desaparecimento do pulso obtinha-se a pressão sistólica. Em 1896, um médico italiano, Riva-Rocci, substituiu a bolsa por um mangueira de borracha e a água pelo ar. A medida da pressão diastólica, contudo só foi possível após 9 anos, quando o médico russo, Nikolai Korotkov descobriu os sons produzidos durante a descompressão da artéria. No final do século XIX, o médico já utilizava três instrumentos básicos no exame do paciente. Além desses três instrumentos, outras ferramentas foram adicionados na prática médica, como o oftalmoscópio, retrator de língua, martelo de reflexo, etc. O aperfeiçoamento do microscópio, por sua vez, proporcionou o surgimento da microbiologia, permitindo identificar os agentes causadores de muitas doenças [27].

A tecnologia médica, propriamente dita, só se desenvolveu no decorrer do século XX, com o diagnóstico por imagens, endoscopia, métodos gráficos, exames de laboratório e provas funcionais. Contudo, como marco inicial da era tecnológica podemos considerar a descoberta por Roentgen, dos raios-X, em 1895. No seu Laboratório de Física, Roentgen obteve a primeira radiografia dos ossos da mão de sua esposa em dezembro de 1895 e em janeiro de 1896 repetia a experiência perante a Sociedade de Física de Würzburg, radiografando a mão do Prof. de Anatomia Albert von Kolliker, que se achava presente. Kolliker propôs que os raios-X fossem chamados de raios Roentgen, denominação ainda usada em países europeus. A descoberta dos raios-X proporcionou a aquisição de novas evidências, que anteriormente eram coletadas através de técnicas invasivas. A emissão de diagnóstico baseada em imagens teve início com os raios-X. Contudo, outros métodos de obtenção de imagens foram desenvolvidos como, por exemplo, cintilografia, ultra-sonografia, tomografia computadorizada, ressonância magnética [27].

Como é possível observar, todas as abordagens anteriores proporcionaram uma redução da complexidade do cenário de tomada de decisão através da aquisição de novas evidências. Contudo, a abordagem adotada no desenvolvimento do sistema de auxílio à

emissão do diagnóstico de patologias neuromusculares raras é bem diferente. A preocupação aqui é desenvolver uma ferramenta de análise de evidências, a qual proporcione ao usuário uma resposta que permita a este emitir um diagnóstico diferencial acerca de patologias que apresentam sintomas em comum.

4.2 Neurologia

A Neurologia é uma das especialidades médicas mais antigas e chegou a apresentar uma evolução muito lenta no decorrer dos anos. A Neurologia é o ramo da medicina que se ocupa das doenças do sistema nervoso em todos os seus aspectos. Contudo, neste trabalho, somente serão consideradas as patologias neurológicas pertencentes ao grupo muscular. Estas serão discutidas na próxima Seção.

4.2.1 Patologias Musculares

As afecções musculares primitivas destroem a estrutura ou o funcionamento das fibras musculares independentemente de sua inervação. As lesões das fibras são disseminadas de maneira anárquica e escapam a uma sistematização em unidades motoras. As alterações do tecido intersticial do músculo são geralmente associadas às lesões dos elementos contráteis. As distrofias musculares progressivas (miopatias) decorrem de um processo degenerativo, geneticamente determinado; as miosites estão sob a dependência de um processo inflamatório adquirido; algumas alterações metabólicas caracterizam-se por suas manifestações musculares predominantes e finalmente, a miastenia resulta de um funcionamento anormal da junção neuromuscular. A Figura 4.1 mostra a região de ocorrência de doenças neuromusculares [4].

Classificação

As patologias neuromusculares são classificadas de acordo com o local de ocorrência. Assim, tem-se que [46]:

- No Neurônio Motor Primário;
- Na Raiz e nervos periféricos;
- Na Junção mioneural: Botulismo, Síndrome Miastênica Congênita, Miastenia Gravis;

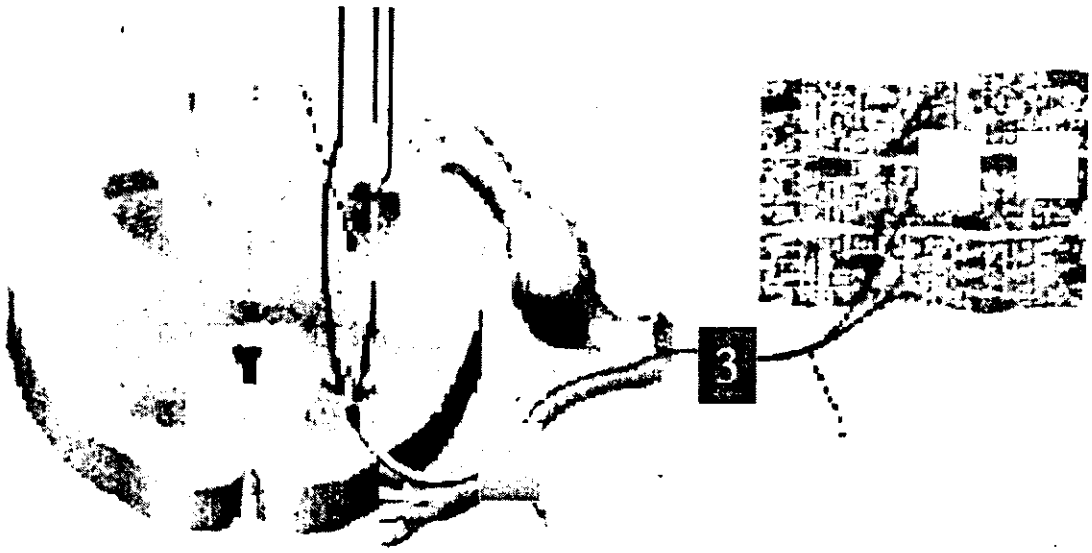


Figura 4.1: Doenças Neuromusculares decorrentes do acometimento primário da Unidade Motora 1-motoneurônio 2-raiz motora 3-nervo periférico 4-junção mioneural 5-músculo

- Na Fibra muscular: Miopatias.

Características das Patologias Neuromusculares

As patologias pertencentes ao grupo neuromuscular apresentam características em comum, por exemplo, é comum aos portadores dessas patologias apresentarem [46]:

- Fraqueza muscular proximal e às vezes atrofia;
- Sensibilidade conservada;
- Reflexos profundos conservados na fase inicial.

Na próxima Seção é apresentado a patologia neuromuscular escolhida como foco do sistema de auxílio à emissão de diagnóstico médico.

4.3 Foco da Aplicação: Miastenia Gravis

A Miastenia Gravis é uma doença caracterizada por fraqueza e fadiga anormal dos músculos estriados, presumivelmente devido a um defeito da transmissão do impulso

nervoso ao nível da junção mioneural, de origem auto-imune. Ela evolui por surtos e sua gravidade é devida ao risco de acidentes respiratórios [4].

Como toda doença auto-imune, a Miastenia acontece em virtude da produção de anticorpos contra elementos do próprio organismo, mais precisamente, contra uma estrutura do músculo chamada de receptor de acetilcolina, que é a região onde o nervo eferente se liga no músculo. As Figuras 4.2 - 4.3 ilustram a redução do número de receptores da acetilcolina em pacientes miastênicos quando comparado a pacientes normais [4].

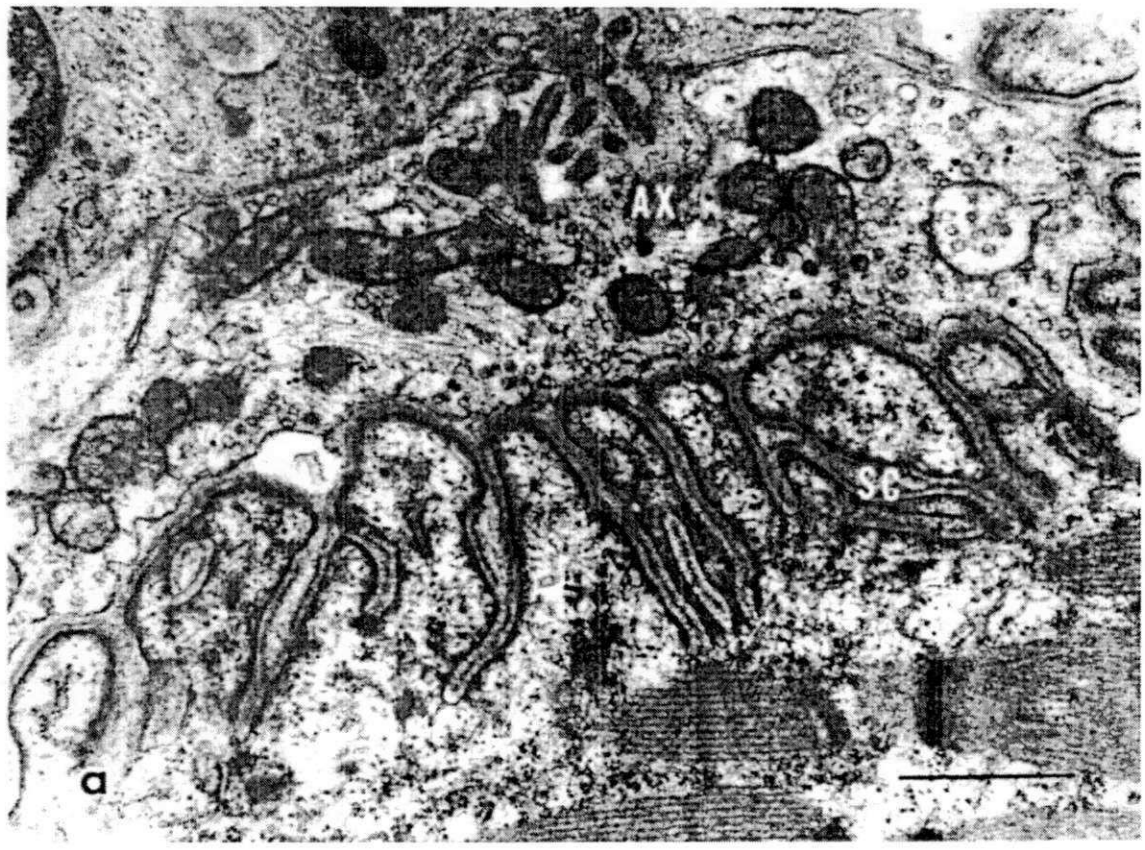


Figura 4.2: Corte da junção neuromuscular mostrando uma quantidade normal de receptores da acetilcolina

O pico de incidência ocorre em adultos jovens (entre a segunda e quarta décadas), afetando mais o sexo feminino antes dos 40 anos. Porém, em idades mais avançadas, ambos os sexos são igualmente afetados. Como já discutido, o principal sintoma é a fadiga, a qual pode aparecer em qualquer músculo do corpo. Logo, a Miastenia pode: produzir sintomas tais como fraqueza nos braços e pernas; e ocasionar dificuldades para

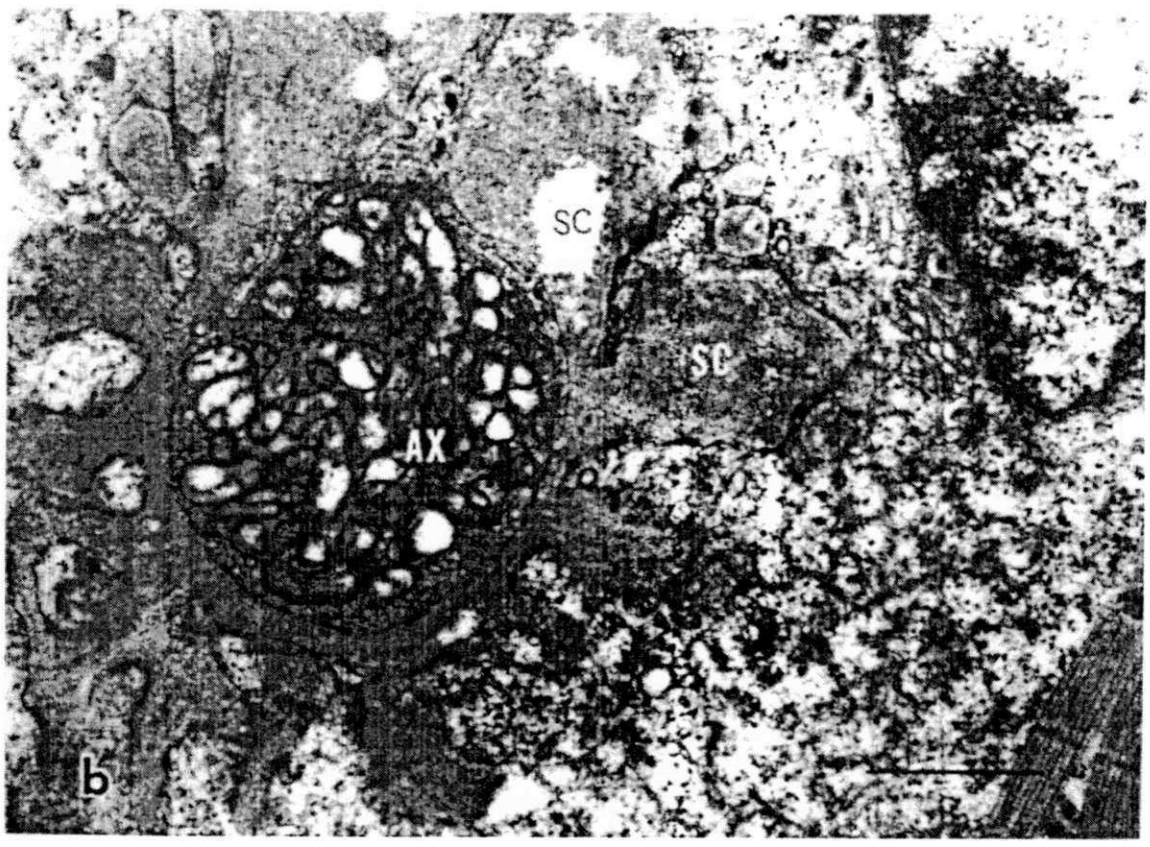


Figura 4.3: Corte da junção neuromuscular mostrando uma quantidade reduzida de receptores da acetilcolina

mastigar e para engolir (disfagia). Quando acomete os músculos do tórax, a Miastenia provoca falta de ar (dispnéia) e voz anasalada (disfonia). Na face, a Miastenia causa queda das pálpebras (ptose palpebral) e visão dupla (diplopia). Em cerca de 50% dos pacientes, os olhos estão inicialmente envolvidos, seja pela ptose palpebral, geralmente assimétrica ou por diplopia, estrabismo decorrente da paralisia da musculatura ocular extrínseca. Caracteristicamente, o comprometimento desta musculatura pode alternar de um lado para outro em exames sucessivos. Na Figura 4.4 é mostrada a presença da Ptose palpebral em um paciente miastênico [4].

