Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação
Departamento de Computação e Automação

Aprendizado em Redes Bayesianas

George Barreto Pereira Bezerra
Pablo Alberto Dalbem de Castro

{bezerra,pablo}@dca.fee.unicamp.br

Campinas
Maio de 2005

Sumário

Capítulo 1

Introdução

As Redes Bayesianas (RBs) são poderosas ferramentas para raciocínio e representação de conhecimento frente a incertezas. Além disso, ela provê uma representação natural para relacionamentos "causa-efeito" entre as variáveis. As RBs estão fundamentadas no teorema de Bayes, que é um método quantitativo para revisão de probabilidades, conforme a chegada de novas informações [Pearl, 1988]. Formalmente, uma RB é um grafo acíclico direcionado em que os nós são as variáveis do problema, os arcos entre os nós indicam uma dependência entre elas e cada variável está associada a uma tabela de probabilidades.

A motivação para usar redes bayesianas é que para descrever um modelo do mundo real não é necessário usar uma enorme tabela de probabilidades conjuntas na qual são listadas as probabilidades de todas as combinações possíveis de eventos. A maioria dos eventos é condicionalmente independente dos outros, portanto suas interações não precisam ser consideradas. Em vez disso, usa-se uma representação mais local, que descreve agrupamentos de

eventos que interagem entre si [Heckerman, 1995].

As redes bayesianas estão sendo utilizadas com sucesso em aplicações, principalmente, que requerem inferência probabilística e diagnósticos. Além disso, sua aplicação em situações em que deseja-se conhecer e modelar a relação causal entre as variáveis do problema, tem se mostrado um linha de pesquisa promissora. Um exemplo mais imediato é na área de bioinformática, em que dadas as expressões gênicas, deseja-se saber qual gene regula qual.

Para se construir uma rede bayesiana, um especialista no domínio do problema deve fornecer a estrutura da rede, bem como as tabelas de probabilidades. Entretanto, nem sempre esse especialista está disponível. Uma alternativa para tal inconveniente é empregar técnicas de aprendizado de RBs, os quais extraem o conhecimento necessário a partir de conjuntos de dados que representam amostras ou exemplos do problema.

O aprendizado de redes bayesianas consiste em encontrar a rede bayesiana que melhor representa o conjunto de dados. Encontrar uma rede bayesiana significa determinar a sua estrutura, bem como as tabelas de probabilidades para cada variável. A melhor rede é aquela que possui maior verossimilhança com os dados.

O aprendizado de redes bayesianas é uma tarefa muitas vezes complexa. Muitas metodologias têm sido propostas para resolvê-la. Este documento apresenta os algoritmos mais conhecidos e utilizados na literatura para realização de tal tarefa.

Este texto está organizado da seguinte forma. No Capítulo 2, alguns conceitos necessários da teoria da probabilidade são apresentados. No Capítulo 3, a definição de Redes Bayesianas, bem com suas aplicações

são apresentadas. O aprendizado de redes bayesianas e os algoritmos mais utilizados são descritos no Capítulo 4. Por fim, no Capítulo 5 são apresentadas algumas considerações finais e direções futuras.

Capítulo 2

Fundamentos da Teoria da Probabilidade

O objetivo deste capítulo é apresentar alguns conceitos básicos da teoria da probabilidade. Tais conceitos são importantes para o entendimento de alguns pontos deste trabalho. As seguintes definições foram baseadas nos livros [Papoulis & Pillai, 2002] e [Leon-Garcia, 1994].

2.1 Espaço Amostral e Evento

Um experimento aleatório, denotado por E, é um experimento em que o resultado não pode ser predito, mesmo que ele seja repetido várias vezes nas mesmas condições. Abaixo estão 3 experimentos aleatórios que serão utilizados no decorrer deste capítulo.

E_1: Jogar uma moeda 3 vezes e observar o número de vezes que aparece "cara" e "coroa".

E_2: Jogar um dado e observar o número na face de cima.

E_3: Uma urna contem bolas numeradas de 1 a 100. Selecionar uma bola e observar seu número.

Uma vez que não é sempre que o mesmo resultado aparece nas realizações de um experimento, é necessário estabelecer o conjunto de todos os possíveis resultados para aquele experimento. Este conjunto é chamado de **espaço amostral**, denotado aqui por S. Se os elementos do espaço amostral são finitos e contáveis, o espaço amostral é dito ser discreto. Caso contrário, ele é contínuo. Os espaços amostrais referentes aos experimentos E_1, E_2 e E_3 são todos discretos e apresentados a seguir:

S_1={1,2,3,4,5,6}

S_2={HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}

S_3={1,2,3,...,98,99,100}

Algumas vezes o interesse não é por algum resultado específico, mas sim pela ocorrência de eventos, isto é, resultados que satisfazem certas condições. Assim, um **evento** é um subconjunto dos possíveis resultados de um experimento. Por exemplo, no experimento de jogar um dado e observar sua face de cima, pode ser de interesse o evento "número menor que 4". As condições de interesse definem um subconjunto do espaço amostral. Assim, o evento "número menor que 4" ocorre se e somente se o resultado do dado for o número 1, 2 ou 3.

Os seguintes exemplos são eventos para os experimentos E_1, E_2 e E_3.

Evento A: "Aparecer um número par". O conjunto de resultados

correspondente a esse evento é A={2,4,6}.

Evento B: "Aparecer duas caras e uma coroa em qualquer ordem". O conjunto de resultados correspondente a esse evento é B={(HHT),(HTH),(THH)}

Evento C: "Aparecer uma bola com número maior que 80". O conjunto de resultados correspondente a esse evento é C={81,82,...,99,100}

Dois eventos de especial interesse são o **evento certo**, S, que consiste de todos os resultados e por isso ocorre sempre, e o **evento nulo**, $\emptyset$, que nunca ocorre, ou seja, os resultados do experimento não satisfazem as condições especificadas.

Dois eventos, A e B, são ditos **mutuamente exclusivos** se eles não possuem elementos em comum. Assim, AB=$\emptyset$. A **partição U** do espaço amostral S é uma coleção de eventos mutuamente exclusivos A_i de S, tal que $A_1 \cup A_2 \cup ... \cup A_n = \text{S}$ e $A_i A_j = \{\emptyset\}$, $i \neq j$.

Para cada evento E é atribuído um número real no intervalo [0,1], *P(E)*, chamado de probabilidade de ocorrer o evento E, o qual satisfaz os seguintes axiomas:

- **Axioma 1:** $P(A) \geq 0$
- **Axioma 2:** $P(S) = 1$
- **Axioma 3:** Se $\text{A} \cap \text{B} = \emptyset$, $P(A \cup B) = P(A) + P(B)$

Se S consiste de N resultados equiprováveis e E é um evento consistindo de r elementos, então a probabilidade do evento E é dada por (2.1).

$$\mathrm{P(E)} = \frac{r}{N} \tag{2.1}$$

Exemplo 1: Uma urna contem 10 bolas numeradas de 0 a 9. Num experimento aleatório é preciso selecionar uma bola da urna e anotar seu número. Encontrar a probabilidade dos eventos:

A = "número da bola é ímpar"

B = "número da bola é múltiplo de 3"

O espaço amostral é S={0,1,...,9} e os resultados correspondentes aos eventos acima são:

$$\mathrm{A}=\{1,3,5,7,9\} \qquad \mathrm{B}=\{3,6,9\}$$

Se for suposto que os resultados são equiprováveis, então:

P(A)=P(1)+P(3)+P(5)+P(7)+P(9)

$= 1/10 + 1/10 + 1/10 + 1/10 + 1/10 = 5/10.$

P(B)=P(3)+P(6)+P(9)

$= 1/10 + 1/10 + 1/10 = 3/10.$

2.2 Probabilidade Conjunta

A **probabilidade conjunta** fornece a probabilidade de dois ou mais eventos ocorrerem simultaneamente. Para isso, é necessário estabelecer a intersecção dos espaços amostrais dos eventos. Considere novamente o Exemplo 1 da seção anterior e os seguintes eventos com seus respectivos espaços amostrais:

A = "número da bola é ímpar" = {1,3,5,7,9}

B = “número da bola é múltiplo de 3” = {3,6,9}

A intersecção destes espaços amostrais é $A \cap B$ ={3,9} e a probabilidade conjunta de A e B, denotado por P(AB), é definida como

$$P(AB) = P(A \cap B) = P(3) + P(9) = 1/10 + 1/10 = 2/10$$

Portanto, ao retirar uma bola da urna a probabilidade de ocorrerem simultaneamente os eventos “número da bola é ímpar” e “número da bola é múltiplo de 3” é 2/10.

2.3 Probabilidade Condicional

Em alguns casos é interessante determinar quando dois eventos estão relacionados, no sentido de que a ocorrência de um evento altera a probabilidade de ocorrência do outro. Isto requer que seja encontrada a **probabilidade condicional** destes dois eventos. A probabilidade condicional de um evento A, dado que ocorreu outro evento M, denotado por *P(A|M)*, é por definição:

$$\mathrm{P}(A|M) = \frac{P(AM)}{P(M)} \tag{2.2}$$

em que *P(M)>0*.

Exemplo 2: Ao jogar um dado, qual a probabilidade de aparecer o número 2 na face superior, dado que ocorreu o evento “par”?

S={1,2,3,4,5,6} A={2} M={par}={2,4,6}

Sabe-se que *P(A)*=1/6, *P(M)*=3/6 e *P(AM)*=1/6. Pela equação (2.2)

tem-se:

$$P(A|M) = \frac{P(AM)}{P(M)} = \frac{1/6}{3/6} = 1/3$$

Dado que ocorreu o evento "par", a probabilidade de ocorrer o evento "aparecer número 2" é 1/3.

2.4 Teorema de Bayes

Probabilidade é capaz de prover um mecanismo para revisar coerentemente a probabilidade de eventos conforme algumas evidências acontecem. Qual a probabilidade de ocorrer o evento A, sabendo que ocorreu o evento B? Existe alguma dependência entre esses eventos? A probabilidade condicional e teorema de Bayes exercem um papel fundamental para responder perguntas nesse contexto. Uma breve introdução desse teorema e dois exemplos são apresentados a seguir.

Para se chegar ao teorema de Bayes, primeiro é necessário recordar a equação de probabilidade condicional, dada por:

$$\mathrm{P}(A|B) = \frac{P(AB)}{P(B)} \tag{2.3}$$

Veja que pela equação (2.3), pode-se escrever a probabilidade conjunta dos eventos A e B como sendo:

$$\mathrm{P(AB)} = \mathrm{P}(A|B)\mathrm{P(B)} = \mathrm{P}(B|A)\mathrm{P(A)} \tag{2.4}$$

A equação (2.4) é conhecida como **teorema da probabilidade total**.

Substituindo a equação (2.4) em (2.3), obtemos

$$\mathrm{P}(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.5}$$

A equação (2.5) é conhecida como **Teorema de Bayes**.

Cada termo na equação (2.5) possui um nome convencional. O termo $\mathrm{P}(A_i)$ é chamado de probabilidade a priori de A_i. O termo $\mathrm{P}(A_i|B)$ é a probabilidade a posteriori de A_i, dado B. O termo $(B|A_i)$ é chamado likelihood de A dado B. O termo P(B) é a probabilidade a priori de B que serve como um fator de normalização, devendo ser positivo. Com esta terminologia, o teorema de Bayes pode ser reescrito como

$$\text{posterior} = \frac{\text{priori * likelihood}}{\text{fator de normalização}}$$

Exemplo 3: Um médico sabe que meningite causa dor no pescoço em 50% dos casos. Ele sabe também alguns fatos incondicionais, como a probabilidade a priori de um paciente ter meningite (M) é 1/50000 e a probabilidade a priori de qualquer paciente ter uma dor no pescoço (S) é 1/20.

Tem-se que:

$\mathrm{P}(S|M) = 1/2$

P(M) = 1/50000

P(S) = 1/20

Um paciente chega ao consultório do médico com dor no pescoço. Qual a probabilidade dele estar com miningite? O que se quer encontrar é $\mathrm{P}(M|S)$.

$$P(M|S) = \frac{P(S|M) * P(M)}{P(S)} = \frac{1/2 * 1/500000}{1/20} = 0.0002$$

Logo, a probabilidade do paciente com dor no pescoço ter meningite é 0.0002.

Exemplo 4: Em uma certa cidade, 30% das pessoas são Conservadoras (Con), 20% Independentes (Ind)e 50% são Liberais (Lib). Registros mostram que na última eleição, 65% dos Conservadores votaram, 82% dos Liberais votaram e 50% dos Independentes votaram. Se uma pessoa é escolhida aleatoriamente na rua e ela diz que não votou na última eleição (NV), qual a probabilidade dessa pessoa ser Liberal?

Deseja-se encontrar a probabilidade

$$P(Lib|NV) = \frac{P(NV|Lib) * P(Lib)}{P(NV)}$$

Tem-se que:

P(Con)= 0.30 P(Ind)=0.20 P(Lib)=0.50

P(NV)=P(Con)*P($Con|NV$)+P(Ind)*P($Ind|NV$)+P(Lib)*P($Lib|NV$)

= 0.30*0.35 + 0.50*0.18 + 0.20*0.50 = 0.295

Logo,

$$P(Lib|NV) = \frac{P(NV|Lib) * P(Lib)}{P(NV)} = \frac{0.18 * 0.50}{0.295} = 0.30$$

Desta forma, a probabilidade dessa pessoa ser Liberal é 30%.

Capítulo 3

Redes Bayesianas

3.1 Definição

Redes Bayesianas (RBs) são poderosos modelos gráficos para representação e raciocínio perante incertezas. Elas representam explicitamente o relacionamento probabilístico entre um conjunto de variáveis de interesse e fornece uma especificação concisa da probabilidade conjunta. As RBs estão fundamentadas no Teorema de Bayes, apresentado no capítulo anterior, para atualizar probabilidades conhecidas conforme a chegada de nova informação.

Formalmente, uma RB para um conjunto $X = \{x_1, x_2, ..., x_n\}$ de variáveis consiste de uma estrutura em rede **S** que representa as dependências condicionais das variáveis em X e de um conjunto **P** de probabilidades para cada variável [Pearl, 1988] [Jensen, 1996] [Cowell et al., 1999] [Jensen, 2001]. Juntos, estes dois componentes definem a probabilidade conjunta de X. A natureza das variáveis pode ser discreta ou contínua.

