Modelos gráficos probabilísticos

Leonardo de Assis da Silva[1]

[1]Departamento Acadêmico de Informática
Universidade Tecnológica Federal do Paraná (UTFPR)
Avenida Sete de Setembro – 3165 – 80.230-901 – Curitiba – PR – Brasil

leosil@alunos.utfpr.edu.br

Abstract. *This paper seeks to give an introduction to graphical probabilistic models through Bayesian networks and Markov random fields. Thus, the key aspects of each are given in turn. Furthermore, a concrete application of Bayesian network is presented afterwards.*

Resumo. *Este artigo descreve brevemente os modelos gráficos probabilísticos com as abordagens de redes Bayesianas e campos aleatórios de Markov, levantando seus principais aspectos. Posteriormente, é exposto uma rede Bayesiana aplicada no mundo real.*

1. Introdução

Aplicações do mundo real muitas vezes devem lidar com situações de incerteza e imprecisão. No entanto, um agente autônomo mesmo diante desse cenário deve ser capaz de determinar suas ações de modo a alcançar seus objetivos. Formas de tratar este problema foram propostas por diversos ramos do campo de Inteligência Artificial, porém a abordagem tratada neste artigo são os modelos gráficos probabilísticos.

Na seção 2 foram listadas as características gerais dos modelos gráficos probabilísticos, nas duas seções seguintes duas especializações foram descritas, iniciando na seção 3 com as redes Bayesianas e, em seguida, os campos aleatórios de Markov na seção 4. Um exemplo concreto utilizando rede Bayesiana foi dado na seção 5, finalizando na seção 6 com uma breve comparação entre as duas abordagens de modelos gráficos probabilísticos e instruções sobre como escolher qual deve ser utilizada.

2. Modelos Gráficos Probabilísticos

Originados a partir da combinação da teoria da probabilidade e teoria dos grafos, os modelos gráficos probabilísticos (MGP) proporcionam uma estrutura intuitiva e, comparativamente às tabelas de distribuição conjunta total, enxuta onde as relações de dependência e suas distribuições de probabilidade são representadas.

2.1. Representação

Um modelo gráfico probabilístico é composto por tabelas de distribuição de probabilidade, conjunto de vértices V e conjunto de arestas E tal que V é o conjunto de todas as variáveis aleatórias e E denota a dependência existente entre duas variáveis.

Um MGP pode tomar a forma de um grafo direcionado acíclico sendo então denominado rede Bayesiana, ou grafo não direcionado chamado campos aleatórios de Markov ou ainda um grafo híbrido que utilize ambas as representações simultaneamente.

2.2. Inferência

O processo de inferência consiste em integrar o processo de cálculo de distribuição de probabilidade com as regras da lógica formal, com o intuito de transformar expressões difíceis de calcular devido à falta de conhecimento sobre distribuições condicionais em outras formas equivalentes mais fáceis de serem tratadas. Esse processo é empregado durante consultas para a descoberta, por exemplo, da probabilidade de determinada variável aleatória ocorrer considerando o acontecimento de determinadas evidências ou a variável aleatória mais provável dado certa evidência.

As técnicas de inferência são categorizadas em exatas e aproximadas. A inferência aproximada embora menos confiável geralmente é a abordagem aplicada no mundo real já que a inferência exata é muitas vezes intratável em modelos probabilísticos de tamanho massivo.

A inferência exata consiste em descobrir a distribuição de uma variável aleatória através da enumeração de todas os eventos em que esta ocorre em conjunto à variáveis ocultas (não relevantes) dado a ocorrência de outras variáveis aleatórias de evidência. Essa abordagem, como dito anteriormente, é intratável na prática por apresentar tempo exponencial no pior caso. Alternativas para esse problema consistem na eliminação de variáveis, inferência aproximada, entre outras.

A inferência aproximada apresenta bons resultados na prática. Uma das técnicas mais simples consiste na utilização de amostras aleatórias, podendo estas serem coletadas através de um algoritmo como o da Cadeia de Markov Monte Carlo, no caso das redes Bayesianas.