Outros sintomas comuns afetam os músculos faciais ou orofaríngeos, ocasionando disartria (dificuldade para articular as palavras), disfagia (dificuldade para engolir) e limitação dos movimentos faciais que determina o aspecto inexpressivo da mímica. Juntas, as fraquezas orofaríngea e ocular causam sintomas em praticamente todos os pacientes portadores de miastenia. Também é comum a fraqueza dos membros e do

Figura 4.4: Presença de ptose unilateral flutuante em paciente miastênico

pescoço. Nos membros, a alteração miastênica predomina sobre os músculos proximais, aparecendo somente durante um esforço mantido ou pertubando permanentemente a atividade do doente [4].

A incidência de Miastenia Gravis na população brasileira é estimada em 4 casos para cada grupo de 100.000 habitantes. A nível mundial, contudo, a baixa freqüência de ocorrência de Miastenia pode variar de 1 a 11 casos para cada grupo de 100.000 habitantes. Quanto as formas clínicas, tem-se [4]:

- Miastenia ocular: só acomete os músculos oculares, podendo haver ptose e paralisia com diplopia. A debilidade dos músculos oculares pode acarretar paralisia ou fraqueza de músculos isolados, paralisia do olhar conjugado ou oftalmoplegia completa em um ou em ambos os olhos. Sua incidência é de aproximadamente 15%;

- Miastenia generalizada - leve ou moderada e que corresponde a 85% dos casos;

- Miastenia fulminante aguda com crise respiratória.

O diagnóstico no início da doença é por vezes difícil, sendo os pacientes muitas vezes considerados neuróticos. O início do quadro é geralmente insidioso e a fraqueza muscular manifesta-se especialmente no fim do dia, quando o paciente está mais

cansado, ou então após exercício físico não comum ou mesmo após quadro infeccioso, denotando assim sua natureza flutuante. A história clínica do paciente é o principal dado para o diagnóstico, juntamente com o exame físico. Como exames complementares temos a eletroneuromiografia que revela a existência de bloqueio miastênico, ou seja, uma diminuição progressiva da amplitude dos potenciais demonstrando o número crescente de junções neuromusculares ineficazes no interior de cada unidade motora; a dosagem de anticorpos anti-receptor de acetilcolina, que está presente na região do receptor neuromuscular; e um teste com injeção de neostigmina ou edrofônio, após o qual há melhora imediata da força [46].

Aproximadamente 65 a 70% dos pacientes têm hiperplasia tímica e 15% têm Timomas (idosos). Ou seja, cerca de 70% das glândulas timo (localizada na parte superior do tórax e que controla a produção dos anticorpos), de pacientes adultos portadores de Miastenia Gravis não involuíram e pesam mais que o normal. Sendo assim, na maioria das vezes se opta por realizar uma cirurgia para retirada do timo [46].

4.4 Exemplo de Caso Clínico e Base de Dados para a Miastenia Gravis

Nesta Seção, tem-se um exemplo de um caso clínico, como ele é fornecido pelo médico, e como os valores das variáveis de entrada para a base de dados são retirados. Este procedimento foi adotado na formação da base de dados formada por 74 casos clínicos apresentada no Apêndice B. Assim, mantida a incidência de 4 casos para cada 100.000 habitantes, o sistema a ser desenvolvido para o auxílio ao diagnóstico das variações de Miastenia, utiliza um aprendizado equivalente a apresentação de 1.850.000 casos ao cérebro do sistema que é a rede bayesiana apresentada no próximo Capítulo.

4.4.1 Exemplo de Caso Clínico

Normalmente, o médico descreve suas observações do paciente como apresentado no exemplo de caso abaixo.

Caso 13 *Paciente do sexo masculino, 38 anos, queixa-se de* ptose palpebral e visão dupla, que se expressa no fim do dia, *há cerca de três meses. Refere-se que a partir desta época tem sensação permanente de cansaço, que piora com os esforços,*

N1	N2	N3	N4	N5	N6	N7	N8
2	2	1	2	2	2	1	3

Tabela 4.1: Exemplo de um caso que compõe a base de dados.

mais pronunciado no fim do dia, ou após atividade física. Queixa-se de dificuldade para falar. Relata que possui uma motocicleta de 750cc e que agora, só com muita dificuldade, consegue manuseá-la.

Exame Neurológico: *observa-se ptose palpebral à esquerda, estrabismo convergente no olho esquerdo. A força muscular está diminuída nos músculos proximais dos quatro membros (grau 4-) e normal nos músculos distais. Sensibilidade e reflexos normais. Coordenação normal. Em 29.09.98, submeteu-se a estimulação repetitiva do nervo auxiliar direito, a 5Hz, com registro no músculo deltóide. Observa-se diminuição da amplitude, aproximadamente 20% entre o 1º e o 5º potencial de ação muscular composto, o que sugere comportamento miastênico da junção mioneural. Iniciou tratamento com piridostigimina 60mg., quatro vezes ao dia, com melhora importante da força muscular, do estrabismo e da ptose palpebral. Em outubro de 1998, o paciente realizou tomografia computadorizada no mediastino que evidenciou a presença de timona. Fez cirurgia para retirada do tumor. Atualmente, está sob uso de piridostigmina assintomático.* **Diagnóstico: Miastenia Gravis- generalizada moderada**

De posse desse caso deve-se extrair as variáveis de interesse e quais são os seus respectivos valores associados. No caso apresentado acima, algumas variáveis estão em negrito no texto, isso foi feito aqui para exemplificar a montagem de um caso e não é comum o médico fornecer o caso dessa forma. Os sintomas observados são associados com os números da Tabela B.1 apresentada no apêndice B e, desta forma, monta-se o *caso* como está mostrado na Tabela 4.1.

4.5 Conclusões

Foi apresentado neste Capítulo uma breve descrição da incorporação de novas ferramentas na Medicina, as quais proporcionaram uma melhoria efetiva da atividade médica. Também foram apresentadas algumas características do grupo neuromuscular, ressaltando particularidades da Miastenia Gravis, que é uma das patologias pertencentes a esse grupo.

Também foi apresentado um caso clínico exemplificando o procedimento adotado na formação da base utilizada no aprendizado da rede bayesiana.

É importante perceber, que a abordagem bayesiana permite incorporar outras patologias do mesmo grupo neuromuscular com a vantagem de apenas ser exigido uma simples extensão na base de casos utilizada. Entretanto, isto não foi possível devido ao fato das outras patologias pertencentes ao grupo neuromuscular serem ainda mais raras, sendo difícil conseguir um número razoável de casos que permitisse o aprendizado da rede.

No Próximo Capítulo é apresentada a metodologia para o desenvolvimento do Sistema de Auxílio à Emissão do Diagnóstico das Variações de Miastenia Gravis, além do procedimento de validação desse sistema.

Capítulo 5

Sistema Desenvolvido

5.1 Introdução

Ao longo das últimas décadas, a medicina vem recebendo notáveis contribuições tecnológicas, as quais permitiram um grande avanço na atividade médica. Em geral, tais contribuições podem ser divididas em sistemas de aquisição de dados (exames clínicos e laboratoriais) e sistemas de avaliação dos dados observados, nos quais destacam-se os Sistemas Especialistas.

Os sistemas especialistas constituem uma abordagem de apoio à tomada de decisão baseada em predicados lógicos. Em geral, a obtenção das regras empregadas no tratamento dos dados observados é feita através de consultas a especialistas do domínio do problema. A característica principal destas regras é a classificação dos dados de entrada no sistema em eventos mutualmente exclusivos, ou seja, a quantização dos dados em níveis previstos pelo especialista consultado. O problema é que em certas áreas do conhecimento, como por exemplo na medicina, a avaliação do especialista é feita de forma cognitiva, sendo difícil dizer que dois especialistas do mesmo domínio guardam regras idênticas para a quantização necessária. Por outro lado, a classificação em eventos mutualmente exclusivos é difícil, já que é comum na medicina a ocorrência de patologias que afetam as observações de maneiras opostas, ou mesmo o fato de diferentes patologias apresentarem sintomas em comum. Todas estas particularidades evidenciam a dificuldade do desenvolvimento de sistemas especialistas na área médica e a necessidade de se dispor de técnicas capazes de permitir o desenvolvimento de sistemas mais robustos ao cenário apresentado.

Neste sentido, as redes bayesianas caracterizam-se pela realização de inferências de

Nós	Pais
N_0	Não tem pais
N_1	N_0
N_2	N_0 e N_1
N_3	N_0, N_1 e N_2
N_4	N_0, N_2 e N_3
N_5	N_0, N_1 e N_3
N_6	N_0, N_1, N_2, N_3, N_4 e N_5
N_7	N_0, N_1, N_2, N_3, N_4, N_5 e N_6

Tabela 5.1: Nós e conjunto da pais da rede bayesiana para a Miastenia Gravis

base probabilística, que é útil quando existe a possibilidades das doenças sob foco apresentarem sintomas em comum e onde é necessário realizar um diagnóstico diferencial. Assim como os sistemas especialistas, as redes bayesianas permitem a incorporação do conhecimento especialista, mas sem que esta exclua a extração do conhecimento humano embutido na base de dados utilizada no aprendizado da rede. Logo, em situações onde é difícil contar com a presença de especialistas, ou mesmo quando existem divergências de opiniões, isto constitui uma vantagem. É importante perceber que a abordagem probabilística evita o uso de regras lógicas, desprovidas de rigor matemático, o que pode levar a aplicações restritas a um domínio específico e cuja aplicação a outro problema implica necessariamente na busca por um conjunto novo e adequado de regras, o que não ocorre na abordagem bayesiana.

5.2 Rede Miastenia Gravis

A base de casos clínicos apresentada no Apêndice B foi utilizada no aprendizado da rede, utilizando para isto o algoritmo K3 e adotando inicialmente a ordem sugerida pelo Prof. Jovany Medeiros. A Tabela 5.1 ilustra os nós e conjuntos de pais da rede bayesiana obtida da base de casos clínicos de Miastenia Gravis. Na Figura 5.1, tem-se uma representação parcial dessa rede (não são mostradas todos os ramos), onde pode ser observada a ocorrência de laços. Nesta figura, as linhas tracejadas foram utilizadas de forma a não sobrecarregar o esboço parcial da rede.

A rede obtida apresenta um alto grau de conectividade, inviabilizando a realização

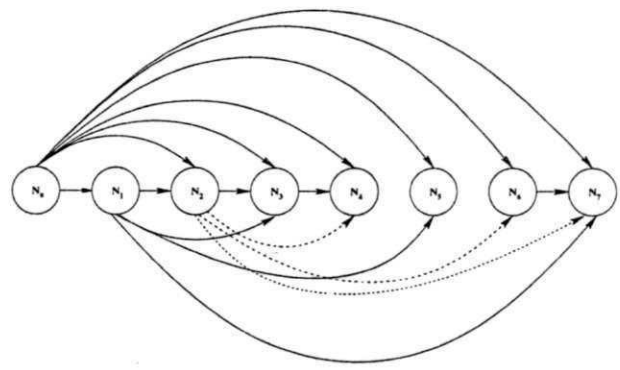


Figura 5.1: Representação parcial da Tabela 5.1

de inferências através do método de passagem de mensagens de Pearl. Como forma de reduzir a complexidade da rede obtida e possibilitar a realização de inferências em tempo polinomial, serão os aplicados os critérios de remoção de arcos propostos, sendo as redes obtidas mostradas na próxima Seção. As redes obtidas serão analisadas de acordo com a medida do comprimento de descrição e quanto à eficácia no auxílio à emissão de diagnóstico médico.

5.2.1 Aplicação da Remoção de Laços à Rede Miastenia Gravis

Aplicando os algoritmos propostos para a remoção de arcos à rede Miastenia Gravis, obtém-se as redes derivadas e apresentadas nas Figuras 5.2, 5.3 e 5.4 para o primeiro, segundo e terceiro critérios propostos.

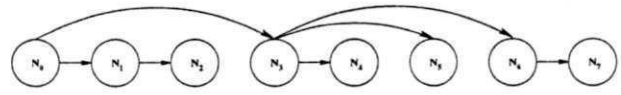


Figura 5.2: Rede obtida após a quebra dos laços da rede apresentada na Figura 5.1 usando o primeiro critério: Divergência de Distribuições.

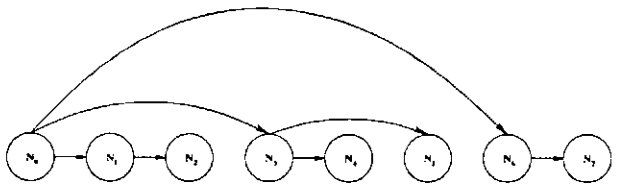


Figura 5.3: Rede obtida após a quebra dos laços da rede apresentada na Figura 5.1 usando o segundo critério: Descrições de Comprimento

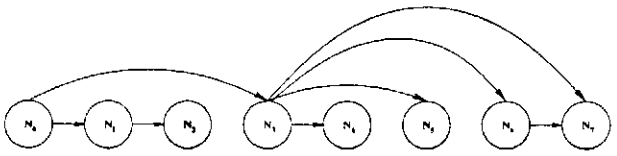


Figura 5.4: Rede obtida após a quebra dos laços da rede apresentada na Figura 5.1 usando o terceiro critério: Entropia Condicionada

A medidas do comprimento de descrição de comprimento para a rede original foi de 306,46 bits; já as redes obtidas do primeiro, do segundo e do terceiro critério apresentaram a mesma medida do comprimento de descrição: 437,19 bits. Esta coincidência pode ser percebida na semelhança entre as topologias apresentadas pelas redes.

As redes foram testadas quanto a eficácia no apoio à tomada decisão médica. A primeira base de testes utilizada foi a própria base de treinamento constituída pelos 74 casos de Miastenia, únicos existentes. Embora esta atitude não seja correta em aplicações correntes na área de reconhecimento de padrões, como por exemplo em sistemas de reconhecimento de manuscritos, ela pode ser justificada pelo fato que observando atentamente a base de 74 casos, percebe-se que esta é formada por alguns padrões de valores de variáveis bem definidos, cujas freqüências de ocorrência somam-se para compor o número de casos existentes. Na verdade, isto já era esperado, visto que como se trata de uma classe de patologias com quatro variações, esta deve apresentar alguns padrões que corretamente a identifique e não outra existente. Já em sistemas de reconhecimento de manuscritos, a variabilidade das formas de letras corridas existentes leva a bases de aprendizado e de testes realmente diferentes, as quais são constituídas por elementos separados por classes, mas estes não padronizados e onde suas freqüências de ocorrência não somam-se para formar a base de treinamento.