Geralmente, a estrutura de rede S é um grafo direcionado acíclico, mas poderia ser também uma estrutura em árvore. Independente da forma de representação, em tal estrutura:

- cada nó no grafo corresponde a uma variável em X
- os arcos ligando os nós indicam as relações de dependência entre as variáveis. Um arco saindo de uma variável X_i e chegando numa variável X_j significa que X_i é pai de X_j.
- cada variável possui uma tabela de probabilidade condicional, a qual descreve as probabilidades daquela variável, dados os diferentes valores que seus pais podem assumir.

Na Figura 3.1 é apresentado um exemplo de Rede Bayesiana que ilustra os conceitos definidos previamente.

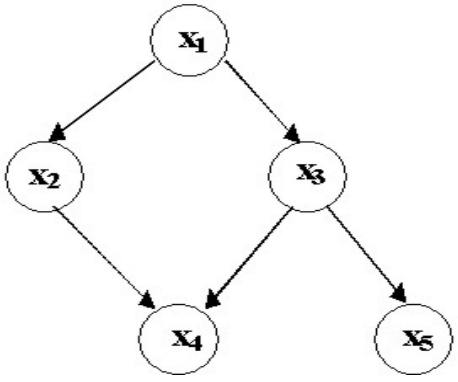

Figura 3.1: Exemplo de Rede Bayesiana

O conjunto de variáveis $X = \{X_1, X_2, X_3, X_4, X_5\}$ retrata as variáveis do modelo e são representadas pelos nós do grafo. As probabilidades para cada variável são apresentadas na Tabela 3.1.

Tabela 3.1: Probabilidades para cada variável

Variavel	Pais	Probabilidade
X1=1	–	0.5
X2=1	X1=0	0.8
	X1=1	0.4
X3=1	X1=0	0.6
	X1=1	0.1
X4=1	X2=0	0.5
	X2=1	0.7
	X3=0	0.3
	X3=1	0.4
X5=1	X3=0	0.9
	X3=1	0.6

Além da explícita dependência condicional entre as variáveis, as RB representam implicitamente as independências condicionais entre elas. Uma variável é dita ser condicionalmente independente de outras que não são suas filhas, dados os seus pais. Na rede da Figura 3.1, a variável X_5 é condicionalmente independente de X_1, X_2 e X_4, dado seu pai X_3. Por exemplo, $P(X_5|X_1, X_2, X_3, X_4)$ é o mesmo que $P(X_5|X_3)$.

Desta forma, as RBs são capazes de expressar a probabilidade conjunta das variáveis de uma forma mais compacta. Assim, a distribuição conjunta das variáveis do modelo da Figura 3.1 pode ser expressa por

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_4|X_2X_3)P(X_2|X_1)P(X_3|X_1)P(X_5|X_3)P(X_1)$$

De uma forma mais geral, a distribuição conjunta de X é dada por

$$P(X) = \prod_{i=1}^{n} P(x_i|\pi_{X_i})$$

em que x_i é uma instância de X, π_{X_i} é o conjunto de pais de x_i e $P(X_i|\pi_{X_i})$ é a probabilidade condicional de X_i dados os valores das variáveis π_{X_i}.

3.2 Porque usar Redes Bayesianas

As redes bayesianas são apropriadas para raciocínio e representação do conhecimento frente a problemas que apresentam alguma forma de incerteza [Ramoni & Sebastiani, 1999].

Conforme apresentado e discutido em [Heckerman & Chickering, 1996] e [Heckerman et al., 1995a], dentre os motivos para se usar RB tem-se que (a) uma RB permite expressar as assertivas de independência de forma visual e fácil de perceber; (b) uma RB representa e armazena uma distribuição conjunta de forma econômica; (c) uma RB torna o processo de inferência eficiente computacionalmente; (d) RBs permitem conhecer o relacionamento causal entre as variáveis do problema; (e) as RB podem trabalhar com conjuntos de dados incompletos (com valores ausentes) e (f) as RB oferecem uma eficiente abordagem para evitar sobre-ajuste dos dados.

Conforme o tipo de aplicação, a utilização prática de uma RB pode ser considerada da mesma maneira que modelos como redes neurais, sistemas especialistas, árvores de decisão, modelos para análise de dados (regressão linear), modelos lógicos, etc. Naturalmente, na escolha do método depende de diferentes critérios, como a facilidade, o custo e a demora na implantação de uma solução. Segundo [Naim et al., 2004], os seguintes aspectos das RBs as fazem preferíveis sobre outros modelos:

- Aquisição de conhecimentos

A possibilidade de juntar conhecimentos de naturezas diversas num mesmo modelo: dados históricos ou empíricos, experiência (expressa na forma de regras lógicas, de equações, de estatísticas ou de probabilidades subjetivas), observações.

- Representação de conhecimentos

A representação gráfica de uma RB é explícita, intuitiva e compreensível para uma pessoa não especialista, o que por sua vez, facilita a validação do modelo, suas evoluções eventuais e sobretudo a sua utilização. Tipicamente, um decisor é mais confiante sobre um modelo no qual ele compreende o funcionamento do que num modelo tipo "caixa preta".

- Utilização de conhecimentos

Uma RB é multi funcional. Pode-se usar o mesmo modelo para avaliar, prever, diagnosticar, ou otimizar as decisões, o que compensa o esforço gasto na construção da RB.

3.3 Aplicações de Redes Bayesianas

Redes Bayesianas estão cada vez mais sendo utilizadas em problemas do mundo real. Elas são úteis em tomada de decisões, para controlar ou prever o comportamento de um sistema, diagnosticar as causas de um fenômeno, etc.

A seguir alguns poucos exemplos de aplicações em que RBs são utilizadas com sucesso [Heckerman et al., 1995b] [Sykacek et al., 1998] [Charniak, 1991].

- Mineração de Dados (*Data Mining*)

Em mineração de dados, as RB são empregadas para descobrir alguma relação entre as variáveis do problema. Recentemente, problemas de bioinformática são os mais visados. Em [Friedman et al., 2000] é apresentado um método bayesiano para análise de expressões gênicas e descobrir as relações entre os genes. Modelagem de redes gênicas usando abordagem bayesiana pode ser encontrada em [Imoto et al., 2003] e [Nariai et al., 2004]. Descobrir o relacionamento entre aminoácidos é a tarefa de uma RB em [Klinger & Brutlag, 1994].

A área de aplicação não se resume a bioinformática. Redes Bayesianas para descobrir relacionamentos meteorológicos são usadas em [Cano et al., 2004]. Mineração de dados com redes bayesianas na área de *marketing* é descrita em [Myllymäki et al., 2001].

Mais trabalhos sobre mineração de dados com abordagens bayesianas são encontrados em [Heckerman et al., 1995a] e [Arnborg, 1999].

- Classificação, Predição e Controle

Uma visão geral de classificação usando abordagem bayesiana é apresentada em [Friedman et al., 1997]. Em [Raval et al., 2002] é apresentado uma abordagem que emprega RB para classificação de proteínas. Em [Ducksbury, 1993] foi usada uma RB para classificar regiões de determinadas cidades.

Reconhecimento de voz por RBs é descrito em [Zweig & Russell, 1998] e [Stephenson et al., 2000]. Já aplicações em visão computacional são encontradas em [Rehg et al., 1999] [Vehtari & Lampinen, 2000]

Em [Habrant, 1999] e [Abramson, 1994] é utilizada RBs para predição de séries temporais financeiras.

Controle de robôs e de ar condicionado são exemplos também de aplicações em [Dean, 1990] e [Nakajima et al., 1998], respectivamente.