2.3. Aprendizagem

Situações de incerteza causadas pela falta de conhecimento teórico e/ou prático sobre o mundo ou pela observacionalidade parcial são aspectos difíceis de serem solucionados diretamente através da modelagem.

Uma alternativa para a modelagem direta é a aprendizagem da estrutura da rede e de seus parâmetros. O método a ser empregado dependerá da combinação desses dois fatores[5], assim é utilizado: Máxima Verossimilhança quando estrutura e parâmetros são conhecidos, Máxima Esperança para estrutura conhecida e parâmetros parcialmente conhecidos, busca no espaço de modelos quando são conhecidos os parâmetros e não a estrutura e Máxima Esperança e busca no espaço de modelos se ambos são desconhecidos.

3. Redes Bayesianas

Para tornar evidente o motivo de utilização das redes Bayesianas primeiramente considere a necessidade de realizar uma consulta no cenário de inexistência de redes Bayesianas, será necessária a declaração da tabela de distribuição conjunta. Esta determina a probabilidade de todos os eventos possíveis, assim, considerando um caso com duas variáveis aleatórias booleanas a distribuição conjunta é calculada como $P(X, Y) = P(X, Y)P(\neg X, Y)P(X, \neg Y)P(\neg X, \neg Y)$, de forma que neste caso, mais simples, utilizando somente variáveis discretas booleanas já seria necessário o preenchimento de $2^n - 1$ entradas.

Outro fator para o uso destas redes é a simplificação gerada ao assumir que as variáveis aleatórias são dependentes somente dos vizinhos presentes na cobertura de Markov, ou seja, dado os pais, os filhos e os pais dos filhos de uma variável aleatória, esta é independente dos demais vértices da rede segundo a hipótese local de Markov. Assim, na rede representada na figura 1 a seguinte distribuição pode ser derivada através da regra da cadeia, cálculo de distribuição conjunta através de probabilidade condicional, assumindo tal hipótese:

$$P(W, X, Y, Z) = P(Z)P(Y)P(X|YX)P(W|X) \tag{1}$$

3.1. Representação

Como foi representado na figura 1, nas Redes Bayesianas os pais de um vértice qualquer v denotam quais variáveis aleatórias este é dependente. De maneira que para o conjunto de arestas E, o conjunto de pais de v, $Pais_v$, é definido por $\forall_{p \in Pais_v}(p, v) \in E$, onde $p, v \in V$, então existe uma tabela de probabilidade condicional (TPC) que contém a distribuição de probabilidade de v para cada combinação de verdadeiro/falso das variáveis aleatórias contidas em $Pais_v$.

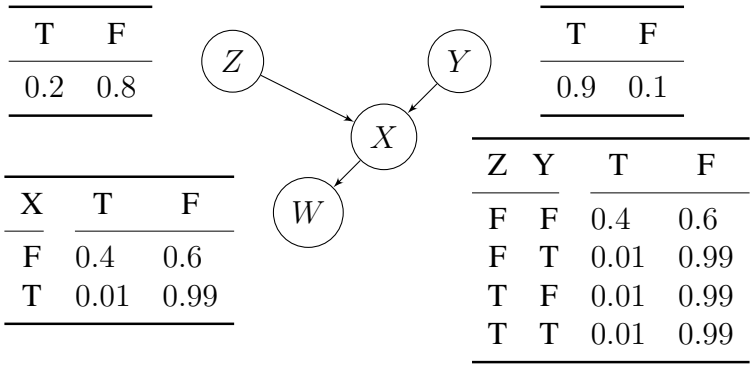

Figura 1. Rede Bayesiana simples com 4 vértices.

Logo, para a criação de uma rede é necessário apenas o conhecimento das probabilidades a priori de cada raiz e das probabilidades condicionais para todos os eventos relatados em cada TPC.

3.1.1. Separação D

Uma interessante informação que pode ser extraída desta representação é a existência, ou não, da independência condicional entre duas variáveis aleatórias. Esta relação pode ser visualizada como um fluxo de caminho de um vértice ao outro através de um terceiro vértice intermediário, desconsiderando a orientação de suas arestas. Quatro tipos de caminhos existem, quando a variável intermediária é conhecida existe dependência entre as demais apenas no caminho de efeito comum, já quando a variável intermediária é desconhecida existe dependência entre as demais nos caminhos de causalidade, de evidência e de causa comum. Nos outros casos as variáveis são condicionalmente independentes.