Assim, testes iniciais de eficácia mostraram resultados satisfatórios, onde as redes obtidas dos critérios de remoção de arcos apresentaram taxas idênticas de acerto,

Rede obtida	Taxa de Acerto		
	"1ª"	"2ª"	"1ª" e "2ª"
Primeiro Critério	60,81%	32,43%	93,24%
Segundo Critério	60,81%	32,43%	93,24%
Terceiro Critério	60,81%	32,43%	93,24%

Tabela 5.2: Teste de eficácia nas redes bayesianas para a Miastenia Gravis obtidas dos critérios de remoção de arcos utilizando como base de testes a base de aprendizagem

conforme mostrada na Tabela 5.2. Nesta Tabela, "1ª" significa que o sistema diagnosticou corretamente a variação da patologia, retornando a resposta com a mais alta probabilidade, enquanto que "2ª" a resposta retornada apresentava a segunda mais alta probabilidade, desde que a diferença entre a primeira e a segunda não fosse superior aos 10%. Esta possibilidade de acerto através da observação da segunda maior probabilidade é razoável quando consideramos que o sistema final desenvolvido não será responsável por emitir um diagnóstico, mas ao invés disto, auxiliará clínicos a fazê-lo. Por outro lado, esta diferença de 10% adotada, pode levar a uma dúvida na emissão do diagnóstico, já que do senso comum valores de resposta são suficientemente próximos, desde que a diferença entre eles seja inferior aos 10%. Logo, como existe a possibilidade durante a entrada de valores correspondendo a "não observado" no sistema, esta consideração pode motivar clínicos a realizarem exames complementares, que proporcionarão um resposta mais precisa. Por outro lado, na ausência de exames complementares, a opção pela segunda mais alta probabilidade pode levar a corretude no diagnóstico, se o médico julgar a resposta do sistema razoável. Somente em 6,76% dos casos o sistema erra no diagnóstico, o que é razoável quando é lembrado que aproximadamente 20% dos casos clínicos obtidos durante a etapa de formação da base de casos apresentavam o diagnóstico médico errado.

5.3 Ordenação de Variáveis na Rede Miastenia Gravis

As medidas de informação apresentadas no Capítulo 3 foram utilizadas para avaliar a ordenação na rede Miastenia Gravis. A ordenação predominante, na ordem decrescen-

Nós	Pais
$N_0' = N_3$	Não tem pais
$N_1' = N_6$	N_0'
$N_2' = N_4$	N_0' e N_1'
$N_3' = N_1$	N_0', N_1' e N_2'
$N_4' = N_0$	N_0', N_1', N_2' e N_3'
$N_5' = N_5$	N_0', N_3' e N_4'
$N_6' = N_2$	N_0', N_3' e N_4'
$N_7' = N_7$	N_0', N_1', N_2', N_3', N_4', N_5' e N_6'

Tabela 5.3: Nós ordenados e conjunto da pais da rede bayesiana para a Miastenia Gravis

tes das medidas, é $N_3, N_6, N_4, N_1, N_0, N_5, N_2, N_7$. Sendo assim, de acordo com estas medidas, a variável aleatória "Fraqueza Muscular" na base de dados contribui com uma quantidade maior de informação na composição do estado de saída, do que a variável "Ptose Palpebral".

A Tabela 5.3 ilustra os nós e o conjunto de pais da rede bayesiana obtida utilizando o algoritmo K3 e esta nova ordenação. Um esboço parcial da rede é mostrada na Figura 5.5. Novamente a rede obtida apresenta um alto grau de conectividade. Assim, aplicando os critérios de remoção de arcos propostos, obtemos as redes apresentadas nas Figuras 5.6, 5.7 e 5.8. A medida do comprimento de descrição (CD) da rede original é de 333,79 bits. Por sua vez, as medidas do CD para as redes obtidas do primeiro e segundo critérios são iguais a 488,38 bits, enquanto a obtida do terceiro critério tem um CD de 535.47 bits. A Tabela 5.4 apresenta os resultados do teste de eficácia, considerando esta nova ordenação.

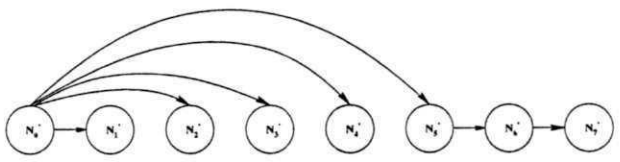


Figura 5.6: Rede obtida após a quebra dos laços da rede apresentada na Figura 5.5 usando o primeiro critério: Divergência de Distribuições.

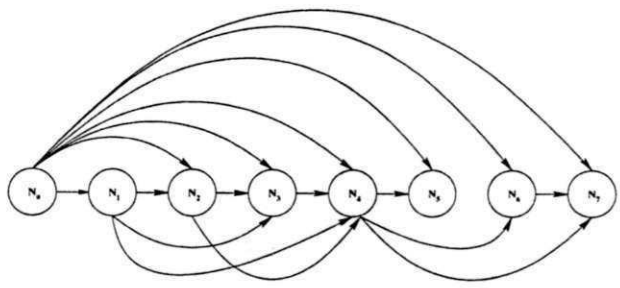


Figura 5.5: Representação parcial da Tabela 5.3

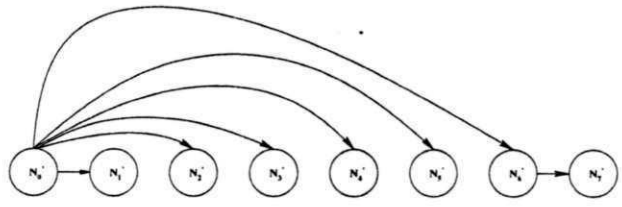


Figura 5.7: Rede obtida após a quebra dos laços da rede apresentada na Figura 5.5 usando o segundo critério: Descrições de Comprimento

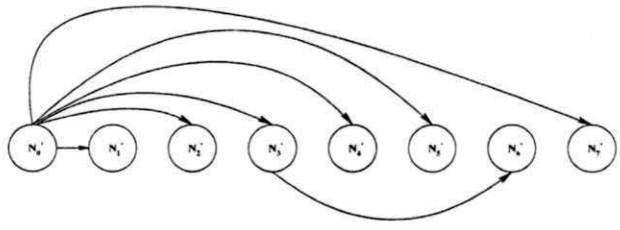


Figura 5.8: Rede obtida após a quebra dos laços da rede apresentada na Figura 5.5 usando o terceiro critério: Entropia Condicionada

Assim, de acordo com a Tabela 5.4 a ordenação de variáveis continua afetando diretamente na eficácia do sistema. O grau de influência observado neste caso é de $\pm 20\%$. A rede obtida pelo critério da Entropia Condicionada apresentou um aumento significativo, chegando próximo aos $80,0\%$, que é razoável para um sistema de auxílio à emissão de diagnóstico médico.

Rede obtida	Taxa de Acerto		
	"1ª"	"2ª"	"1ª" e "2ª"
Primeiro Critério	37,84%	27,02%	64,86%
Segundo Critério	37,84%	27,02%	64,86%
Terceiro Critério	79,73%	13,51%	93,24%

Tabela 5.4: Teste de eficácia nas redes bayesianas para a Miastenia Gravis obtidas dos critérios de remoção de arcos e que tiveram suas variáveis ordenadas de acordo com as medidas apresentadas no Capítulo 3

5.4 Validação do Sistema de Auxílio à Emissão do Diagnóstico Médico

Os testes de eficácia apresentados na Seção anterior serviram apenas para escolher a rede bayesiana que seria utilizada no desenvolvimento do sistema de auxílio à emissão do diagnóstico médico. Como pode ser visto através das Tabelas 5.2 e 5.4, a rede bayesiana cujos a ordenação dos nós foi feita de acordo com as medidas de informação apresentadas no Capítulo 3 e onde os arcos foram removidos de acordo com o critério baseado na Entropia Condicionada de Shannon, apresentou eficácia superior, o que nos permite selecioná-la para a próxima etapa de testes, cujo o objetivo agora é validar o sistema de auxílio proposto à tomada de decisão.

A base de casos de miastenia adotada na composição da validação do sistema é formada por 30 casos clínicos pertencentes ao cenário considerado pelo especialista no diagnóstico de Miastenia e Doutor em Neurologia, o Prof. Jovany Medeiros. Este cenário de testes avalia o diagnóstico em cerca de 750.000 ocorrências. Desta forma, temos uma base de testes que corresponde a 40% da base utilizada no aprendizado da rede. Nesta base, encontram-se as quatro variações da Miastenia Gravis: Ocular, Generalizada Leve, Generalizada Moderada e Grave, as quais estão representadas na base de testes através de 7, 8, 11, 4 casos, respectivamente. O número de casos para cada uma das variações foi obtido de acordo com as probabilidades a *priori* observadas na base utilizada no aprendizado da rede original.

A metodologia adotada para os testes consistiu de colocar na entrada do sistema os valores pertencentes ao caso sob foco e observar as respostas do sistema quanto à primeira e segunda maior probabilidades, as quais são obtidas a partir atualização das

crenças (probabilidades a *posteriori*).

As Tabelas 5.5 e 5.6 ilustram uma parte do procedimento e metodologia observadas na validação do sistema desenvolvido para auxílio à tomada de decisão. Como pode ser observado, as respostas obtidas são bastantes razoáveis, influenciando o usuário dos sistema para uma tomada de decisão segura e sensata.

Analisando as respostas do sistema de acordo com o procedimento adotado para os testes de eficácia presentes na Seção anterior, encontramos que a Taxa de Acerto "**1ª**" é de aproximadamente 86, 67%, enquanto que a "**2ª**" é de 13, 33%. Desta forma, no teste de validação o sistema não apresentou nenhuma resposta que possa ser considerada como errada.

5.5 Produto Final

O programa de realização de inferências foi construído utilizando a linguagem de programação C, o qual permite gerar executáveis funcionado sobre o sistema operacional DOS. Acontece que o público-alvo dos sistema é composto por usuários da classe médica, uma interface funcional baseada em menus típicos do DOS seria pouco atraente. Assim, com o objetivo de dotar a interface humana com o sistema mais amigável, um aplicativo *windows* foi construído utilizando a linguagem C++ . As entradas são as observações dos sintomas e a saída é a atualização da crença sobre o nó de interesse: o da miastenia com suas quatro variações. As Figuras 5.9 e 5.10 ilustram a execução do programa e um exemplo de consulta realizado no sistema, respectivamente.

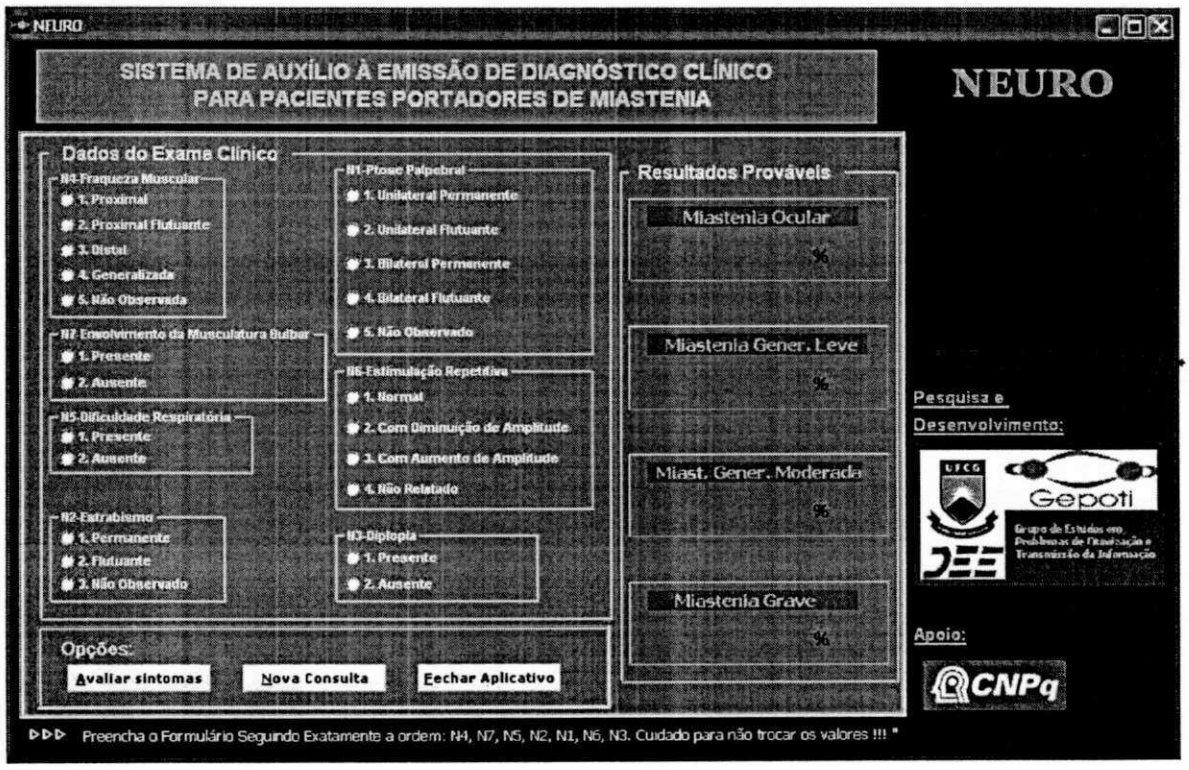


Figura 5.9: Interface do programa de realização de inferências.

N_3	N_6	N_4	N_1	N_0	N_5	N_2	Diagnóstico Médico	Sistema Desenvolvido Resposta "**A**" Resposta"**B**" prob("**A**") prob("**B**")
5	2	2	2	2	2	1	Ocular	Ocular leve 79,12% 12,50%
5	2	2	2	4	2	1	Ocular	Ocular leve 79,12% 12,50%
5	2	2	2	2	1	1	Ocular	Ocular leve 79,12% 12,50%
5	2	2	2	4	1	1	Ocular	Ocular leve 79,12% 12,50%
5	2	2	3	2	1	1	Ocular	Ocular leve 79,12% 12,50%
5	2	2	3	4	1	1	Ocular	Ocular leve 79,12% 12,50%
5	2	2	3	4	2	1	Ocular	Ocular leve 79,12% 12,50%
2	2	2	2	2	2	1	leve	leve moderada 44,74% 44,74%
2	2	2	3	2	2	1	leve	leve moderada **44,74%** 44,74%
2	2	2	2	4	2	1	leve	leve moderada 44,74% 44,74%
2	2	2	3	4	2	2	leve	leve moderada 44,74% 44,74%
2	2	2	3	2	2	2	leve	leve moderada 44,74% 44,74%
2	2	2	2	2	1	1	leve	leve moderada 44,74% 44,74%
2	2	2	3	5	2	2	leve	leve moderada 44,74% 44,74%
2	2	2	2	2	2	1	leve	leve moderada 44,74% 44,74%

Tabela 5.5: Validação do Sistema de Auxílio à Emissão do Diagnóstico Médico: Variações Ocular e Leve.