Outros trabalhos que empregam redes bayesianas para classificação, predição e controle podem ser encontrados em [Ezawa & Schuermann, 1995] e [Vehtari & Lampinen, 1999].

- Diagnóstico

Diagnóstico médico é o campo mais explorado por redes bayesianas. Trabalhos para diagnosticar câncer pode ser encontrado em [Antal et al., 2003] e para diagnóstico em exames de mamografia são apresentados em [Kahn et al., 1995].

Diagnóstico de falhas em computadores é uma área que está crescendo. Trabalhos nessa linha são descritos em [Breese & Blake, 1995], [Horvitz et al., 2001] e [Shortliffe, 1976]. Ainda na área de computação, as RBs estão sendo empregadas para filtrar *spam* [Androutsopoulos et al., 2000] [Sahami et al., 1998].

Abordagens para outros tipos de diagnóstico podem ser encontrados em [Peng & Reggia, 1987].

- Aprendizado de Máquina

Redes Bayesinas estão sendo empregadas também para em Inteligência Artificial na área de aprendizado de máquina para modelar redes neurais [Lampinen & Vehtari, 2000] [MacKay, 1996] [Neal, 1996] ou sistemas baseados em regras [Andersen, 1989] [Duda et al., 1976]. Pode-se citar ainda aplicações de RB em seleção de atributos [Inza et al., 2000] e mistura de classificadores [Bahle & Navarro, 2000].

Capítulo 4

Aprendizado em Redes Bayesianas

A construção de uma Rede Bayesiana é feita por um especialista no domínio do problema que se está tentando modelar. Entretanto, dependendo do problema, esse processo pode ser difícil e demorado ou o especialista pode não estar disponível. Dessa forma, nos últimos anos têm sido utilizadas técnicas automáticas que aprendem RB a partir de dados numéricos que representam amostras ou exemplos do problema.

A *aprendizagem de RBs* consiste em induzir, a partir de uma amostra de dados, as distribuições de probabilidades simples e condicionais e/ou identificar as relações de interdependência entre as variáveis. Esse processo de aprendizagem indutiva pode ser de dois tipos. A primeira é quando a estrutura da rede é conhecida e então a tarefa se resume em aprender apenas as tabelas de probabilidades para cada variável. A segunda situação é quando a estrutura é desconhecida e, portanto, além das probabilidades, é necessário

aprender a estrutura também.

Nas próximas seções estes dois cenários serão explorados em detalhes.

4.1 Estrutura Conhecida

Este é o caso mais simples, estudado e compreendido da literatura de RBs [Heckerman, 1995]. A estrutura da RB é especificada, e só é necessário estimar os parâmetros numéricos (distribuição de probabilidade simples e conjunta). O problema é bem definido e os algoritmos computacionalmente eficientes.

As tabelas de probabilidade são geradas, visando maximizar o *likelihood*, isto é, a verossimilhança entre o modelo e os dados.

A aprendizagem deste tipo é realizada calculando estimativas de máxima verossimilhança para as entradas nas tabelas de probabilidade das variáveis. Estimativas de máxima verossimilhança não consideram conhecimento a priori sobre as distribuições de probabilidade, utilizam somente os dados disponíveis.

Trabalhos para aprendizado das tabelas de probabilidade para redes bayesianas são descritos em [Greiner et al., 1997].

4.2 Estrutura Desconhecida

Neste cenário, a partir de dados disponíveis deseja-se construir uma estrutura para a RB cuja distribuição conjunta melhor represente a verdadeira distribuição subjacente aos dados.

Uma idéia simples deste tipo de aprendizagem seria, baseando-se nos dados, determinar relações entre as variáveis, adicionar ou excluir arcos, estabelecer direções, enfim definir um grafo para o qual serão calculadas distribuições de probabilidades. Este é um problema NP-complexo [Buntine, 1996]. Em geral, há dois paradigmas principais de aprendizagem de redes bayesianas a partir de dados [Cheng et al., 1997]: **paradigma de busca e pontuação** e o **método de independência condicional**.

4.2.1 Método de Busca e Pontuação

No paradigma de busca e pontuação, a aprendizagem se realiza buscando uma estrutura que seja aderente aos dados, isto é, a estrutura que melhor representa o conjunto de dados. A busca por estruturas é realizada por heurísticas, dado que varrer todo o espaço de busca é um problema NP-complexo. Em geral se inicia com um grafo sem arcos, então, usa-se algum método de busca gulosa que adicione um arco ao grafo. O próximo passo consiste em usar uma função de pontuação para determinar se a nova estrutura é melhor que a anterior. Se for melhor, o novo arco adicionado é mantido e tenta-se adicionar outro. Este processo continua até que nenhuma nova estrutura seja melhor que as anteriores.

Na Figura 4.1 está ilustrado o esquema desta busca para um problema com 4 variáveis. O algoritmo começa sem ligação entre os nós (a) e uma avaliação desta rede não conectada é feita. Em seguida, adiciona-se um arco e a avalia novamente (b). Se a nova estrutura for melhor, mantenha o arco. Caso contrário, retire-o. Este processo continua até que nenhuma estrutura

seja melhor que a anterior. No exemplo, a rede (e) é retornada como solução.

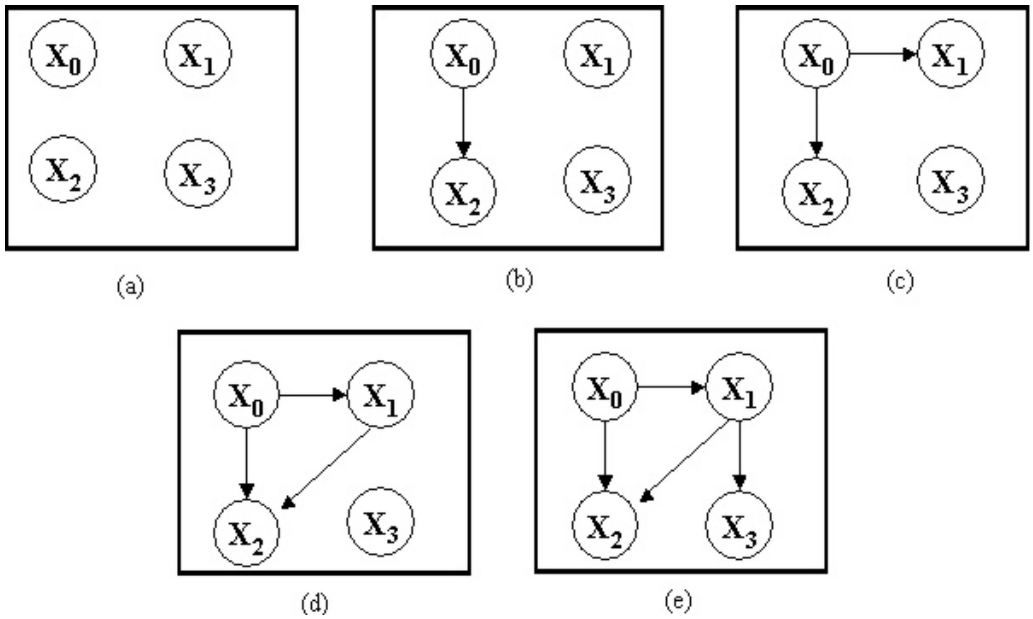

Figura 4.1: Esquema dos métodos de busca e pontuação

Diferentes algoritmos tem sido propostos para geração de redes bayesianas seguindo o raciocínio previamente apresentado. O que geralmente os diferem é a função de pontuação, a estrutura da rede e o algoritmo de busca empregado.