3.2. Inferência Exata

Neste tipo de rede a inferência exata acontece através da enumeração de todos os eventos com a presença de determinadas evidências em que a variável aleatória consultada ocorre, ou seja, há o somatório da distribuição de probabilidade dos eventos em que variáveis ocultas assumem todas suas possíveis valorações. Esse processo pode ser melhorado através da eliminação de variáveis, programação dinâmica ou a formação de agrupamentos para a obtenção de uma poliárvore.

As poliárvores, ou redes unicamente conectadas, tratam-se de grafos subjacente não direcionados no quais existem somente um caminho entre quaisquer dois vértices. A rede Bayesiana da figura 1 é uma poliárvore.

3.2.1. Complexidade

Como a inferência exata utiliza implicitamente os valores presentes em uma tabela de distribuição conjunta total, citada anteriormente por seu tamanho exponencial, o algoritmo apresenta complexidade de tempo de $O(2^n)$ no caso geral.

Ademais, embora o caso geral seja NP-difícil, ao restringirmos a inferência exata à poliárvores a complexidade é reduzida para linear ao número de vértices.

3.3. Inferência Aproximada

Existem diversas técnicas de inferência aproximada, sendo citadas em [5] os métodos variacionais, amostragens, propagação de crença em laço, condicionamento de corte limitado e métodos de aproximação paramétricas. Neste artigo é descrito brevemente uma técnica de amostragem, o algoritmo de Cadeia de Markov Monte Carlo (CMMC).

3.3.1. Cadeia de Markov Monte Carlo

O CMMC é uma técnica formada pela junção da integração Monte Carlo, cadeia de Markov e o algoritmo Metropolis-Hastings [2].

A integração Monte Carlo consiste em estimar valores de integrais a partir de amostras aleatórias. Como a população pode não estar padronizada a realização deste processo depende da garantia de existência de uma distribuição estacionária, estando esta presente nas cadeias de Markov. Uma sequência de variáveis aleatórias na forma $X_1, X_2, ..., X_t$ na qual os estados são gerados a partir da anterior e independente das demais é denominada cadeia de Markov, ou seja, a distribuição de X_{t+1} é determinada através de $P(X_{t+1}|X_t)$. Por último, o algoritmo Metropolis-Hastings fornece uma maneira de construir uma cadeia de Markov que apresente distribuição estacionária compatível à distribuição real da população da rede.

Uma variação do algoritmo Metropolis-Hastings foi proposto por Gibbs, este algoritmo consiste em retirar amostras de distribuições condicionais totais.

3.3.2. Complexidade

Embora a inferência aproximada apresente bons resultados na prática, ela possui complexidade #P-difícil, isto é, um problema de contagem estritamente mais difícil que a classe NP-completo[7].

3.4. Aplicação

As Redes Bayesianas são utilizadas principalmente como sistemas causais, de modo que entre suas aplicações mais habituais esta a atribuição de probabilidades à sistemas especialistas, tendo como exemplo o sistema de diagnósticos QMR-DT, o assistente Clippy da Microsoft, o sistema meteorológico Vista da Nasa e os classificadores ingênuos Bayesianos. Além disso, várias especializações com diversas utilidades foram derivadas das redes Bayesianas, como os modelos Ocultos de Markov que permitem o reconhecimento de voz, as redes Bayesianas dinâmicas utilizadas na robótica e os sistemas dinâmicos lineares, que com o uso de filtros de Kalman, são capazes de implementar atividades de rastreamento (GPS).

4. Campos Aleatórios de Markov

Campos aleatórios de Markov, também chamados de redes de Markov (RM), trata-se de grafos não direcionados que representam variáveis aleatórias (vértices) e as relações probabilísticas existentes entre elas (arestas). Diferencia-se das redes Bayesianas por representar a independência de forma mais simples e a causalidade de forma menos intuitiva.