N_3	N_6	N_4	N_1	N_0	N_5	N_2	Diagnóstico Médico	Sistema Desenvolvido Resposta "**A**" Resposta"**B**" prob("**A**") prob("**B**")
2	1	2	2	2	2	1	moderada	moderada leve 44,74% 44,74%
2	1	2	3	2	2	2	moderada	moderada leve 44,74% 44,74%
2	1	2	3	5	2	2	moderada	moderada leve 44,74% 44,74%
2	1	2	2	2	2	1	moderada	moderada leve 44,74% 44,74%
2	1	2	2	2	2	2	moderada	moderada leve 44,74% 44,74%
4	1	2	2	2	2	1	moderada	moderada leve 41,67% 37,50%
4	1	2	2	4	2	1	moderada	moderada leve 41,67% 37,50%
4	1	2	3	4	2	2	moderada	moderada leve 41,67% 37,50%
4	1	2	3	5	2	2	moderada	moderada leve 41,67% 37,50%
4	1	2	2	5	2	1	moderada	moderada leve 41,67% 37,50%
4	1	2	3	5	2	2	moderada	moderada leve 41,67% 37,50%
4	1	1	2	4	2	1	grave	grave moderada 37,50% 41,67%
4	1	1	3	4	2	2	grave	grave moderada 37,50% 41,67%
4	1	1	2	2	2	1	grave	grave moderada 37,50% 41,67%
4	1	1	3	5	2	2	grave	grave moderada 37,50% 41,67%

Tabela 5.6: Validação do Sistema de Auxílio à Emissão do Diagnóstico Médico: Variações Moderada e Grave.

Figura 5.10: Interface do programa de realização de inferências: exemplo de consulta.

5.6 Conclusões

Este Capítulo apresentou a metodologia para o desenvolvimento do sistema de auxílio à emissão do diagnóstico das variações de Miastenia Gravis, além do procedimento de validação do sistema. Os resultados obtidos com redes bayesianas são bastante promissores, favorecendo esta abordagem quando comparada àquelas utilizadas no desenvolvimento de sistemas especialistas.

O próximo Capítulo é um espaço dedicado às considerações finais e às perspectivas para trabalhos futuros.

Capítulo 6

Conclusões

Neste trabalho é avaliado o desempenho da inferência probabilística em redes bayesianas. Esta abordagem foi utilizada no desenvolvimento de um sistema de auxílio à emissão do diagnóstico médico, o qual proporcionou vantagens sobre as abordagens tradicionais dos sistemas especialistas.

Neste sentido, aprender uma rede bayesiana consiste em escolher a hipótese que melhor modela os dados presentes na base. Esta abordagem é particularmente interessante quando não é possível contar com a presença de especialistas do domínio do problema, ou quando já está disponível um grande volume de dados coletados e, deseja-se extrair o conhecimento humano embutido nesta base. Isto por sua vez, não exclui a possibilidade de ser adicionado o conhecimento especialista na busca pela melhor rede e, ao contrário, sempre que este estiver disponível deve ser utilizado pois isto proporciona uma obtenção do modelo, dentro do espaço de busca, de forma mais rápida, retornando também uma rede mais precisa.

Por outro lado, especificada a rede, esta pode ser usada para a realização de inferências. A inferência probabilística em redes bayesianas consiste na atualização das crenças segundo os valores observados. O método de inferência a ser usado, contudo, é função da topologia da rede. Assim, quando a topologia obtida do aprendizado não pertence a classe associada ao algoritmo de inferência escolhido, deve-se realizar alterações na rede obtida do aprendizado, com o objetivo de adequá-la a esta classe. Para isto, foram propostos critérios de remoção de arcos baseados em conceitos da Teoria da Informação. O objetivo comum a estes critérios é avaliar o impacto das remoções dos arcos na perda de informação associada a simplificação da rede.

Bases de dados de tamanhos diferentes foram utilizadas, sendo as redes bayesianas

aproximadas pelos critérios de remoção de arcos analisadas de acordo com a eficácia obtida na realização de inferências. Apesar da perda de informação inerente ao processo, os resultados obtidos foram bastante satisfatórios.

Com o objetivo de avaliar o comportamento da ordenação dos nós na realização de inferências, em redes bayesianas, foram utilizadas medidas de informação presentes na Teoria da Informação. Conforme esperado, a inferência bayesiana é alterada de acordo com a ordenação adotada para as variáveis presentes na rede. Os testes de eficácia até este momento, foram realizados utilizando a mesma base de treinamento, já que o objetivo é selecionar a rede de maior eficácia para o desenvolvimento do sistema de auxílio à emissão de diagnóstico de pacientes portadores de variações de miastenia gravis.

O sistema desenvolvido possui uma interface amigável e de fácil manuseio, sendo validado utilizando uma base de testes diferente e que corresponde a 40% da base de treinamento. Os resultados obtidos são promissores, possibilitando ao usuário uma tomada de decisão mais segura diante de um cenário de incertezas.

6.1 Perspectivas para Trabalhos Futuros

Como continuação das atividades de pesquisa realizadas, podem ser citadas as seguintes sugestões as quais consideram as particularidade presentes em aplicações práticas que consistem no auxílio à tomada de decisão e as vantagens proporcionadas pelo desenvolvimento de um método que utilize algoritmos iterativos com grafos contendo ciclos

- Desenvolver um método para utilização de algoritmos em grafos contendo ciclos,
- Desenvolver um método de ordenação de variáveis para aumentar a eficácia dos algoritmos envolvidos e
- Aplicar estes métodos a problemas que envolvem a tomada de decisão, proporcionando um auxílio à emissão de diagnóstico

Apêndice A

Fundamentação Teórica

Este Apêndice foi incluído a fim de facilitar a leitura do texto, evitando o leitor buscar fontes externas. Apresenta-se aqui definições e conceitos elementares da Teoria das Probabilidades, extraídos parcialmente do Capítulo 2 do trabalho de Leonardo Matos [37], reproduzidos com o consentimento do autor.

A.1 Revisão de Matemática e Estatística

A definição de probabilidade empregada na maioria dos livros texto em probabilidade e estatística deve-se a Kolmogorov que define probabilidade como um número satisfazendo a um conjunto de axiomas. Na formalização destes axiomas faz-se necessário apresentar algumas definições preliminares.

Definição 14 (Espaço Amostral)

- *Chama-se espaço amostral, denotado por Ω, o conjunto não necessariamente finito de todos os resultados de um experimento aleatório.*

Definição 15 (Evento)

- *Um evento é um conjunto, denotado como E, contido no espaço amostral, $E \subset \Omega$. O subconjunto $E = \Omega$ é também referenciado como um evento certo.*

Para qualquer espaço amostral, Ω, define-se probabilidade de um evento E, $E \subset \Omega$, representado por $P(E)$, um número que satisfaz o seguinte grupo de axiomas

Axioma 1 $0 \leq P(E) \leq 1$

Axioma 2 $P(\Omega) = 1$

Axioma 3 Se $E_1, E_2, ...$ é uma seqüência enumerável de eventos mutuamente exclusivos, isto é, $E_i \cap E_j = \emptyset$, $\forall i, j | i \neq j$, então

$$P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i) \tag{A.1}$$

A partir destes axiomas, pode ser realizado uma série de desenvolvimentos que fundamentam as disciplinas de probabilidade e estatística tal como são conhecidas. Algumas desses desenvolvimentos serão abordados a seguir.

Definição 16 (Probabilidade condicional)

- *Seja B um evento não vazio, o que implica em $P(B) > 0$, a probabilidade condicional de um evento A em relação a B, expressa como $P(A|B)$, é definida como.*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{A.2}$$

Utilizando a notação produto para designar intercessão entre conjuntos, a expressão (A.2) pode ser escrita como

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{A.3}$$

A noção intuitiva de probabilidade condicional, $P(A|B)$, está relacionada com a medição da probabilidade de ocorrência de um evento A dado que um evento B já tenha ocorrido. Esta noção se estende para grupos de vários eventos, assim tem-se que

$$P(A_3|A_1, A_2) = \frac{P(A_1, A_2, A_3)}{P(A_1 A_2)} \tag{A.4}$$

Os termos na Equação (A.4) podem ser reagrupados e combinados com (A.3), conhecida como regra da cadeia,

$$P(A_1, A_2, A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \tag{A.5}$$

de um modo geral tem-se que

$$P(A_1, A_2, ..., A_n) = P(A_1)P(A_2|A_1)...P(A_n|A_1, A_2, ..., A_{n-1})$$

Definição 17 (Independência Estatística)

- *Dois eventos A e B, não vazios, são ditos ser independentes se e somente se*

$$P(A|B) = P(A)$$

Se $P(A|B) = P(A)$ então também é válido que

$$P(B|A) = P(B).$$

Teorema 5 (Teorema de Bayes) *: Sejam $A_1, A_2, A_3, ...$ eventos mutuamente exclusivos e coletivamente exaustivos, isto é, $A_1 \cup A_1 \cup A_1 ... = \Omega$ com $A_i \cap A_j = \emptyset$, $\forall i, j | i \neq j$. Seja B um evento tal que $P(B) > 0$, então*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum P(A_i)P(B|A_i)}$$

O Teorema de Bayes é particularmente importante por ser um meio de relacionar probabilidades a *priori*, $P(A_i)$ com probabilidades a posteriori, $P(A_i|B)$.

A.1.1 Variáveis Aleatórias

A apresentação do Teorema de Bayes normalmente encerra as discussões preliminares sobre cálculo de probabilidades na maioria dos livros textos em Matemática Estatística e probabilidade. Dentro desta seqüência lógica comumente adotada, esta discussão encerra-se quando o leitor está apto a ser apresentado à definição de variáveis aleatórias. O conceito de variáveis aleatórias é fundamental para o desenvolvimento da teoria sobre

funções de variáveis aleatórias ou funções de probabilidades. Particularmente, para os objetivos deste texto a apresentação de algumas destas funções de probabilidade é uma parte muito importante, pois estas foram muito utilizadas nos desenvolvimentos realizados no Capítulo 2.

Definição 18 (Variáveis Aleatórias-va.)

- *Chama-se variável aleatória uma função que mapeia o conjunto Ω (espaço amostral) na reta real* - $\mathbb{R}$.

São exemplos de variáveis aleatórias:

- Número de caras no lançamento de uma moeda;
- Peso dos recém-nascidos em uma maternidade m um dado período;
- Valor em dinheiro na conta bancária de um indivíduo em um dado período.

As variáveis aleatórias são caracterizadas quanto ao tipo de mapeamento que realizam como contínua, se mapeiam em um subconjunto não enumerável da reta real, ou discreta, se o mapeamento é realizado sobre um conjunto enumerável de $\mathbb{R}$. Nos exemplos acima, o primeiro e o último são exemplos de variáveis aleatórias discretas, ao passo que o segundo item exemplifica uma variável aleatória contínua.

Definição 19 (Função de massa de probabilidade)

- *Uma função $f(x) : \mathbb{R} \to [0, 1]$ é chamada função de massa de probabilidade, ou distribuição de probabilidade, se e somente se possuir as seguintes propriedades:*

  i) $f(x) \geq 0$

  ii) $\sum f(x) = 1$

Uma função de massa de probabilidade, $f(x)$, modela a distribuição de probabilidade de uma va. discreta X se $f(x) = P(X = x)$.

Definição 20 (Função de densidade de probabilidade)

- *Uma função* $f(x) : \mathbb{R} \rightarrow \mathbb{R}$ *é chamada função densidade de probabilidade, também referenciada como densidade ou distribuição de probabilidade, se e somente se possuir as seguintes propriedades:*

  i) $f(x) \geq 0$

  ii) $\int_{\infty}^{\infty} f(x)dx = 1$

Uma função densidade de probabilidade, $f(x)$, modela a distribuição de probabilidade de uma va. contínua X se $f(x)$ for definida como:

$$f(x) = \lim_{\epsilon_0 \rightarrow 0} P(x - \epsilon_0 < X < x + \epsilon_0)$$

O termo distribuição é usado neste texto para denotar tanto massa como densidade de probabilidade. Essa ambigüidade deverá ser dirimida pelo contexto, que esclarecerá qual dos conceitos o termo estará fazendo referência.

A.1.2 Distribuições de Probabilidade

Esta Seção apresenta algumas equações de referência que serão muito úteis para os desenvolvimentos posteriores.

Definição 21 (Distribuição binomial)

- *Uma va. discreta* X *é dita ter distribuição binomial se e somente se possuir função de massa de probabilidade definida como*

$$f(x) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^x (1 - \theta)^{n-x}, \tag{A.6}$$

  para $x = 0, 1, ..., n$.

  Os argumentos n *e* θ *são chamados parâmetros da binomial. O parâmetro* θ *está associado a uma proporção, por conseguinte,* $\theta \in [0, 1]$ *e* n *corresponde ao valor máximo que* X *pode assumir. A distribuição binomial, em função de seus parâmetros, é denotada alternativamente como*

$$f(x) = B(x; n, \theta)$$

Definição 22 (Distribuição multinomial)

- *Um conjunto de k de va.'s discretas $\mathbf{X} = \{X_1, X_2, ..., X_k\}$ é dito ter distribuição multinomial se e somente se possuir função de massa de probabilidade conjunta definida como*

$$f(x_1, x_2, ..., x_k) = f(x_1, x_2, ..., x_k; \theta_1, ..., \theta_k, n) = \begin{pmatrix} n \\ x_1, x_2, ..., x_k \end{pmatrix} \theta_1^{x_1} ... \theta_k^{x_k}, \quad (A.7)$$

em que $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} \theta_i = 1$ *e*

$$\begin{pmatrix} n \\ x_1, x_2, ..., x_k \end{pmatrix} = \frac{n!}{x_1!, ..., x_k!}$$

A distribuição multinomial desempenha um papel importante no estudo de redes bayesianas pois estas lidam com variáveis discretas cuja distribuição pode ser modelada por uma distribuição multinomial.