Dentre os métodos de pontuação mais utilizados estão o método de pontuação Bayesiana [Buntine, 1994] [Cooper & Herskovits, 1992] [Heckerman et al., 1995a], o método baseado na entropia [Acid & Campos, 1996a] e o método de comprimento mínimo de descrição (MDL, do inglês *Minimum Description Length*) [Grünwald, 2000] [Lam & Bacchus, 1994] [Suzuki, 1996]. Estas funções de pontuação avaliam cada rede com base na verossimilhança do modelo com os dados, junto com alguma possível restrição de complexidade do modelo.

Outros métodos clássicos para seleção de modelos também podem ser empregados para avaliação das redes bayesianas, como o **BIC** (Bayesian Information Criterion) [Schwarz, 1978] e **AIC** (Akaike Information Criterion) [Akaike, 1974]. A seguir, algumas propostas de algoritmos referentes às principais classes de pontuação serão apresentadas.

Pontuação baseada na entropia

O método **Chow-Liu**, apresentado em [Chow & Liu, 1968], foi um dos primeiros trabalhos sobre aprendizado de redes bayesianas. O algoritmo trabalha com uma estrutura de árvore para k variáveis. A idéia básica é estimar uma distribuição de probabilidade conjunta e procurar a árvore que possui uma distribuição de probabilidade mais próxima da estimada. Um grafo não direcionado é formado iniciando com um grafo sem arcos e adicionando um arco entre dois nós com máxima entropia. Depois, um arco com máxima entropia associada é adicionado, desde que não crie um ciclo no grafo. Este processo é repetido até que não seja possível adicionar arcos. 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 muito restrita a uma pequena área.

O algoritmo **Rebane-Pearl**, descrito em [Rebane & Pearl, 1987], é uma extensão direta do algoritmo de Chow-Liu. Este algoritmo é aplicado no caso em que a representação gráfica das probabilidades é em forma de poliárvore. Uma poliárvore, 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. Usando o método de Chow-Liu é gerada a estrutura básica da árvore e, em seguida, aplicando-se nesta estrutura o algoritmo de poliárvores, obtem-se a representação gráfica da distribuição. Os autores também desenvolveram um método para encontrar a direcionalidade dos arcos no grafo. A idéia subjacente a este método é usada por muitos algoritmos que são capazes de orientar arcos [Cheng et al., 1997].

O algoritmo **Kutató**, apresentado em [Cooper & Herskovits, 1992] utiliza medida de entropia para encontrar a melhor RB. O algoritmo Kutató aplica um algoritmo de busca que utiliza a técnica "gulosa" entre estruturas de rede, selecionando aquela com a menor entropia associada, a qual representa a distribuição de probabilidades mais expressiva. Neste algoritmo é necessário uma ordenar as variáveis em uma lista, de modo que os pais de uma determinada variável apareçam antes dela na lista. Este procedimento requer que o usuário tenha uma noção de quais variáveis podem influenciar algumas outras. Os autores adotaram o problema de aprendizagem como uma aproximação de função de probabilidade conjunta "verdade" dos dados usando uma estrutura de Rede Bayesiana que tem mínima perda de informação, isto significa entropia máxima.

Mais algoritmos que utilizam medida de entropia para geração de redes bayesianas são encontrados em [Geiger, 1992].

Pontuação MDL

Este método é bastante utilizado por engenheiros e cientistas da computação, pois são pessoas que possuem um bom embasamento teórico na área de codificação e teoria da informação. Uma das maiores vantagens

apontadas pelos defensores deste método é que ele dispensa a existência de conhecimento à priori. De uma forma geral, os algoritmos que realizam este tipo de pontuação dão prioridade às redes bayesianas mais compactas. O princípio do MDL é fazer uma equalização entre complexidade e acuidade do modelo. A seguir estão duas descrições de algoritmos para aprendizado de RB que utilizam o MDL como medida de qualidade da rede.

O algoritmo **Lam-Bacchus** [Lam & Bacchus, 1994] não precisa de uma ordenação das variáveis e pode orientar os arcos usando um método de busca. Os autores usam o conjunto de dados ALARM [Beinlich et al., 1989] para avaliar seu algoritmo, encontrando uma estrutura com três arcos ausentes e dois arcos orientados erroneamente, quando comparado com a rede ideal. A rede ALARM é um conjunto de dados amplamente aceito para algoritmos de aprendizagem.

O algoritmo de aprendizado de Suzuki [Suzuki, 1996], diferentemente da maioria dos outros algoritmos de busca e pontuação, não utiliza uma busca heurística e garante que uma estrutura ótima é encontrada. Como o espaço de busca é enorme, o autor desenvolveu uma técnica "ramificar e podar", que calcula uma ramificação mínima depois da adição de um arco à estrutura e determina se uma busca adicional neste ramo é necessária. Em teste com dados da rede ALARM, mostrando que com cem, duzentos, quinhentos e mil casos, o algoritmo é mais eficiente e preciso do que o algoritmo K2. Todavia, é menos eficiente quando o número de casos cresce para muitos milhares.

Pontuação Bayesiana

Os algoritmos deste tipo avaliam a qualidade das redes bayesianas pela sua distribuição posterior. Dado n redes M_i, i=1...n, e um conjunto de dados D, calcula-se a pontuação para estes modelos usando o teorema de Bayes da seguinte forma:

$$\mathrm{P}(M_i|D) = \frac{P(M_i)P(D|M_i)}{P(D)} \tag{4.1}$$

em que $P(M_i|D)$ é a probabilidade posterior da rede, dado os dados, $P(M_i)$ é a probabilidade a priori da rede, $P(D|M_i)$ é o *likelihood* e P(D) é a probabilidade dos dados. A rede M_i que possuir maior probabilidade posterior será selecionada.

As principais propostas nesta categoria para aprendizado de redes bayesianas são listadas a seguir.

O algoritmo **K2** [Cooper & Herskovits, 1992] é o algoritmo mais representativo dos algoritmos baseados em busca e pontuação para aprendizado em Redes Bayesianas. Este algoritmo é muito conhecido devido aos resultados precisos obtidos quando aplicado ao conjunto de dados da rede ALARM [Beinlich et al., 1989].

Com a finalidade de simplificar a equação (4.1), as seguintes assunções são consideradas. Primeiramente, como todas as redes são avaliadas em função do mesmo conjunto de dados, o denominador da equação (4.1) pode sumir. Em segundo, todas as redes possuem a mesma probabilidade a priori, isto é, todos $P(M_i)$=1. Desta forma, a equação (4.1) se reduz a

$$P(M_i|D) = P(D|M_i)$$

Ou seja, a probabilidade posterior é igual ao *likelihood*. A tarefa, então, se resume a encontrar a rede com maior verossimilhança com os dados. A função de verossimilhança é dada por [Geiger & Heckerman, 1995]:

$$P(D|M_i) = \Pi_{i=1}^{n} \Pi_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \Pi_{k=1}^{r_i} N_{ijk}!$$

em que n é número de variáveis do problema; q_i é o número de instâncias dos pais da variável X_i; r_i são os valores de X_i; N_{ijk} é o número de casos em que a variável X_i assume valor k e seus pais valor j; e $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

O K2 utiliza um algoritmo de busca gulosa para varrer o espaço de busca de todas as possíveis redes. Entretanto, outros algoritmos de busca podem ser utilizados, como por exemplo, algoritmos genéticos [Etxeberria et al., 1997] e [Larrañaga et al., 1996].