4.1. Representação

Na rede de Markov as distribuições de probabilidades não são definidas somente para os vértices mas também para as arestas. No entanto, nesta a distribuição de um conjunto X não é calculada simplesmente encontrando sua probabilidade condicional, e sim fatorizando cliques maximais de variáveis (subgrafos completos) e definindo suas funções potenciais não negativas que atendam as restrições de distribuição da rede e considerando uma transformação logarítmica, de modo a evitar o aumento exponencial da representação. A equação 2 descrita é referida como função de energia e a equação 3 apresenta o cálculo de distribuição de probabilidade :

$$\epsilon C = -ln\phi C \tag{2}$$

$$P(X) = \frac{1}{Z}\prod_{i=1}^{n}\phi_i C_i \tag{3}$$

Essa equação define a distribuição de Gibbs, onde X é o subconjunto formado por $X_1, X_2, ..., X_n$, C_i é uma clique $\in X$, ϕ_i uma função potencial e Z a função de partição, utilizada para normalizar a distribuição. A instância específica da rede de Markov onde distribuição é calcula considerando apenas variáveis individuais e par de variáveis é denominada rede de Markov emparelhada.

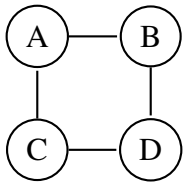

Figura 2. Rede de Markov simples com 4 vértices.

Conforme representado na figura 2, a diferença em relação a rede Bayesiana mais evidente é falta de direção na relação entre dois vértices, isto é, não existe uma hierarquia definida o que por consequência delimita a cobertura de Markov aos vértices diretamente vizinhos de uma variável aleatória. Essa relação de independência é composta por duas propriedades similares à separação D das redes Bayesianas: independência local de Markov e independência global de Markov.

4.1.1. Independência local de Markov

A independência local de Markov é determinada para cada variável $X_i \in X$ e denota a separação entre esta e as restantes dado a cobertura de Markov, ou seja, a existência de uma independência condicional entre a variável aleatória e a demais, denotado por $P(X_i \perp [X - X_i - X_{cobertura}] | X_{cobertura})$ onde $\perp$ representa a independência condicional.

4.1.2. Independência global de Markov

A relação de independência global de Markov entre dois subconjuntos é estabelecida com base na não presença de *caminhos ativos* entre os dois considerando determinada evidência. Caminhos são conjuntos de vértices conectados sequencialmente, sendo este declarado ativo caso nenhum de seus vértices tenha sido observado.

4.2. Técnicas de inferência

4.2.1. Cortes em grafos

Essa técnica consiste em construir um grafo direcionado com um nó inicial e final, s e t, tal que o custo s-t (soma dos custos das arestas do caminho de s à t) é equivalente à representada na rede de Markov. Logo, a inferência pode ser realizada através de um algoritmo que minimize o custo s-t em tempo polinomial, como o algoritmo de Ford-Fulkerson.

Entretanto, essa técnica funciona somente em grafos como as redes de Markov emparelhadas. Estes grafos são denominados grafos representáveis e se destacam por permitirem extrair um mínimo absoluto se a função de energia for submodular. Em casos onde a função de energia não é submodular essa minimização é NP-difícil[9].

4.2.2. Propagação de crenças em laço

A propagação de crenças funciona através da troca de mensagens localmente, sendo capaz de encontrar solução exatas em RMs em forma de árvore e soluções aproximada em grafos com laço[9]. A troca de mensagens denota as variações de crença que uma variável aleatória possui sobre as outras. Essa troca ocorre de maneira iterativa, quando uma variável tem sua distribuição atualizada ela propaga a mudança às demais, com exceção da variável que causou sua mudança, e estas realizam o mesmo processo até que a rede se estabilize.

4.2.3. Complexidade

Analisando as técnicas elencadas, a inferência exata muitas vezes funciona apenas nas RMs emparelhadas, com o caso geral sendo pertencente à classe NP-difícil. Todavia, o caso geral pode ser satisfeito utilizando técnicas de inferência aproximada.