Definição 23 (Função gama)

- *A função gama, denotada como $\Gamma(\cdot)$, é definida pela expressão*

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy,$$

para $\alpha > 0$

Algumas importantes propriedades da função gama são:

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
2. $\Gamma(1) = 1$
3. $\Gamma(\alpha) = (\alpha - 1)!$ para α inteiro
4. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Definição 24 (Distribuição gama)

- *Uma va. contínua X possui distribuição gama se e somente se sua função de densidade for definida como*

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & \text{se } x > 0 \\ 0, & \text{caso contrário} \end{cases} \tag{A.8}$$

em que $\alpha > 0$ e $\beta > 0$.

Um caso particular da distribuição gama é a distribuição exponencial negativa que ocorre quando $\alpha = 1$, cuja função de densidade é dada por

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\beta}, & \text{se } x > 0 \\ 0, & \text{caso contrário} \end{cases} \tag{A.9}$$

em que $\theta > 0$.

Definição 25 (Função beta)

- *A função beta, denotada como $B(\cdot)$, é definida por*

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx,$$

com $\alpha > 0$ e $\beta > 0$.

A função beta relaciona-se com a função gama através da expressão

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Definição 26 (Distribuição beta)

- *Uma va. contínua, X, possui distribuição beta se e somente se sua função de densidade for definida como*

$$f(x;\alpha,\beta) = Beta(x;\alpha,\beta) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{para } 0 < x < 1 \\ 0, & \text{caso contrário} \end{cases} \tag{A.10}$$

em que $\alpha > 0$ e $\beta > 0$.

A distribuição beta é particularmente importante por se tratar de uma distribuição complementar da binomial. O sentido de complementariedade se origina do fato de que a distribuição Beta modela o comportamento das proporções p e $q = 1 - p$ de uma binomial $B(x; n, p)$, considerando n fixo.

Teorema 6 : *Seja X uma va. com distribuição $Beta(\alpha, \beta)$. O valor esperado de X é dado por*

$$E[X] = \frac{\alpha}{\alpha + \beta}$$

Prova:

$$E[X] = \frac{1}{B(\alpha, \beta)} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

$$= \frac{1}{B(\alpha, \beta)} B(\alpha + 1, \beta)$$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + 1)\Gamma(\beta)}{\Gamma(\alpha + \beta) + 1}$$

$$= \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta) + 1}$$

$$= \frac{\alpha}{\alpha + \beta}$$

■

Um caso particular da distribuição beta é a distribuição uniforme no intervalo [0,1], definida por $f(x) = 1$, $0 \leq x \leq 1$, que ocorre quando os valores de α e β são ambos iguais a 1.

Definição 27 (Distribuição Dirichlet)

- *Um conjunto de k va.'s contínuas $\theta = \{\theta_1, \theta_2, ..., \theta_k\}$ possui distribuição de probabilidade conjunta Dirichlet se sua função de distribuição for expressa como*

$$f(\theta_1, \theta_2, ..., \theta_k; \nu_1, \nu_2, ..., \nu_k, \nu_{k+1}) = \tag{A.11}$$

$$= D(\theta_1, \theta_2, ..., \theta_k; \nu_1, \nu_2, ..., \nu_k, \nu_{k+1})$$

$$= \frac{\Gamma(\nu_1 + \nu_2 + ... + \nu_k + \nu_{k+1})}{\Gamma(\nu_1)\Gamma(\nu_2)...\Gamma(\nu_k)\Gamma(\nu_{k+1})} \theta_1^{\nu_1 - 1} ... \theta_k^{\nu_k - 1} (\theta_{K+1})^{\nu_{k+1} - 1}$$

com $\sum_i = 1$.

Pode-se constatar que a distribuição beta é um caso particular da distribuição Dirichlet, pois quando $k = 1$, $D(\theta, 1 - \theta; \nu_1, \nu_2)$ iguala-se a $Beta(x; \nu_1, \nu_2)$.

A.1.3 Inferência Estatística

Inferência é a parte da estatística que estuda métodos que procuram estabelecer o valor de um parâmetro populacional com base em dados amostrais. Este texto apresenta dois dos principais métodos de inferência estatística que são: método de máxima verossimilhança e inferência bayesiana. O segundo método é o mais importante para os propósitos do estudo de redes bayesianas.

Método da máxima verossimilhança

O método da máxima verossimilhança proposto R. Fisher em 1922, forma a principal técnica de estimação de parâmetros da escola clássica de estatística. A idéia deste método possui um apelo intuitivo bastante simples, o estimador, $\widehat{\theta}$, do parâmetro populacional desconhecido, é aquele que maximiza a probabilidade de em uma amostra $\{X_1, ..., X_n\}$ as variáveis aleatórias terem sido obtidas de uma população com parâmetro Θ.

No caso genérico, o estimador $\widehat{\theta}$ corresponde ao valor que maximiza a distribuição $f(x_1, ..., x_n; \theta)$. Esta função de distribuição, representada por $L(\theta; x_1, ..., x_n)$ é denominada função de máxima verossimilhança. Para manter coerência com a notação usada nas funções de distribuição de probabilidade, a função de máxima verossimilhança deverá ser escrita como $L(\theta; x_1, ..., x_n)$. O valor de $L(\theta; x)$ é dado por

$$L(\theta; x) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1 - \theta)^{n-x}, \tag{A.12}$$

Usualmente, $\widehat{\theta}$ é encontrado pela minimização da função logaritmo de L, dado que a função logaritmo é monotonicamente crescnte e sua derivada possui propriedades matemáticas atraentes. Assim, o valor de $\widehat{\theta}$ seria obtido pela solução de

$$\widehat{\theta} = \arg\max_{\theta}\{\log \binom{n}{x} \theta^x (1-\theta)^{n-x}\} \tag{A.13}$$

A derivada de A.13 em relação a θ é dada por:

$$L(\theta; x) = \log[\binom{n}{x} \theta^x (1-\theta)^{n-x}]$$

$$\Rightarrow \frac{dL(\theta;x)}{d\theta} = \frac{d\log\binom{n}{x}}{d\theta} + x\frac{d\log\theta}{d\theta} + (n-x)\frac{d\log(1-\theta)}{d\theta}$$

$$\Rightarrow \frac{dL(\theta;x)}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

Igualando $\frac{dL(\theta;x)}{d\theta}$ a zero, obtém-se

$$\frac{x}{\widehat{\theta}} = \frac{n-x}{1-\widehat{\theta}}$$

$$\Rightarrow \widehat{\theta} = \frac{x}{n}.$$

Formalmente, define-se a função e estimador de máxima verossimilhança como

Definição 28 (Função de Máxima Verossimilhança)

- *Dado uma amostra de tamanho n,* $\mathbf{X} = \{X_1, ..., X_n\}$, *define-se função de máxima verossimilhança, denotada por* $L(\cdot)$, *a distribuição conjunta de* $\mathbf{X}$ *em relação a um parâmetro* θ, *isto é,* $f(x_1, ..., x_n; \theta)$. *Como L é função de* θ, *utiliza-se a notação* $L(\theta; x_1, ..., x_n)$ *para exprimir uma relação de ordem que indica ser* θ *uma variável e* $\{x_1, ..., x_n\}$ *valores constantes.*

Definição 29 (Estimador de Máxima Verossimilhança)

- *Seja* $L(\theta) = L(\theta; x_1, ..., x_n)$ *a função de máxima verossimilhança para uma amostra* $X_1, ..., X_n$. *Define-se o estimador de máxima verossimilhança de* θ *para os valores amostrais obtidos, denotado por* $\widehat{\theta}$, *como argumento que maximiza* $L(\theta)$, *isto é,*

$$\widehat{\theta} = \arg\max_{\theta}\{L(\theta)\}$$

A.1.4 Inferência Bayesiana

A estimação bayesiana difere da estimação clássica essencialmente pela interpretação do parâmetro estimado, denotado como no caso anterior pela letra $\widehat{\theta}$. A escola clássica interpreta $\widehat{\theta}$ como um valor constante, fixo, ao passo que a escola bayesiana interpreta $\widehat{\theta}$ como uma variável aleatória, assim, a crença no verdadeiro valor do parâmetro populacional estimado pode ser mudada em rodadas diferentes de um experimento aleatório. As escolas clássicas e bayesianas concordam no procedimento de levantamento de dados para realizar a estimação, ambas utilizam um processo de amostragem, mas diferem quanto a interpretação dos dados obtidos. A estimação bayesiana objetiva determinar qual o valor que a variável aleatória θ pode ter assumido ao ter sido obtida a amostra.

Para dirimir quaisquer dúvidas quanto à discussão levantada no parágrafo anterior, um breve estudo de caso pode ser apresentado e analisado sob a luz do dissernimento entre as diferenças ligadas às escolas clássica e bayesiana. Considere que um processo fábrico produz uma certa proporção de peças defeituosas. Coletado uma amostra, o método clássico estimaria um determinado valor para a verdadeira proporção de peças defeituosas produzidas, acreditando ser esta valor fixo. O método bayesiano também utilizaria os dados amostrais para realizar a estimação, mas ao contrário da escola clássica, assumiria que esta proporção não tem um valor fixo e que, portanto, deve sofrer variações com base em uma função de probabilidade. Com base nos dados amostrais, a escola bayesiana determinaria o valor da proporção de peças defeituosas que estaria sendo produzida no instante em que a amostra fora coletada.

A idéia empregada na estimação bayesiana pode ser descrita como segue. Parte-se de um conhecimento a priori sobre a distribuição de θ, este conhecimento a priori é vago. Com base nos dados amostrais, determina-se uma distribuição $f(\theta|\mathbf{x})$, que concentra em uma região mais estreita do eixo horizontal grande parte da área da curva. A informação oriunda de $f(\theta|\mathbf{x})$ é, portanto, mais específica e restringe a um intervalo

pequeno os valores mais prováveis que θ pode assumir. A estimação do parâmetro populacional, $\widehat{\theta}$, constitui o valor esperado de θ com base na distribuição $f(\theta|\mathbf{x})$. Este valor tamb´m possui uma interpretação geométrica, $\widehat{\theta}$ corresponde ao centro de massa da curva $f(\theta|\mathbf{x})$, que também pode ser interpretado como uma média ponderada infinita dos valores de θ, em que os pesos são dados pelos valores de $\frac{f(\theta|\mathbf{x})}{\int f(\mathbf{x}|\theta)d\theta}$.

No caso geral, assume-se que θ tenha uma distribuição contínua. Adicionalmente, se a variável X possui distribuição binomial a estimação de θ pode ser realizada rapidamnte com base nas propriedades enunciadas nos Teoremas 7 e 8, ambos apresentados sem prova.

Teorema 7 : *Dado que X seja uma variável aleatória discreta com distribuição binomial $B(n,\theta)$, dado que a distribuição $h(\theta)$ seja Beta, com parâmetros α e β, então $h(\theta)$ é beta com parâmetros $\alpha + x$ e $\beta + x$.*

Teorema 8 : *Dado que $h(\theta)$ possui distribuição Beta, com parâmetros α e β, e dado que $f(x)$ seja binomial, $B(x;n,\theta)$, então o valor esperado de $E_{h(\theta|\mathbf{x})}[\theta]$ é,*

$$E_{h(\theta|\mathbf{x})}[\theta] = \frac{\alpha + x}{\alpha + \beta + n} \tag{A.14}$$

Os Teoremas 7 e 8 são particularmente úteis porque fornecem a fundamentação matemática para estimação das probabilidades condicionais $P(X|Y)$ entre nós adjacentes em uma rede bayesiana com variáveis aleatórias binomiais.

Os Teoremas 9 e 10, a seguir, generalizam os resultados anunciados nos Teoremas 7 e 8.

Teorema 9 : *Seja $\mathbf{X} = \{X_1, ..., X_n\}$ um conjunto de variáveis aleatórias discretas com distribuição conjunta multinomial com parâmetros $\mathbf{\Theta} = \{\Theta_1, ..., \Theta_n\}$. Dado que $h(\theta)$, a distribuição conjunta de $\{\Theta_1, ..., \Theta_n\}$ seja Dirichlet com parâmetros $\{\nu_1, ..., \nu_n\}$, então $h(\theta|\mathbf{x})$ é Dirichlet com parâmetros $\{x_1 + \nu_1, x_2 + \nu_2, ..., x_n + \nu_n\}$.*

Teorema 10 : *Seja $\mathbf{\Theta} = \{\Theta_1, ..., \Theta_n\}$ um conjunto de variáveis aleatórias contínuas com distribuição conjunta de Dirichlet com parâmetros $\{\nu_1, ..., \nu_n\}$. Sendo $\mathbf{X} = \{X_1, ..., X_n\}$ um conjunto de variáveis aleatórias discretas com distribuição conjunta multinomial com parâmetros $\mathbf{\Theta}$, então o valor esperado de θ_i dado $\mathbf{x}$, isto é, $E_{h(\theta|\mathbf{x})}[\theta_i]$ é dado por*

$$E_{h(\theta|\mathbf{x})}[\theta_i] = \frac{\nu_i + x_i}{\sum_i^n \nu_i + \sum_{i=1}^n x_i} \tag{A.15}$$

No caso geral, quando as variáveis de uma rede bayesiana podem assumir mais de dois valores discretos, o cálculo das probabilidades condicionais envolvendo nós adjacentes é feita com base nas propriedades dos Teoremas 9 e 10, cujos resultados são análogos quando considerando o par de distribuições complementares Multinomial/Dirichlet.

A.2 Prova do Teorema 2:

O objetivo deste Teorema é encontrar uma expressão que fornece o valor de $P(B_S, D)$, então, partindo-se da definição de $P(B_S, D)$ tem-se que:

$$P(B_S, D) = P(B_S)P(D|B_S) = P(B_S) \int_{B_P} P(D|B_S, B_P) f(B_P|B_S) d(B_P) \quad \text{(A.16)}$$

A notação acima é adequada, visto que a integral na expressão A.16 pode ser analisada como uma soma ponderada infinita das probabilidades dos dados terem sido originados de uma rede $B = \{B_S, B_P\}$, mantendo-se fixo B_S e variando-se B_P, sendo esta variação sobre todo o espaço onde B_P é definido e os pesos são dados pelo valor de $f(B_P|B_s)dB_P$, que ser comparado à probabilidade de B_P dado B_S. Portanto, a integral corresponde ao valor esperado de $P(D|B_S, B_P)$. Utiliza-se o valor esperado porque supõe-se que não seja conhecido B_P dado B_S e, assim, calcula-se $P(D|B_S)$ considerano-se todos os possíveis valores de B_P.