Ordenação das variáveis é um pré-requisito também para a execução do K2. Na seção 4.2.3 é descrito um algoritmo que propõe a ordenação das variáveis com base em uma análise de independência.

O algoritmo **HGC** (as siglas vem das inicias dos autores, Heckerman, Geiger e Chickering) [Heckerman et al., 1995a] é um algoritmo baseado em pontuação Bayesiana. A importância deste trabalho é que pelo estudo de propriedades consistentes e suposições de métodos de pontuação, eles encontraram duas suposições, chamadas modularidade de parâmetros e

equivalência de eventos, que têm sido ignoradas por outros pesquisadores. Os autores mostram que um método direto pode ser obtido a partir da combinação do conhecimento de usuários e de dados estatísticos.

O método Bound and Collapse (BC), recentemente proposto em [Ramoni & Sebastiani, 2001], adota a medida de pontuação Bayesiana e é indicado para conjuntos de dados com valores ausentes. O método BC baseia-se na mesma idéia do algoritmo K2 para aprender a estrutura e as probabilidades de um Rede Bayesiana a partir de base de dados possivelmente incompletos. Os autores desenvolveram um sistema que executa duas tarefas: a) extrai o grafo da rede a partir da informação disponível em uma base de dados através de um algoritmo de busca; e b) estimar as probabilidades condicionais da rede extraída utilizando algum algoritmo para estimar dados ausentes, como o algoritmo Maximização da Esperança (EM, do inglês *Expectation Maximization*) e o algoritmo *Gibbs Sampling*.

4.2.2 Método por Independência Condicional

No paradigma de análise de independência condicional, o problema da aprendizagem é abordado de maneira diferente. Uma vez que a estrutura traz embutida muitas dependências do modelo subjacente, os algoritmos deste paradigma tentam descobrir as dependências a partir dos dados, e então usar essas dependências para inferir a estrutura. As relações de dependência são avaliadas pelo uso de alguma classe de teste de Independência Condicional (IC).

De acordo com [Cheng et al., 1997], são poucos os algoritmos

nesta categoria e os existentes são computacionalmente custosos e quase impraticáveis. O algoritmo **Wermuth-Lauritzen**, proposto em [Wermuth & Lauritzen, 1983], requer que as variáveis estejam ordenadas e é indicado para grandes conjuntos de dados com poucas variáveis, uma vez que o teste de independência condicional entre as variáveis é complexo. O algoritmo **SRA**, proposto em [Sampath et al., 1990] tenta aliviar o teste de independência para todas as variáveis usando um algoritmo de busca para selecionar quais variáveis serão testadas e, conseqüentemente, adicionadas ao grafo. Entretanto, os resultados deste algoritmo não são convincentes. O algoritmo **SGS**, apresentado em [Spirtes et al., 1991], é similar ao Wermuth-Lauritzen, entretanto não requer uma ordenação das variáveis. Ele possui a mesma desvantagem do primeiro, o que o torna impraticável. De acordo com [Cheng et al., 1997], outros algoritmos baseados em análise de independência que não são aplicáveis são **CDL**, **Constructor** e **Boundary DAG**.

4.2.3 Métodos Híbridos

Ambos os paradigmas, pontuação e análise de independência, têm vantagens e desvantagens. Por serem bastante específicas às suas aplicações, não se pode eleger um paradigma como sendo o melhor. Em geral, o paradigma de análise de independência é mais eficiente quando usado para aprender redes esparsas (que não são densamente conectadas) com poucas variáveis. Mas, muitos destes algoritmos requerem um número exponencial de testes IC, o que exige elevado esforço computacional. Por outro lado, os

algoritmos de pontuação são capazes de encontrar uma rede muito próxima da rede ótima. Entretanto, eles possuem o inconveniente de pré-determinar a ordenação das variáveis antes de aplicá-los.

O algoritmo **CB** [Singh & Valtorta, 1995] é uma tentativa recente de solucionar o problema de ordenação de variáveis de algoritmos baseados em busca e pontuação. Como os algoritmos baseados em análise de dependência possuem habilidade para orientar arcos, os autores desenvolveram um algoritmo híbrido que primeiramente emprega o algoritmo baseado em testes de independência condicional para propor uma ordenação dos nós e, em seguida, aplica o algoritmo K2 para construir a Rede Bayesiana a partir da ordenação gerada na primeira fase. Assim, o algoritmo evita a necessidade da informação de uma ordenação das variáveis por especialistas.

Outros trabalhos que sugerem métodos heurísticos ou a combinação de algoritmos para ordenar as variáveis podem ser encontrados em [Hruschka & Ebecken, 2003], [Acid & de Campos, 2001], [Dash & Druzdzel, 1999] e [Acid & Campos, 1996b].

Capítulo 5

Considerações Finais

Este trabalho apresentou uma revisão bibliográfica básica dos principais algoritmos para aprendizado de redes bayesianas a partir de conjuntos de dados.

Foi visto que há dois principais paradigmas de aprendizado de redes bayesianas a partir de dados. O paradigma de busca e pontuação e o de independência condicional. Ambos os paradigmas têm vantagens e desvantagens. Por serem bastante específicas às suas aplicações, não se pode eleger um como sendo o melhor. Em geral, o paradigma de análise de independência é mais eficiente quando usado para aprender redes esparsas (que não são densamente conectadas) e com poucas variáveis. Mas, muitos destes algoritmos requerem um número exponencial de testes de independência condicional, o que exige elevado esforço computacional. Por outro lado, os algoritmos de pontuação têm a capacidade de encontrar uma rede muito próxima da rede ótima. Entretanto, eles possuem o inconveniente de ter que pré-determinar a ordenação das variáveis antes de aplicá-los.

As limitações dos dois tipos de algoritmos têm motivado pesquisadores a criar algoritmos híbridos, que combinam as vantagens dos métodos de pontuação com os de análise de independência.

Dos algoritmos apresentados aqui, a maioria requer que as variáveis sejam discretas e o conjunto de dados esteja completo, isto é, sem valores ausentes. Entretanto, nos problemas do mundo real, isso nem sempre é possível e são poucos os algoritmos para aprendizado de redes bayesianas que conseguem atender tais especificações.

Referências Bibliográficas

[Abramson, 1994] Abramson, B. "The design of belief network-based systems for price forecasting". *Comput. Electr. Eng.* **20**(2), pp. 163–180, 1994.

[Acid & Campos, 1996a] Acid, S.; Campos, L. M. "An algorithm for finding minimum d-separating sets in belief networks". *Proc. of 12th Conference of Uncertainty in Artificial Intelligence* **1**, 1996.

[Acid & Campos, 1996b] Acid, S.; Campos, L. M. "Benedict: An algorithm for learning probabilistic belief networks". *Proc. of Sixth International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems* , pp. 974–979, 1996.

[Acid & de Campos, 2001] Acid, S.; de Campos, L. M. "A hybrid methodology for learning belief networks: Benedict". *International Journal of Approximate Reasoning* **27**(3), pp. 235–262, 2001.

[Akaike, 1974] Akaike, H. A. "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* **AC-19**, pp. 716–723, 1974.

[Andersen, 1989] Andersen, S. "Hugin - a shell for building bayesian belief universes for expert systems". *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* , pp. 1080–1085, 1989.