4.3. Aplicação

As redes de Markov são utilizadas principalmente nos campos de visão computacional e física, tendo se originado em sistemas de modelagem interação de partículas[8]. As RMs também possuem variações populares, como o modelo Ising que reproduz o comportamento de imãs, as redes Hopfield utilizadas como memórias associativas, modelo de Potts de segmentação de imagem, campos aleatórios de Markov Gaussiano e redes lógicas de Markov[6].

5. MGP: Answer Wizard

O Answer Wizard, descrito em detalhes em [3], trata-se de um assistente ao usuário do Microsoft Office que oferece alternativas de instruções candidatas à responder uma dúvida do usuário representada em uma consulta em *free-text*. A necessidade de atender esse tipo de consulta surge da falta de conhecimento de termos técnicos por parte do usuário leigo ao expressar seu problema em forma de texto.

Contanto, outros problemas como o uso de termos que atendem à diversos problemas ao mesmo tempo e de palavras ambíguas dificultam e tornam este cenário incerto. Por essa razão, os pesquisadores da Microsoft Research propuseram uma rede Bayesiana para a recuperação de informação.

A função da rede criada era realizar a inferência da probabilidade de o usuário ter um objetivo específico dado que ele digitou determinada consulta: $P(objetivo|consulta)$. Para isso foi utilizado uma metodologia onde a consulta era separada em termos, e heurísticas foram adotadas, como a avaliação apenas da raiz das derivações de um termo, interesse limitado aos termos presentes na base de conhecimento e diferenciação entre substantivo próprio e comum (e.g. *Word* e *word*). Assim, com o auxílio de especialistas foram estabelecidas as probabilidades condicionais de um termo ocorrer considerando um determinado objetivo, essa rede é representada na figura extraída de [3].

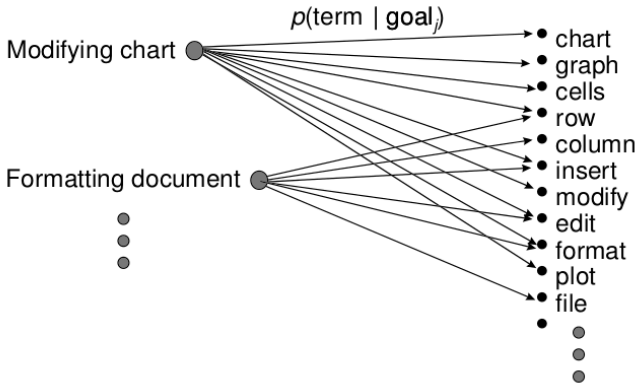

Figura 3. Rede Bayesiana com objetivos representados como pais dos termos.

Na rede da figura acima, as ligações foram limitadas às relações ”relevantes”, ou seja, com o intuito de otimizar a rede foi descartada a representação fiel de todas as possibilidades existentes. Dessa forma, foi necessário a definição de uma probabilidade padrão de vazão (em inglês, *leak probability*) que representasse a chance de um termo ocorrer mesmo que este não estivesse conectado ao objetivo.

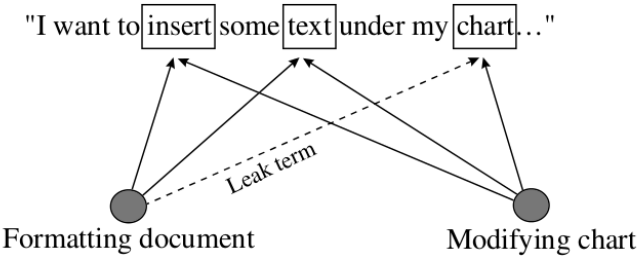

Figura 4. Representação de um termo de vazão. Retirado de [3].

Finalmente, a equação definitiva da probabilidade condicional do objetivo do usuário considera quatro situações: objetivo conectado à termos não presentes na consulta, objetivo conectado à termos presentes na consulta, objetivo não conectado à termos presentes na consulta e objetivo não conectados à termos não presentes na consulta. Essa relação é denotada pela seguinte equação

$$P(objetivo_i|termo^+termo^-) = P(objetivo_i)\prod_j P(termo_j|objetivo_i) \prod_k [1 - P(termo_k|objetivo_i)]\epsilon^l(1-\epsilon)^m \quad (4)$$

Onde $termo^+$ e $termo^-$ é o conjunto de termos presentes na consulta e os não presentes, respectivamente, $termo_k$ são os conectados ao objetivo e não presentes na consulta, $termo_j$ são os conectados e presentes, l o número de termos presentes na consulta que não são ligados ao objetivo e m o número de termos não conectados ao objetivo e não presentes na consulta.