O termo $P(D|B_S, B_P)$ pode ser expandido. Suponha que o conjunto D seja completo, isto é, em cada caso existe observações sobre cada variável X_i. Considere que os dados possam ser organizados da seguinte forma: admita que o conjunto de dados D é obtido a partir de uma rede com n variáveis. Uma vez que B_P é o elemento desconhecido e que se conhece B_S, suponha de agora em diante, que a base de dados possa ser reagrupada conforme o arranjo abaixo.

$$D = \left\{ \begin{array}{cccc} N_{111}(x_{11}, \mathbf{pa_{11}}), & N_{112}(x_{12}, \mathbf{pa_{11}}), & \ldots & N_{11r_i}(x_{1r_i}, \mathbf{pa_{11}}) \\ N_{121}(x_{11}, \mathbf{pa_{12}}), & N_{122}(x_{12}, \mathbf{pa_{12}}), & \ldots & N_{11r_i}(x_{1r_i}, \mathbf{pa_{12}}) \\ \vdots & \vdots & \vdots & \vdots \\ N_{1q_i1}(x_{11}, \mathbf{pa_{1q_i}}), & N_{1q_i2}(x_{12}, \mathbf{pa_{1q_1}}), & \ldots & N_{1q_1r_1}(x_{1r_1}, \mathbf{pa_{1q_1}}) \\ \vdots & \vdots & \vdots & \vdots \\ N_{n11}(x_{n1}, \mathbf{pa_{n1}}), & N_{n12}(x_{n2}, \mathbf{pa_{n1}}), & \ldots & N_{n1r_n}(x_{nr_n}, \mathbf{pa_{n1}}) \\ \vdots & \vdots & \vdots & \vdots \\ N_{nq_n1}(x_{n1}, \mathbf{pa_{nq_n}}), & N_{nq_n2}(x_{n2}, \mathbf{pa_{nq_n}}), & \ldots & N_{nq_nr_n}(x_{nr_n}, \mathbf{pa_{nq_n}}) \end{array} \right\}. \quad \text{(A.17)}$$

Utiliza-se o índice r_i para denotar a quantidade máxima de instâncias de um nó X_i, enquanto q_i denota a quantidade máxima de instâncias do conjunto $\mathbf{Pa_i}$. Ao longo desta seção, as letras i, j e k serão usadas com o seguinte sentido: a letra i denota o índice do nó, o qual pode variar de 1 até n, a letra j denota o índice dos pais de um nó e k, o índice de uma instância de um nó.

Admitindo que os dados em D possam ser organizados desta forma, e supondo que as tuplas que formam D sejam geradas independentemente umas das outras, tem-se que:

$$P(D|B_S, B_P) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (P(x_{ik}|\mathbf{pa_{ij}}, B_P))^{N_{ijk}} \quad \text{(A.18)}$$

Substituindo-se a Equação(A.18) em Equação(A.16), tem-se que:

$$P(B_S, D) = \int_{B_P} \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} (P(x_{ik}|\mathbf{pa_{ij}}, B_P))^{N_{ijk}} f(B_P|B_S) d(B_P). \quad \text{(A.19)}$$

Fazendo $\theta_{ijk} = P(x_{ik}|\mathbf{pa_{ij}})$ e supondo qu $f(\theta_{ij1}, ..., \theta_{ijr_i})$ seja marginalmente independente de $f(\theta_{i'j'1}, ..., \theta_{i'j'r'_i})$ com $i \neq i'$ e $j \neq j'$. Isto é, θ_{ijk} é independente de $\theta_{i'j'k'}$ se $i \neq i'$ ou $j \neq j'$, assim, tem-se que:

$$f(B_P|B_S) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} f(\theta_{ij1}, ..., \theta_{ijr_i}). \quad \text{(A.20)}$$

Substituindo a Equação (A.20) na Equação (A.19) resulta em:

$$P(B_S, D) = P(B_S) \int_{\theta_{ijk}} \dots \int \left[\prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right] \left[\prod_{i=1}^{n} \prod_{j=1}^{q_i} f(\theta_{ij1}, ..., \theta_{ijr_i}) \right] d\theta_{111} \dots d\theta_{nq_nr_n} =$$

$$= P(B_S) \int_{\theta_{ijk}} \dots \int \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\left(\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right) f(\theta_{ij1}, ..., \theta_{ijr_i}) \right] d\theta_{111} \dots d\theta_{nq_nr_n} \quad \text{(A.21)}$$

uma vez que $f(\theta_{ij1}, ..., \theta_{ijr_i})$ é marginalmente independente de $f(\theta_{i'j'1}, ..., \theta_{i'j'r'_i})$ então:

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int_{\theta_{ijk}} \dots \int \left[\left(\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right) f(\theta_{ij1}, ..., \theta_{ijr_i}) \right] d\theta_{1j1} \dots d\theta_{ijr_i} \quad \text{(A.22)}$$

Supondo que para qualquer i e j, $f(\theta_{ij1}, ..., \theta_{ijr_i})$ seja uniforme, isto é, uma vez que não existe conhecimento *a priori* sobre B_P, supõe-se que as proporções $\frac{N_{ijk}}{N_{ij}}$ sejam equiprováveis. Então tem-se que:

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int_{\theta_{ijk}} \dots \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right] C_{ij} d\theta_{1j1} \dots d\theta_{ijr_i}, \quad \text{(A.23)}$$

em que C_{ij} é uma constante.

Como $f(\theta_{ij1}, ..., \theta_{ijr_i})$ forma uma distribuição uniforme e $\sum_{k=1}^{r_i} \theta_{ijk} = 1$ então $f(\Theta)$ é dirichlet com parâmetros $\nu_{ij1} = \nu_{ij2} = ... = \nu_{ijr_i} = 1$, logo:

$$\begin{aligned} C_{ij} &= f(\theta_{ij1}, ..., \theta_{ijr_i}; \nu_{ij1}, \nu_{ij2}, ..., \nu_{ijr_i}) = \\ &= \frac{\Gamma(\nu_{ij1} + ... + \nu_{ijr_i})}{\Gamma(\nu_{ij1}) ... \Gamma(\nu_{ijr_i})} \theta_{ij1}^{\nu_{ij1}-1}, ..., \theta_{ijr_i}^{\nu_{ijr_i}-1} = \\ &= \frac{\Gamma(\overbrace{1 + ... + 1}^{r_i vezes})}{\Gamma(1) ... \Gamma(1)} \\ &= \Gamma(r) = (r-1)! \end{aligned} \quad \text{(A.24)}$$

Substituindo-se a Equação (A.24) na Equação (A.23), tem-se que

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} (r_i - 1)! \int_{\theta_{ijk}} \dots \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right] d\theta_{1j1} \dots d\theta_{ijr_i}, \quad \text{(A.25)}$$

A integral múltipla expressa na Equação (A.25) é uma integral Dirichlet, portanto, tem-se que:

$$\int_{\theta_{ijk}} \cdots \int \frac{\Gamma(\nu_{ij1} + ... + \nu_{ijr_i})}{\Gamma(\nu_{ij1})...\Gamma(\nu_{ijr_i})} \theta_{ij1}^{N_{ij1}} ... \theta_{ijr_i}^{N_{ijr_i}} d\theta_{1j1} \ldots d\theta_{ijr_i} = 1, \tag{A.26}$$

Logo $\nu_{ijk} = N_{ijk} + 1$, o que implica em:

$$\frac{\Gamma(N_{ij1} + 1 + N_{ij2} + 1... + N_{ijr_i} + 1)}{\Gamma(N_{ij1})\Gamma(N_{ij2})...\Gamma(N_{ijr_i+1})} \int_{\theta_{ijk}} \cdots \int \theta_{ij1}^{N_{ij1}} ... \theta_{ijr_i}^{N_{ijr_i}} d\theta_{1j1} \ldots d\theta_{ijr_i} = 1$$

$$\Rightarrow \int_{\theta_{ijk}} \cdots \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right] d\theta_{1j1} \ldots d\theta_{ijr_n} = \frac{\Gamma(N_{ij1})\Gamma(N_{ij2})...\Gamma(N_{ijr_i+1})}{\Gamma(N_{ij1} + 1 + N_{ij2} + 1... + N_{ijr_i} + 1)} \tag{A.27}$$

Como N_{ijk} é inteiro, usando a Propriedade III da função Gama, tem-se que

$$\int_{\theta_{ijk}} \cdots \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \right] d\theta_{111} \ldots d\theta_{nq_nr_n} = \frac{\prod_{k=1} r_i N_{ijk}!}{(N_{ij} + r_i - 1)!}, \tag{A.28}$$

com $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

Substituindo Equação (A.28) em (A.25) e reagrupando os termos que não variam com k, tem-se que

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{A.29}$$

■

A.3 Demonstração da Métrica MDL (Definição 10):

Baseado na Equação (2.4), a medida de descrição de B_S e D é obtida aplicando-se o negativo do logaritmo na base 2 de $P(B_S, D)/P(D)$. Como já mencionado anteriormente, interessa apenas conhecer uma expressão para $P(B_S, D)$, já que a busca para obter a medida de descrição mínima é realizada no espaço de hipóteses, $\mathcal{H}$, que portanto não depende do termo $P(D)$.

O trabalho de Bouckaert se baseia em trabalhos anteriores de Cooper e Herskovits [9] [25] tomando como ponto de partida a expressão de $P(B_S, D)$ fornecida pelo Teorema 2. Faz-se inicialmente:

$$\begin{aligned} L(B_S, D) &= -\log P(B_S, D) = \\ &= -\log\left(P(B_S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}!\right) = \\ &= -\log P(B_S) - \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left\{\log(r_i-1)! - \log(N_{ij}+r_i-1)! + \sum_{k=1}^{r_i}\log N_{ijk}!\right\} \end{aligned} \tag{A.30}$$

Considere inicialmente a contribuição do termo limitado a uma instância de x_i e $\mathbf{pa_{ij}}$, isto é, considere a expressão em:

$$\log(r_i-1)! - \log(N_{ij}+r_i-1)! + \sum_{k=1}^{r_i}\log N_{ijk}! \tag{A.31}$$

Expandindo o termo $\log(N_{ij}+r_i-1)!$ tem-se que:

$$\begin{aligned} &\log(r_i-1)! - \log(N_{ij}+r_i-1)! + \sum_{k=1}^{r_i}\log N_{ijk}! = \\ &= \log(r_i-1)! - \log(N_{ij}+r_i-1) - \log(N_{ij}+r_i-2) - \ldots - \log(N_{ij})! + \sum_{k=1}^{r_i}\log N_{ijk}! \end{aligned} \tag{A.32}$$

Considerando apenas os dois últimos termos da expressão acima e considerando a aproximação de Stirling para $N!$ quando N é suficientemente grande, $N! \approx \sqrt{2\pi N}(N/e)^N$, tem-se que:

$$
\begin{aligned}
-\log N_{ij}! + \sum_{k=1}^{r_i} \log N_{ijk}! \approx & \\
& -\log \sqrt{2\pi N_{ij}} (\frac{N_{ij}}{e})^{N_{ij}} + \sum_{k=1}^{r_i} \log \sqrt{2\pi N_{ijk}} (\frac{N_{ijk}}{e})^{N_{ijk}} \\
& = -\frac{1}{2}\log 2\pi - (N_{ij} + \frac{1}{2}) \log N_{ij} + N_{ij} \log e \\
& + \sum_{k=1}^{r_i} \frac{1}{2} \log 2\pi - \sum_{k=1}^{r_i} N_{ijk} \log e + \frac{1}{2} \sum_{k=1}^{r_i} \log N_{ijk} + \sum_{k=1}^{r_i} N_{ijk} \log N_{ijk}
\end{aligned}
\tag{A.33}
$$

Como $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ então os termos $N_{ij} \log e$ e $-\log e \sum_{k=1}^{r_i} N_{ijk}$ se anulam na expressão acima. O termo $-N_{ij} \log N_{ij}$ pode ser reescrito como $\sum_{k=1}^{r_i} -N_{ijk} \log N_{ij}$. Reagrupando-os tem-se que:

$$
\begin{aligned}
-\log N_{ij}! + \sum_{k=1}^{r_i} \log N_{ijk}! \approx & \\
& \frac{(r_i - 1)}{2} \log 2\pi - \frac{1}{2} \log N_{ij} + \frac{1}{2} \sum_{k=1}^{r_i} \log N_{ijk} + \sum_{k=1}^{r_i} N_{ijk} \log N_{ijk}
\end{aligned}
\tag{A.34}
$$

Para N suficientemente grande o termo $\frac{(r_i-1)}{2} \log 2\pi$ pode ser negligenciado em relação ao restante da equação. Realizando esta aproximação, adiciona-se um erro de ordem $O(1)$. Tem-se assim que:

$$
\begin{aligned}
-\log N_{ij}! + \sum_{k=1}^{r_i} \log N_{ijk}! \approx & \\
& -\frac{1}{2} \log N_{ij} + \frac{1}{2} \sum_{k=1}^{r_i} \log N_{ijk} + \sum_{k=1}^{r_i} N_{ijk} \log N_{ijk} + O(1)
\end{aligned}
\tag{A.35}
$$

Para N suficientemente grande o valor de $-\log \prod_{l=1}^{r_i-1} (N_{ij} + l)$ em (A.35) pode ser aproximada por $-\log N_{ij}^{r_i-1}$. Esta aproximação introduz um erro $\sum_{l=1}^{r_i} \log \frac{N_{ij}+l}{N_{ij}}$. Ocorre que $\log \frac{N_{ij}+l}{N_{ij}} < \log l$ para $l > 1$. Como l é limitado superiormente por $r_i - 1$ então:

$$\sum_{l=1}^{r_i} \log \frac{N_{ij}+l}{N_{ij}} < \sum_{l=1}^{r_i} \log(r_i - 1) = (r_i - 1)\log(r_i - 1). \tag{A.36}$$

Logo, o erro introduzido é de ordem $O(1)$, pois sendo limitado superiormente por uma constante não depende de N. Tem-se portanto que:

$$-\log \prod_{l=1}^{r_i-1} (N_{ij}+l) = -\log N_{ij}^{(r_i-1)} + O(1). \tag{A.37}$$