[Androutsopoulos et al., 2000] Androutsopoulos, I.; Koutsias, J.; Chandrinos, K.; Paliouras, G.; Spyropoulos, C. "An evaluation of naive bayesian anti-spam filtering". *Proc. of the Workshop on Machine*

Learning in the New Information Age, 11th European Conference on Machine Learning , pp. 9–17, 2000.

[Antal et al., 2003] Antal, P.; Fannes, G.; Timmerman, D.; Moor, B. D.; Moreau, Y. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor 24 classification with rejection. *Artificial Intelligence in Medicine* **29**, pp. 39–60, 2003.

[Arnborg, 1999] Arnborg, S. "A survey of bayesian data mining - part I: Discrete and semi-discrete data matrices". Relatório Técnico T99-08, 1999.

[Bahle & Navarro, 2000] Bahle, D.; Navarro, L. Methods for combining heterogeneous sets of classifiers. *17th National Conference on Artificial Intelligence (AAAI)* , 2000.

[Beinlich et al., 1989] Beinlich, I. A.; Suermondt, H. J.; Chavez, R. M.; Cooper, G. F. "The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks". *Proceedings of the Second Conference on Artificial Intelligence in Medicine* , pp. 247–256, 1989.

[Breese & Blake, 1995] Breese, J.; Blake, R. "Automating computer bottleneck detection with belief nets". *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **1**, pp. 36–45, 1995.

[Buntine, 1994] Buntine, W. "Operations for learning with graphical models". *Journal of Artificial Intelligence Research* **2**, pp. 159–225, 1994.

[Buntine, 1996] Buntine, W. "A guide to the literature on learning probabilistic networks from data. *Proc. of IEEE Transactions on Knowledge and Data Engineering* **8**, pp. 195–210, 1996.

[Cano et al., 2004] Cano, R.; Sordo, C.; Gutiérrez, J. M. "Applications of bayesian networks in meteorology". In et al., G. (ed), *Advances in Bayesian Networks*, pp. 309–327. Springer, 2004.

[Charniak, 1991] Charniak, E. "Bayesian network without tears". *AI Magazine* **12**(4), pp. 50–63, 1991.

[Cheng et al., 1997] Cheng, J.; Bell, D. A.; Liu, W. "Learning belief networks from data: an information theory based approach". In *Proceedings of the sixth international conference on Information and knowledge management*, pp. 325–331, New York, NY, USA. ACM Press, 1997.

[Chow & Liu, 1968] Chow, C. J.; Liu, C. N. "Approximating discrete probability distributions with dependence trees". *IEEE Trans. on Information Theory* **14**(3), pp. 462–467, 1968.

[Cooper & Herskovits, 1992] Cooper, G.; Herskovits, E. "A bayesian method for the induction of probabilistic networks from data". *Machine Learning* **9**, pp. 309–347, 1992.

[Cowell et al., 1999] Cowell, R. G.; Dawid, A. P.; Lauritzen, S. L.; Spiegelhalter, D. J. *"Probabilistic networks and expert systems"*. Springer, 1999.

[Dash & Druzdzel, 1999] Dash, D.; Druzdzel, M. J. "A hybrid anytime algorithm for the construction of causal models from sparse data". In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, pp. 142–149, 1999.

[Dean, 1990] Dean, T. "Coping with uncertainty in a control system for navigation and exploration". *Proceedings of the Ninth National Conference on Artificial Intelligence* , pp. 1010–1015, 1990.

[Ducksbury, 1993] Ducksbury, P. G. "Parallel texture region segmentation using a pearl bayes network". *Proceedings of the 4th British Machine Vision Conference* , pp. 187–195, 1993.

[Duda et al., 1976] Duda, R.; Hart, P.; Nilsson, N. "Subjective bayesian methods for rule-based inference systems". *Proceedings of the American Federation of Information Processing Societies National Computer Conference* , pp. 1075–1082, 1976.

[Etxeberria et al., 1997] Etxeberria, R.; Larrañaga, P.; Pikaza, J. M. "Analysis of the behaviour of the genetic algorithms when searching bayesian networks from data". *Pattern Recognition Letters* **18**(11-13), pp. 1269–1273, 1997.

[Ezawa & Schuermann, 1995] Ezawa, K. J.; Schuermann, T. "Fraud/uncollectable debt detection using a bayesian network based learning system: A rare binary outcome with mixed data structures". *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **1**, pp. 157–166, 1995.

[Friedman et al., 1997] Friedman, N.; Geiger, D.; Goldszmidt, M. "Bayesian networks classifiers". *Machine Learning* **29**, pp. 131–163, 1997.

[Friedman et al., 2000] Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. "Using bayesian networks to analyze expression data". *Proc. Fourth Annual International Conference on Computational Molecular Biology* **1**, pp. 127–135, 2000.

[Geiger, 1992] Geiger, D. "An entropy-based learning algorithm of bayesian conditional trees". In *Proceedings of the eighth conference on Uncertainty in Artificial Intelligence*, pp. 92–97, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 1992.

[Geiger & Heckerman, 1995] Geiger, D.; Heckerman, D. "A characterization of the dirichlet distribution with application to learning bayesian networks". In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, pp. 196–207, 1995.

[Greiner et al., 1997] Greiner, R.; Grove, A.; Schuurmans, D. "Learning bayesian nets that perform well". *Proceedings of Conference on Uncertainty in Artificial Intelligence* , 1997.

[Grünwald, 2000] Grünwald, P. "Model selection based on minimum description length". *Journal of Mathematical Psychology* **44**, pp. 133–152, 2000.

[Habrant, 1999] Habrant, J. Structure learning of bayesian networks from databases by genetic algorithms: Application to time series prediction in finance. *Proceedings of the 1st International Conference on Enterprise Information Systems* **1**, pp. 225–231, 1999.

[Heckerman, 1995] Heckerman, D. "A Tutorial on Learning Bayesian Networks". Relatório Técnico MSR-TR-95-06, 1995.

[Heckerman & Chickering, 1996] Heckerman, D.; Chickering, M. "Efficient approximation for the marginal likelihood of incomplete data given a bayesian network". Relatório Técnico MSR-TR-96-08, 1996.

[Heckerman et al., 1995a] Heckerman, D.; Geiger, D.; Chickering, D. "Learning bayesian networks: The combination of knowledge and statistical data". *Machine Learning* **20**, pp. 197–243, 1995.

[Heckerman et al., 1995b] Heckerman, D.; Mamdani, A.; Wellman, M. P. "Real-world applications of bayesian networks". *Communication of ACM* **38**(3), pp. 24–26, 1995.

[Horvitz et al., 2001] Horvitz, E.; Ruan, Y.; Gomes, C. P.; Kautz, H.; Selman, B.; Chickering, D. M. A bayesian approach to tackling hard computational problems. pp. 235–244, 2001.

[Hruschka & Ebecken, 2003] Hruschka, E.; Ebecken, N. F. F. "Variable ordering for bayesian networks learning from data". *International Conference on Computational Intelligence for Modelling, Control and Automation - CIMCA 2003* , 2003.

[Imoto et al., 2003] Imoto, S.; Kim, S.; Goto, T.; Aburatani, S.; Tashiro, K.; Kuhara, S.; Miyano, S. "Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology* **1**(2), pp. 231–252, 2003.