Melhorias adicionais foram implementadas, entre elas a abstração de termos que consiste em formar agrupamentos de palavras de semântica similares, cálculo da probabilidade de uma palavra expressar um substantivo ou verbo (e.g. *print*) e o cálculo da probabilidade de uma consulta se referir à um objeto existente ou um objeto que o usuário deseja criar através da análise de artigos definidos, artigos indefinidos e pronomes possessivos. Com essas otimizações foi possível tornar a rede Bayesiana de recuperação de informação em um assistente presente na versão comercial da suíte Microsoft Office 95, onde a rede possuía 1.000 tópicos, 5.000 termos e 145.000 dependências.

6. Conclusão

Como visto anteriormente, os modelos probabilísticos proporcionam a capacidade de agir em cenários incertos típicos do mundo real, entretanto nos casos gerais, seja em grafos direcionados, seja em grafos não direcionados, a complexidade da inferência exata é exponencial. Logo, o uso de otimizações para a inferência exata, como as apresentadas em [4], e as técnicas de inferência aproximada são cruciais para o bom aproveitamento da abordagem de raciocínio probabilístico.

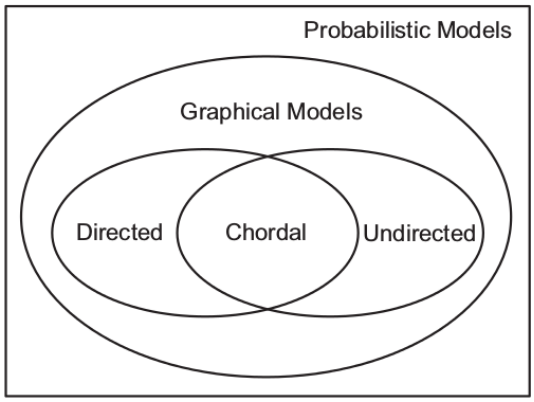

Figura 5. Diagrama de Venn de PGMs. Retirado de [6].

Analisando a imagem 5 é justificável dizer que ambos os tipo de modelos gráficos não são equivalentes e que tampouco existe um modelo que possua maior poder de representação que o outro. De fato, as redes de Markov não são capazes de reproduzir o comportamento da estrutura de causa comum (*estrutura v*) das redes Bayesianas, enquanto estas não conseguem representar relações de influência mútua devido à sua propriedade acíclica.

A existência de categorias de MGP torna necessária a análise de qual modelo é mais adequado para a adoção em determinada aplicação. Algumas características básicas sobre o aplicação que podem ser consideradas durante a escolha são: existência de condicionamento mútuo entre variáveis aleatórias, necessidade de avaliação da causalidade ou a realização de diagnósticos.

Referências

[1] Eugene Charniak. Bayesian networks without tears. *AI magazine*, 12(4):50, 1991.

[2] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

[3] David Heckerman and Eric Horvitz. Inferring informational goals from free-text queries: A bayesian approach. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 230–237. Morgan Kaufmann Publishers Inc., 1998.

[4] Daphne Koller, Nir Friedman, Lise Getoor, and Ben Taskar. Graphical models in a nutshell. *Introduction to statistical relational learning*, pages 13–55, 2007.

[5] Kevin P Murphy. A brief introduction to graphical models and bayesian networks.

[6] Kevin P Murphy. *Machine learning: a probabilistic perspective*. 2012.

[7] P. Norvig and S. Russell. *Inteligência Artificial: Tradução da 3a Edição*. Elsevier Brasil, 2015.

[8] Padhraic Smyth. Belief networks, hidden markov models, and markov random fields: a unifying view. *Pattern recognition letters*, 18(11):1261–1268, 1997.

[9] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, 2013.