Substituindo (A.36) e (A.37) em (A.35) tem-se que:

$$\begin{aligned}
&\log(r_i-1)! - \log(N_{ij}+r_i-1)! + \sum_{k=1}^{r_i} \log N_{ijk}! \\
&= -(r_i-1)\log N_{ij} - \frac{1}{2}\log N_{ij} + \frac{1}{2}\sum_{k=1}^{r_i} \log N_{ijk} + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= -(r_i - \frac{1}{2})\log N_{ij} + \frac{1}{2}\sum_{k=1}^{r_i} \log N_{ijk} + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= -(r_i - \frac{1}{2})\log N_{ij} + \frac{1}{2}\log \prod_{k=1}^{r_i} N_{ijk} + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= \log\left(N_{ij}^{-(r_i-\frac{1}{2})} (\prod_{k=1}^{r_i} N_{ijk})^{\frac{1}{2}} \right) + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= \log\left(\frac{\sqrt{N_{ij}}}{N_{ij}^{r_i}} \sqrt{\prod_{k=1}^{r_i} N_{ijk}} \right) + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= \log\left(\frac{\sqrt{\frac{N_{ij}}{N}} N^{\frac{1}{2}} \sqrt{\prod_{k=1}^{r_i} \frac{N_{ijk}}{N}} N^{\frac{r_i}{2}}}{(\frac{N_{ij}}{N})^{r_i} N^{\frac{r_i}{2}}} \right) + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= \log\left(\frac{\sqrt{\prod_{k=1}^{r_i} \frac{N_{ijk}}{N}} \sqrt{\frac{N_{ij}}{N}}}{(\frac{N_{ij}}{N})^{r_i}} N^{\frac{r_i}{2}+\frac{1}{2}-r_i} \right) + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1) \\
&= \frac{1}{2}\log \prod_{k=1}^{r_i} \frac{N_{ijk}}{N} + \frac{1}{2}\log \frac{N_{ij}}{N} - r_i \log \frac{N_{ij}}{N} + (\frac{1}{2} - \frac{r_i}{2})\log N + \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} + O(1)
\end{aligned} \tag{A.38}$$

Para N suficientemente grande as proporções $\frac{N_{ijk}}{N}$ e $\frac{N_{ij}}{N}$ convergem respectivamente para as probabilidades $P(x_{ik}|\mathbf{pa_{ijk}})$ e $P(\mathbf{pa_{ij}})$. Logo, os três primeiros termos da ex-

pressão anterior convergem para valores constantes que não dependem de N, podendo ser eliminados pela introdução de mais um erro de ordem O(1).

$$\begin{aligned} &\log(r_i-1)! - \log(N_{ij}+r_i-1)! + \sum_{k=1}^{r_i} \log N_{ijk}! \\ &= -\frac{(r_i-1)}{2}\log N + \sum_{k=1}^{r_i} N_{ijk}\log\frac{N_{ijk}}{N_{ij}} + O(1) \end{aligned} \tag{A.39}$$

Aplicando-se o somatório sobre i e j, chega-se a expressão:

$$L(B_S, D) = -\log P(B_S) - \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(-\frac{(r_i-1)}{2}\log N + \sum_{k=1}^{r_i} N_{ijk}\log\frac{N_{ijk}}{N_{ij}}\right) + O(1) \tag{A.40}$$

$$=$$

Como $q_i = \prod_{X_j \in \mathbf{Pa_i}} r_j$ e $\frac{(r_i-1)}{2}\log N$ não dependem de j, então:

$$\begin{aligned} L(B_S, D) = &-\log P(B_S) - \\ &-\sum_{i=1}^{n}\left[\left(\prod_{X_j\in\mathbf{Pa_i}} r_j\right)(r_i-1)\frac{\log N}{2} + N\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\frac{N_{ijk}}{N}\log\frac{N_{ijk}}{N_{ij}}\right] + C \end{aligned} \tag{A.41}$$

A Expressão (A.41) pode ser reescrita como:

$$L(B_s, D) = -\log P(B_S) + NH(B_s, D) + \frac{\log N}{2}K + C. \tag{A.42}$$

Sendo

$$H(B_s, D) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} -\frac{N_{ijk}}{N}\log\frac{N_{ijk}}{N_{ij}}, \tag{A.43}$$

e,

$$K = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} - \left(\prod_{X_j \in \mathbf{Pa_i}} r_j \right) (r_i - 1). \tag{A.44}$$

■

A.4 Demonstração do Teorema 4:

O Teorema é provado para o caso em que cada nó tem precisamente 2 filhos. O caso de uma árvore arbitrária é então uma genralização direta. Seja D_x o subconjunto de $\mathtt{A}$ contendo os membros de $\mathtt{A}$ que estão na sub-árvore enraizada em X (logo, inclui X se $X \in \mathtt{A}$); denote N_X pelo subconjunto de $\mathtt{A}$ contendo todos os membros de $\mathtt{A}$ que não descendem de X. Recorde que X não descende de X; logo, este conjunto inclui X se $X \in \mathtt{A}$. Esta situação é esboçada na Figura A.1. Logo, tem-se para cada valor de x,

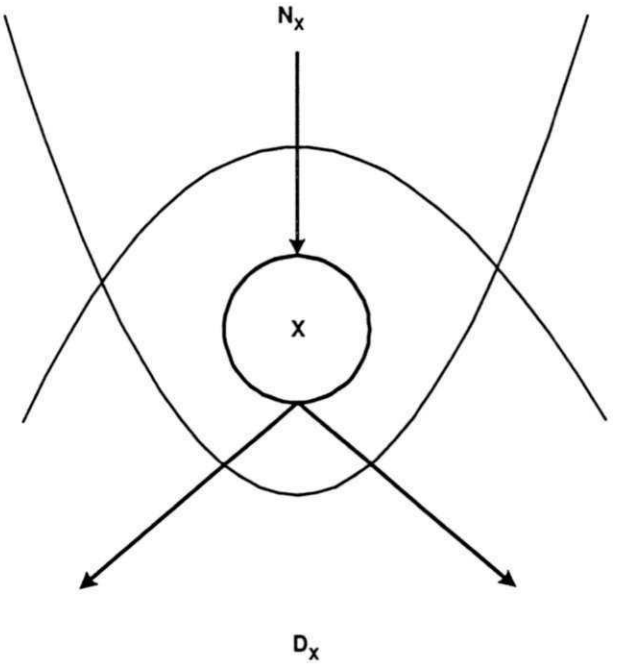


Figura A.1: Conjunto de Variáveis instanciadas $A = N_X \cup D_X$. Se $X \in \mathtt{A}$, X está tanto em N_X quanto D_X.

$$
\begin{aligned}
P(x|a) &= P(x|d_x, n_x) \\
&= \frac{P(d_x, n_x|x)P(x)}{P(d_x, n_x)} \\
&= \frac{P(d_x|x)P(n_x|x)P(x)}{P(d_x, n_x)} \\
&= \frac{P(d_x|x)P(x|n_x)P(n_x)P(x)}{P(x)P(d_x, n_x)} \\
&= \beta P(d_x|x)P(x|n_x),
\end{aligned}
\tag{A.45}
$$

onde β é uma constante que não depende do valor x. A segunda e a quarta igualdade são obtidas da aplicação do Teorema de Bayes. A terceira igualdade segue diretamente da d-separação.

É possível desenvolver funções $\lambda(x)$ $\pi(x)$ tais que:

$$\lambda(x) \backsimeq P(d_x|x)$$

e

$$\pi(x) \backsimeq P(x|n_x),$$

onde $\backsimeq$ significa "proporcional a". Isto é, $\pi(x)$, por exemplo, pode não igualar com $P(x|n_x)$, mas certamente ele iguala por uma constante vezes $P(x|n_x)$, sendo esta constante não dependente de x. Uma vez procedendo desta forma, devido a igualdade mantida pela Equação (A.45), tem-se que:

$$P(x|a) = \alpha\lambda(x)\pi(x)$$

1. Desenvolvendo $\lambda(x)$:

   É necessário mostrar que,

$$\lambda(x) \backsimeq P(d_x|x) \tag{A.46}$$

   Caso 1: $X \in \mathtt{A}$ e o valor de X é $\widehat{x}$. Desde que $X \in D_X$,

$$P(d_x|x) = 0$$

para todo $x \neq \widehat{x}$.

Desta forma, para encontrar a proporcionalidade expressa pela Equação (A.46), é possível ajustar:

$$\begin{aligned} \lambda(\widehat{x}) &\equiv 1 \\ \lambda(x) &\equiv 0, \end{aligned} \tag{A.47}$$

para todo $x \neq \widehat{x}$.

Caso 2: $X \notin \mathtt{A}$ e X é uma folha. Neste caso, $d_x = \emptyset$ e então:

$$P(d_x|x) = P(\emptyset|x) = 1,$$

para todos os valores de x.

Logo, para encontrar a proporcionalidade assegurada pela Equação (A.46), é possível ajustar:

$$\lambda(x) \equiv 1,$$

para todos os valores de x.

Caso 3: $X \notin \mathtt{A}$ e X não é uma folha. Seja Y o filho esquerdo de X e W o filho direito de X. Então, desde que $X \notin \mathtt{A}$,

$$D_X = D_Y \cup D_W.$$

Esta situação é esboçada na Figura A.2. Tem-se que:

$$\begin{aligned} P(d_x|x) &= P(d_y, d_w|x) \\ &= P(d_y|x)P(d_w|x) \\ &= \sum_y P(d_y|y)P(y|x) \sum_w P(d_w|w)P(w|x) \\ &\backsimeq \sum_y P(y|x)\lambda(y) \sum_w P(w|x)\lambda(w) \end{aligned} \tag{A.48}$$

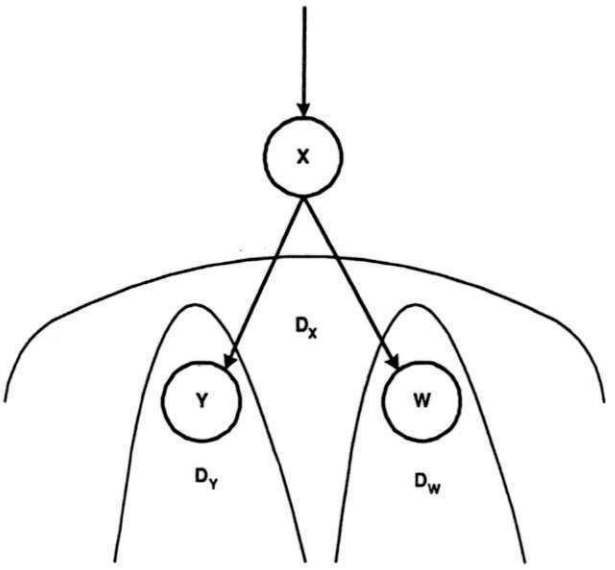


Figura A.2: $D_X = D_Y \cup S_W$

A segunda igualdade é devido a d-separação e a terceira a lei da probabilidade total. Logo, é possível encontrar a proporcionalidade expressa pela Equação (A.46) definindo para todos os valores de x,

$$\lambda_Y(x) \equiv \sum_y P(y|x)\lambda(y)$$

e

$$\lambda_W(x) \equiv \sum_w P(w|x)\lambda(w)$$

e ajustando $\lambda(x) = \lambda_Y(x)\lambda_W(x)$.

2. Desenvolvendo $\pi(x)$:

É necessário mostrar que,

$$\pi(x) \backsimeq P(x|n_x) \tag{A.49}$$

Caso 1: $X \in \mathtt{A}$ e o valor de X é $\widehat{x}$. Devido ao fato de que $X \in N_X$,

$$P(\widehat{x}|n_x) = P(\widehat{x}|\widehat{x}) = 1$$

e

$$P(x|n_x) = P(x|\widehat{x}) = 0,$$

para todo $x \neq \widehat{x}$.

Logo, nós podemos encontrar a proporcionalidade expressa na Equação (A.49) ajustando:

$$\pi(\widehat{x}) = 1$$

$$\pi(x) = 0,$$

para todo $x \neq \widehat{x}$.

Caso 2: $X \notin \mathtt{A}$ e X é raiz. Neste caso, $n_X = \emptyset$, então:

$$P(x|n_x) = P(x|\emptyset) = P(x),$$

para todos os valores de x.

Logo, nós podemos encontrar a proporcionalidade expressa na Equação (A.49) ajustando:

$$\pi(x) \equiv P(x),$$

para todos os valores de x.

Caso 3: $X \notin \mathtt{A}$ e X não é raiz. Sem perda de generalidade assuma que X é o filho da direita de Z e seja T o filho da esquerda de Z. Então, $N_X = N_Z \cup D_T$. Esta situação é esboçada na Figura A.3. Tem-se que:

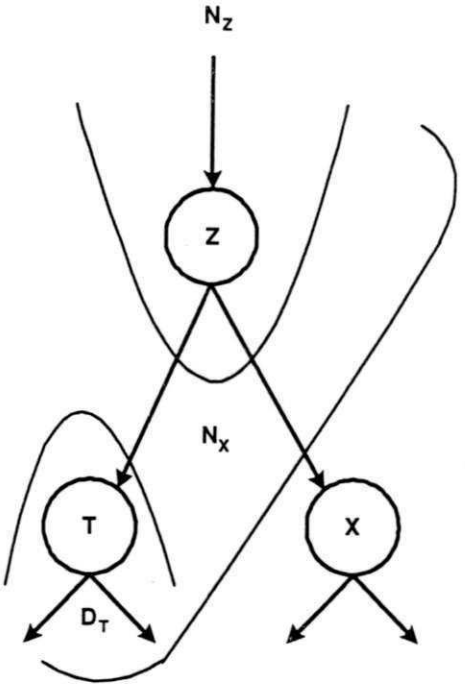


Figura A.3: $N_X = N_Z \cup D_T$

$$\begin{aligned} P(x|n_x) &= \sum_z P(z|n_x)P(x|z) \\ &= \sum_z P(x|n_Z, d_T)P(x|z) \\ &= \sum_z P(x|z)\frac{P(z|n_z)P(n_Z)P(d_T|z)P(z)}{P(z)P(n_Z, d_T)} \\ &= \gamma \sum_z P(x|z)\pi(z)\lambda_T(z) \end{aligned} \tag{A.50}$$

Logo, é possível encontrar a proporcionalidade expressa na Equação (A.49), definindo-se para todos os valores de z,

$$\pi_X(z) = \pi(z)\lambda_T(z),$$

e ajustando:

$$\pi(x) \equiv \sum_{z} P(x|z)\pi_X(z),$$

para todos os valores de x. Isto completa a prova.

■

A.5 Conclusões

Este apêndice foi dedicado a uma revisão matemática, a qual encerra algumas importantes conclusões: apesar dos modelos de inferência clássico e bayesiano diferirem em seus princípios, um não pode ser considerado mais correto que o outro, pois ambos são matematicamente consistentes com suas premissas e com o cálculo de probabilidades. Outra conclusão importante é que o modelo bayesiano sugere um processo de revisão de crença, já que a distribuição a priori é reajustada com base nos dados amostrais. Além disso, o modelo bayesiano não concebe um valor fixo, estável, para o parâmetro estimado e baseia-se na prerrogativa de que este valor pode sofrer variações causai, mapeadas por uma distribuição de probabilidade ais características do modelo bayesiano são atraentes em Inteligência Artificial por poder ser associadas a um processo de aprendizado.