[Inza et al., 2000] Inza, I.; Larrañaga, P.; Etxeberria, R.; Sierra, B. Feature subset selection by bayesian network-based optimization. *Artificial Intelligence* **123**, 2000.

[Jensen, 1996] Jensen, F. V. *"An Introduction to Bayesian Networks"*. UCL press, London, 1996.

[Jensen, 2001] Jensen, F. V. *"Bayesian Networks and Decision Graphs"*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[Kahn et al., 1995] Kahn, C. E.; Roberts, L. M.; Wang, K.; Jenks, D.; Haddawy, P. Preliminary investigation of a bayesian network for mammographic diagnosis of breast cancer. *Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care* **1**, pp. 208–212, 1995.

[Klinger & Brutlag, 1994] Klinger, T.; Brutlag, D. "Discovering structural correlations in alpha-helices". *Protein Science* **3**, pp. 1847–1857, 1994.

[Lam & Bacchus, 1994] Lam, W.; Bacchus, F. "Learning bayesian belief networks: An approach based on the mdl principle". *Computational Intelligence* **10**(4), pp. 269–293, 1994.

[Lampinen & Vehtari, 2000] Lampinen, J.; Vehtari, A. "Bayesian techniques for neural networks - review and case studies". In Gabbouj, M.; Kuosmanen, P. (eds), *Proceedings of Eusipco'2000, X European Signal Processing Conference*, volume 2, pp. 713–720, Tampere, Finland, 2000.

[Larrañaga et al., 1996] Larrañaga, P.; Kuijpers, C.; Murga, R.; Yurramendi, Y. "Learning bayesian network structures by searching for the best ordering with genetic algorithms". *IEEE Transactions on System, Man and Cybernetics* **26**(4), pp. 487–493, 1996.

[Leon-Garcia, 1994] Leon-Garcia, A. *Probability and Random Processes for Electrical Engineering.* Addison-Wesley, 2nd edition, 1994.

[MacKay, 1996] MacKay, D. J. C. "Bayesian methods for back-propagation networks". In Domany, E.; van Hemmen, J. L.; Schulten, K. (eds), *Models of Neural Networks III*, p. 309. Springer-Verlag, 1996.

[Myllymäki et al., 2001] Myllymäki, P.; Silander, T.; Tirri, H.; Uronen, P. "Bayesian data mining on the web with b-course", 2001.

[Naim et al., 2004] Naim, P.; Wuillemim, P.; Leray, P.; Pourret, O.; Becker, A. *Réseaux Bayésiens*. Eyrolles, Paris, 2004.

[Nakajima et al., 1998] Nakajima, Y.; Sugi, J.; Saito, M.; Hamagishi, H.; Hattori, D.; Matsumoto, T. "Hierarchical bayesian neural nets for air-conditioning - load prediction: Nonlinear dynamics approach". *IEEE World Congress on Computacional Intelligence* **1**, 1998.

[Nariai et al., 2004] Nariai, N.; Kim, S.; Imoto, S.; Miyano, S. "Using protein-protein interactions for refining gene networks estimated from microarray data by bayesian networks", 2004.

[Neal, 1996] Neal, D. H. *"Bayesian learning for neural networks"*. Cambridge University Press, 1996.

[Papoulis & Pillai, 2002] Papoulis, A.; Pillai, S. U. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Boston, MA, 4th edition, 2002.

[Pearl, 1988] Pearl, J. *"Probabilistic reasoning in intelligent systems: networks of plausible inference"*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Peng & Reggia, 1987] Peng, Y.; Reggia, J. A probabilistic causal model for diagnostic problem solving - part 2: Diagnostic strategy. *IEEE Trans. on Systems, Man, and Cybernetics: Special Issue for Diagnosis* **17**, pp. 395–406, 1987.

[Ramoni & Sebastiani, 1999] Ramoni, M.; Sebastiani, P. *Bayesian methods in Intelligent Data Analysis. An Introducction*. Physica Verlag, Heidelberg, 1999.

[Ramoni & Sebastiani, 2001] Ramoni, M.; Sebastiani, P. "Robust learning with missing data". *Machine Learning* **45**(2), 2001.

[Raval et al., 2002] Raval, A.; Ghahramani, Z.; Wild, D. L. "A bayesian network model for protein fold and remote homologue recognition. *Bioinformatics* **18**(6), pp. 88–780, 2002.

[Rebane & Pearl, 1987] Rebane, G.; Pearl, J. "The recovery of causal polytrees from statistical data". *Proc. of Third Workshop of Uncertainty in Artificial Inteligence* , pp. 222–228, 1987.

[Rehg et al., 1999] Rehg, J.; Murphy, K.; Fieguth, P. Vision-based speaker detection using bayesian networks. *Computer Vision and Pattern Recognition* **1**, 1999.

[Sahami et al., 1998] Sahami, M.; Dumais, S.; Heckerman, D.; Horvitz, E. "A bayesian approach to filtering junk e-mail". *Workshop on Learning for Text Categorization* , 1998.

[Sampath et al., 1990] Sampath, S.; Stuart, R.; Alice, A. "Automated construction of sparse bayesian networks from unstructured probabilistic models and domain information". In *Proceedings of the 5th Annual Conference on Uncertainty in Artificial Intelligence (UAI-90)*, New York, NY. Elsevier Science Publishing Company, Inc., 1990.

[Schwarz, 1978] Schwarz, G. "Estimating the dimension of a model". *Annals of Statistics* **6**, pp. 461465, 1978.

[Shortliffe, 1976] Shortliffe, E. *Computer-Based Medical Consultations: MYCIN*. American Elsevier, 1976.

[Singh & Valtorta, 1995] Singh, M.; Valtorta, M. "Construction of bayesian belief networks from data: a brief survey and an efficient algorithm". *International Journal of Approximate Reasoning* **12**(2), pp. 111–131, 1995.

[Spirtes et al., 1991] Spirtes, P.; Glymour, C.; Scheines, R. "An algorithm for fast recovery of sparse causal graphs". *Social Science Computer Review* **9**, pp. 62–72, 1991.

[Stephenson et al., 2000] Stephenson, T.; Bourlard, H.; Bengio, S.; Morris, A. Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables. In *International Conference on Spoken Language Processing, October 2000, vol. II*, pp. 951–954, 2000.

[Suzuki, 1996] Suzuki, J. "Learning bayesian belief networks based on the mdl principle: An efficient algorithm using the branch and bound technique". *Proc. of International Conference on Machine Learning* **1**, 1996.

[Sykacek et al., 1998] Sykacek, P.; Dorffner, G.; Rappelsberger, P.; Zeitlhofer, J. Experiences with bayesian learning in a real world application. *Advances in Neural Information Processing Systems* **10**, 1998.

[Vehtari & Lampinen, 1999] Vehtari, A.; Lampinen, J. Bayesian neural networks for industrial applications, 1999.

[Vehtari & Lampinen, 2000] Vehtari, A.; Lampinen, J. "Bayesian MLP neural networks for image analysis". *Pattern Recognition Letters* **21**(13-14), pp. 1183–1191, 2000.

[Wermuth & Lauritzen, 1983] Wermuth, N.; Lauritzen, S. "Graphical and recursive models for contingency tables". *Biometrika* **72**, pp. 537–552, 1983.

[Zweig & Russell, 1998] Zweig, G.; Russell, S. J. Speech recognition with dynamic bayesian networks. In *American Association for Artificial Intelligence*, pp. 173–180, 1998.