Apêndice B

Base de Casos Clínicos

Seguindo a orientação do Prof. Jovany Medeiros, que é o componente do projeto especialista em Neurologia, foram escolhidos os sintomas e as variações da Miastenia apresentados na Tabela B.1 como ponto de partida para o sistema a ser produzido. A ordem em que os sintomas aparecem na tabela reflete a escolha do especialista e será a ordem dos nós na rede formada sem a avaliação do efeito das ordenações dos nós, segundo as medidas de informação apresentadas no Capítulo 3.

Os casos para montar a base foram em parte fornecidos Prof. Jovany Medeiros e em parte obtidos por buscas na Internet e na literatura especializada. A base de dados com casos da Miastenia é formada até o momento por 74 casos. Vale observar que essa pequena quantidade de casos se deve ao fato que essa doença é rara. A seguir, temos a base formada e utilizada no desenvolvimento do sistema de apoio à emissão do diagnóstico médico.

74

1 1 0 1 1 1 0 2

1 1 0 1 1 1 0 2

1 1 0 4 1 1 1 0

3 2 0 1 1 1 1 1

1 1 0 4 1 1 1 0

3 1 0 1 1 3 0 1

3 1 0 1 1 3 0 2

3 1 0 1 1 1 1 1

1 1 0 3 0 1 0 3

3 1 0 3 0 1 0 3

Nós	Valores
N1 – Ptose palpebral	1. Unilateral Permanente 2. Unilateral Flutuante 3. Bilateral Permanente 4. Bilateral Flutuante 5. Não Observado
N2 – Estrabismo	1. Permanente 2. Flutuante 3. Não Observado
N3 – Diplopia	1. Sim 2. Não
N4 – Fraqueza Muscular	1. Proximal 2. Proximal Flutuante 3. Distal 4. Generalizda 5. Não Observada
N5 – Dificuldade Respiratória	1. Sim 2. Não
N6 – Estimulação Repetitiva	1. Normal 2. Com Diminuição da Amplitude 3. Com Aumento de Amplitude 4. Não Observado
N7 – Envolvimento da Musculatura Bulbar	1. Sim 2. Não
N8 – Miastenia Gravis	1. Ocular 2. Generalizada Leve 3. Generaliza moderada 4. Grave

Tabela B.1: Variáveis e seus valores para a montagem da base de casos da Miastenia Gravis.

3 2 1 3 1 1 0 2
3 1 0 4 1 1 1 0
4 1 0 3 0 1 0 2
3 1 0 1 1 1 1 1
1 1 0 4 1 3 1 0
3 1 0 4 1 3 1 0
1 1 0 1 1 1 0 2
3 1 0 3 1 1 0 2
1 1 0 4 1 0 1 0
3 1 0 1 0 1 0 2
3 1 0 4 1 3 1 0
1 2 1 4 1 3 1 0
3 1 0 4 1 1 1 0
4 2 1 3 0 1 0 3
3 1 0 1 0 1 0 3
1 1 0 4 1 0 1 0
3 1 1 3 1 1 0 1
3 1 0 1 1 1 0 2
3 1 0 1 1 1 0 2
4 1 1 3 0 1 0 2
1 2 1 3 1 1 1 1
2 1 1 4 1 3 1 0
4 0 1 4 1 3 1 0
4 2 1 3 0 0 0 2
4 1 0 1 1 1 1 1
3 2 1 1 1 1 1 1
4 1 0 3 0 1 0 2
4 2 1 1 1 1 1 1
3 2 1 3 0 1 0 3
3 1 0 1 1 1 1 1
3 1 0 1 1 1 0 2
3 1 0 1 1 1 1 1
3 1 0 3 0 1 0 3
3 1 0 1 1 1 1 1

4 1 0 3 1 1 0 2
3 2 1 1 0 1 0 3
3 1 0 1 1 3 1 1
1 1 0 1 0 3 1 1
3 2 1 4 1 1 1 0
3 1 0 1 1 1 1 1
4 1 0 1 1 3 1 1
3 1 0 1 1 1 0 2
3 1 0 1 1 1 0 2
3 1 0 1 1 1 0 1
3 1 0 1 0 1 0 2
4 1 0 4 1 3 1 0
2 0 1 4 1 1 0 1
3 1 0 3 1 3 0 2
3 2 1 4 1 3 0 1
4 2 1 3 0 3 0 3
1 1 0 1 1 3 0 2
4 1 0 4 1 3 1 0
3 1 0 3 1 3 0 2
3 1 0 1 1 3 0 2
3 1 0 1 1 0 1 1
3 1 0 3 0 3 0 3
3 1 0 1 0 1 0 2
3 1 0 3 1 1 1 1
2 1 0 1 1 1 0 2
1 1 0 4 1 1 1 0
3 1 0 4 1 3 1 0
2 2 1 4 1 1 1 0
3 2 0 3 0 1 0 3
3 1 0 1 1 1 0 2

Bibliografia

[1] Air Force Institute of Technology. *Artificial Intelligence Laboratory.* Disponível por WWW em http://www.afit.af.mil/Schools/EN/ENG/LABS/AI/BayesianNetworks/research3.html, 05/janeiro/1997.

[2] Remco R Bouckaert. *Probabilistic Network Construction Using The Minimum Description Length.* Technical Report RUU-CS-94-27, Ultrecht University. Department of Computer Science, Ultrecht University,The Netherlands,1994.

[3] Remco R Bouckaert. *Bayesian Belief Networks: from inference to construction. PhD Thesis.* Faculteit Winskunde en Informatica, Ultrech University, June 1995.

[4] J. Cambier, M. Masson; H. Dehem. *Manual de Neurologia.* Ed.Masson, 2ª edição. 1997.

[5] Cecilia Dias Flores e Charles Lenadro Höher. *Uma Experiência do Uso de Redes Probabilísticas no Diagnóstico Médico, 3er. Simpósio de Informática y Salud.* Argentina, 2001.

[6] Cormen, T. H., Leiserson, C.E. and Rivest, R.L.. *Introduction to Algorithms.* MIT-Press, Cambridge, Massachusetts,1989.

[7] Gregory G. Cooper. *The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks.* Artificial Intelligence.Amsterdam: Elsevier, v.42, p.393-405, 1990.

[8] Gregory G. Cooper and E. Herskovits. *A Bayesian Method for Constructing Bayesian Belief Networks from Databases.* Proceedings of the Conference on Uncertainly in Artificial Intelligence, 86-94, 1991.

[9] Gregory G. Cooper and E. Herskovits. *A Bayesian Method for Induction of Probabilistic Networks from Data.* Machine Learning, 9:309-347, 1992.

[10] C.K. Chow and C.N. Liu. *Approximating Discrete Probability Distributions With Dependence Trees.* IEEE Transactions on Information Theory,14(3):462-467. 1968.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* New York, USA, John Wiley & Sons, Inc.1991.

[12] Paul Dagum and Michael Luby. *Approximating Probabilistic Inference in Bayesian Belief Networks is NPHard.* Artificial Intelligence. Amsterdam: Elsevier, v.60, p.141153, 1993.

[13] Francisco Marcos de Assis. *Caracterização Operacional da Entropia de Tsallis Comparada à Entropia de Rényi. Universidade Federal de Campina Grande.* Campina Grande, Brasil, 2002.

[14] Francisco Marcos de Assis. *Aplicações da Teoria da Informação: Códigos para Transmissão, Compressão e Criptografia. Universidade Federal de Campina Grande.* Campina Grande, Brasil, 2003.

[15] Francisco Marcos de Assis, Ernesto L. Pinto e Jovany L. A. de Medeiros. *Proposta de Sistema Computacional para Auxílio de Diagnóstico de Doenças Neuromusculares. Universidade Federal de Campina Grande.* Campina Grande, Brasil, 2001.

[16] Luiz Gonzaga de Q. Silveira Júnior, Edmar Candeia Gurjão, Ernesto L. Pinto e Francisco Marcos de Assis. *Um Critério para Retirada de Arcos em Redes Bayesianas.* VI Congresso Brasileiro de Redes Neurais, pp 187-192. Centro Universitário da FEI. São Paulo, Brasil.

[17] Luiz Gonzaga de Q. Silveira Júnior, Edmar Candeia Gurjão, Ernesto L. Pinto, Jovany L. Medeiros e Francisco Marcos de Assis. *Critérios para Remoção de Arcos em Redes Bayesianas.* VI Simpósio Brasileiro de Automação Inteligente, pp . Bauru-SP, Brasil, Setembro de 2003.

[18] Luiz Gonzaga de Q. Silveira Júnior, Edmar Candeia Gurjão, Ernesto L. Pinto, Jovany L. Medeiros e Francisco Marcos de Assis. *Critérios para Remoção de Arcos em Redes Bayesianas.* Learning and Nonlinear Models-Revista da Sociedade Brasileira de Redes Neurais, 2003.

[19] R. A. Engelen. *Approximating Bayesian Belief Network by Arc Removal.* IEEE Trans. on Pattern Recgnition and Machine Intelligence, vol. 19, no. 8, pp. 916–920, August 1997.

[20] E. Charniak. *Bayesian Networks Without Tears.* AI Magazine, 12(4):50-63,1991.

[21] Randall Davis; Bruce Buchanan; E. Shortliffe. *Production Rules as a Representation for a KnowledgeBased Consultation Program.* Artificial Intelligence. Amsterdam: NorthHolland, v.8, 1977, p.1545.

[22] Max Henrion; John S. Breese; Eric J. Horvitz. *Decision Analysis and Expert Systems.* AI Magazine. AAAI Press: Winter 1991, p.6491.

[23] David Heckerman ; HORVITZ, E. and NATHWANI, B.. *Towards normative experts systems: Part I. the Pathfinder project..* In Methods of Information in Medicine, 31, p.90-105, 1992.

[24] David Heckerman. *A Tutorial on Learning Bayesian Networks.* In Technical Report MSR-TR–95-06, Advanced Technology Division, Microsoft Corporation, 1995.

[25] E. Herskovits. *Computer-Based Probabilistic-Network Construction.* In PhD Thesis. Medical Informatics, Stanford University,1991.

[26] E. Herskovits. *A Bayesian Approach to Learning Causal.* In Technical Report MSR-TR–95-04, Microsoft Research, March, 1995.

[27] Joffre M. de Rezende. *Caminhos da Medicina: Uso da Tecnologia no Diagnóstico Médico e suas Conseqüências.* In Anais do XIV Encontro Científico do Acadêmicos de Medicina. Goiânia, 2002.

[28] Finn V. Jensen and Stig K. Andersen. *Use of Causal Probabilistic Networks as High Level Models in Computer Vision.* Technical Report R-90-39, Univ. of Aalborg, Denmark, 1990

[29] Finn V. Jensen, Kristian G. Olsen, Stig K. Andersen. *An Algebra of Bayesian Belief Universes for KnowledgeBased Systems Networks.* New York: John Wiley & Sons, Inc., v.20, p.637659, 1990.

[30] Finn V. Jensen, Steffen L. Lauritzen, Kristian G. Olsen. *Bayesian Updating in Causal Probabilistic Networks by Local Computations.* Computational Statistics Quarterly. Heidelberg: PhysicaVerlag v.4, p.269282, 1990.

[31] Finn V. Jensen, Frank Jensen. *Optimal Junction Trees.* 10th CONF. ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE (UAI 94). Proc. ... San Francisco: Morgan Kaufmann, p.360366, 1994.

[32] Finn V. Jensen, Frank Jensen and Søren L. Dittmer. *From Influence Diagrams to Junction Trees.* 10th CONF. ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE (UAI 94). Proc.... San Francisco: Morgan Kaufmann, p.367373, 1994.

[33] Finn V. Jensen. *Bayesian Networks Basics.* van der GAAG, Linda C. (ed.) AISB Quarterly, v.94, Winter 1995/Springer 1996, p.922 (Newsletter of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, Special Theme: Bayesian Belief Networks).

[34] Finn V. Jensen. *An Introduction to Bayesian Networks.* London: UCL Press, 1996.

[35] Cristian Koehler e Sílvia Modesto Nassar. *Modelagem de Redes Bayesianas a partir de Base de Dados Médicas, 3er. Simpósio de Informática y Salud.* Argentina, 2001.

[36] S. Kullback and R. A. Leibler. *On Information and Sufficiency.* Annals of Mathematical Statistics,22:76-86,1951.

[37] Leonardo Nogueira Matos. *Projeto de Pesquisa II: Estudo de Redes Bayesianas e sua Aplicação em Reconhecimento de Padrões.* Universidade Federal de Campina Grande. Junho de 2002.

[38] Steffen L. Lauritzen and David J. Spiegelhalter. *Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion).* Journal Royal Statistical Society, Series B, v.50, n.2, p.425448, 1988.

[39] Judea Pearl. *Fusion, Propagation, and Structuring in Belief Networks.* Artificial Intelligence. North Holland, v.29, p.241288, 1986.

[40] Judea Pearl. *A Constraint Propagation Approach to Probabilistic Reasoning.* KANAL, L.N.; LEMMER, J.F. (Eds.). Uncertainty in Artificial Intelligence. Amsterdam: North Holland, p.357370, 1986.

[41] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference.* San Mateo: Morgan Kaufmann, 1988; revised second printing, 1991.

[42] Judea Pearl. *Belief Networks Revisited.* Artificial Intelligence.Amsterdam: Elsevier, v.59, p.4956, 1993.

[43] G. Rebane and Judea Pearl. *The Recovery of Causal Polytrees from Statistical Data.* Proceedings of the Conference on Uncertainly in Artificial Intelligence, 222-228, 1987.

[44] Richard E. Neapolitan. *Learning Bayesian Networks.* Prentice Hall Series in Artificial Intelligence, 2003

[45] R. D. Robinson. *Counting Unlabeled Acyclic Digraphs.* Proceedings of the Fifth Australian Conference on Combinatorial Mathematics, 28-43, 1976.

[46] L.P. Rowland. *Tratado de Neurologia de Merrit.* Ed. Guanabara Koogan, 9a. edição. Rio de Janeiro, 1997.

[47] Jorma Rissanen. *Modeling by Shortest Data Description.* Automatica, 14:465-471, 1978.

[48] R.E. Tarjan. *Depth-first search and linear graph algorithms.* SIAM Journal on Computing, 1(2):146– 160, 1972.

[49] S.M. Aji and R. J. McEliece. *The Generalized Distributive Law.* IEEE Transactions On Information Theory, 325-343 ,March, 2000.

[50] Wai Lam and Fahiem Bacchus. *Learning Bayesian Belief Networks An Approach Based on the MDL Principle.* Computational Intelligence, 10:4, 1994.

[51] S.L. Zabell. *W. E. Johnson's Sufficientness Postulate.* Annals of Statistics, 10(4):1091-1099, 1982.