LEANDRO ALVES DE SOUZA

APLICAÇÃO DE REDES BAYESIANAS A DOIS PROBLEMAS ATUAIS

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

LAVRAS
MINAS GERAIS – BRASIL
2008

LEANDRO ALVES DE SOUZA

APLICAÇÃO DE REDES BAYESIANAS A DOIS PROBLEMAS ATUAIS

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Área de Concentração:
Inteligência Artificial

Orientador:
Prof. Rudini Menezes Sampaio

LAVRAS
MINAS GERAIS – BRASIL
2008

Ficha Catalográfica preparada pela Divisão de Processos Técnicos

da Biblioteca Central da UFLA

Souza, Leandro Alves de

Aplicação de Redes Bayesianas a Dois Problemas Atuais/ Leandro Alves de Souza. Lavras – Minas Gerais, 2004. 74 p. : il.

Monografia de Graduação – Universidade Federal de Lavras. Departamento de Ciência da Computação.

1. Informática. 2. Inteligência Artificial. 3. Redes Bayesianas. I. Souza, L.A. II. Universidade Federal de Lavras. III. Título.

LEANDRO ALVES DE SOUZA

APLICAÇÃO DE REDES BAYESIANAS A DOIS PROBLEMAS ATUAIS

Monografia de Graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Aprovada em *20 de Novembro de 2008*

Prof. Cristiano Leite de Castro

Prof. Thiago de Souza Rodrigues

Prof. Rudini Menezes Sampaio
(Orientador)

Prof. Ahmed Ali Abdala Esmin
(Co-Orientador)

LAVRAS
MINAS GERAIS – BRASIL
2008

”Precisamos de Santos sem véu ou batina. Precisamos de Santos de calças jeans e tênis. Precisamos de Santos que coloquem Deus em primeiro lugar, mas que se "lascam"na faculdade. Precisamos de Santos que tenham tempo todo dia para rezar e que saibam namorar na pureza e castidade, ou que consagrem sua castidade. Precisamos de Santos que estejam no mundo; e saibam saborear as coisas puras e boas do mundo, mas que não sejam mundanos ” (João Paulo II).

Aos meus pais, Wellington e Maria Aparecida, que com todo esforço, carinho, dedicação e amor, lutaram para educação de nós filhos. Aos meus irmãos Lucas e Gustavo que são presenças maravilhosas em minha vida. À querida Janaina, por ter ”marcado presença” em momentos difíceis e por ter sido tão compreeensiva e carinhosa. Aos meus irmãos em Cristo Jesus, da Missão Maria de Nazaré, em especial pelos presentes de Deus que são o Eduardo e a Carla. Aos eternos amigos da república Galo Doido e república Espartano, em especial ao Moisés Habib e ao Thiago Gomes Gontijo que iniciaram esta caminhada juntos comigo.

Dedico.

Agradeço a Jesus Cristo, Deus verdadeiramente vivo na minha vida e a minha mãe Nossa Senhora, pelo presente de viver e por eu ter alcançado a universidade pública. Ao imenso apoio de toda família. Aos funcionários e professores do DCC, em especial os meus orientadores Rudini Sampaio e Ahmed Esmin. À professora Maria Cristina do Unilavras, pelo apoio. Agradeço aos amigos, pelos momentos de companheirismo, alegria e pelas noites não dormidas juntos, em especial ao Thiago Cristian de Souza, Luis Paulo Magalhães, Wagner Gonçalves, Tony Harley Avelar, André Gomes e Elisa Boari. Aos amigos do Ministério Universidades Renovadas. Agradeço aos amigos de república, que sempre foram um imenso apoio em horas difíceis.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Problemas encontrados em diversas áreas do conhecimento podem ter soluções com o auxílio da matemática e da computação. Neste trabalho foram abordados dois problemas. O primeiro, da área biológica, trata de testes de paternidade. O objetivo foi elaborar um modelo, e implementá-lo em software que atendesse também situações nas quais o suposto pai é desconhecido. O segundo problema aborda a contrução de sistemas de recomendação. Trabalhos encontrados na literatura tem demonstrado ineficiência das arquiteturas já existentes para construção destes sistemas. Foi proposto um modelo que minimizasse essas dificuldades encontradas. Para abordagem dos dois problemas, foi proposto a utilização das Redes Bayesianas. Estas são grafos que expressam variáveis e distribuições de probabilidade e possui capacidade de lidar com a incerteza.

Palavras-Chave: Redes Bayesianas; Sistemas de Recomendação; Teste de Paternidade.

Abstract

Problems found in various areas of knowledge may have solutions with the aid of mathematics and computing. In this work were addressed two problems. The first, the biological area, comes to a paternity test. The objective was to develop a model, and implement it in software to take account also situations where the supposed father is unknown. The second issue deals with the construction of recommendation systems. Papers found in the literature has demonstrated inefficiency of existing architectures for building these systems. It was proposed a model that minimizes the difficulties encountered. To approach the two problems, was offered the use of Bayesian Networks. These are graphs that express variables and probability distributions, and has capacity to deal with the uncertainty.

Keywords: Bayesian Network; Paternity Test, Recommender System.

Capítulo 1

INTRODUÇÃO

1.1 Considerações Iniciais

Problemas em diversas áreas do conhecimento são tratados com o auxílio da computação e da matemática. Neste trabalho foram abordados dois temas atuais: teste de paternidade para determinar o pai biológico e proposta de um modelo de sistemas de recomendação de informações adequadas ao perfil do usuário.

O primeiro, da área biológica, pode ser resumido em como verificar se um determinado homem é pai de uma criança. Assim, foi necessário revisar conceitos de composição do material genético, genética de populações e hereditariedade. O objetivo principal foi tratar este problema quando o suposto pai é desconhecido e não é conhecido o seu perfil genético.

No segundo problema, há propostas e análises de modelos de sistemas de recomendação para web. Atualmente, há um imenso volume de informações disponíveis na internet. Com isso, os usuários nem sempre conseguem encontrar a informação desejada em curto espaço de tempo. Sistemas de recomendação levam em conta o perfil do usuário que está buscando a informação (ou produto) para identificar sua necessidade. O objetivo foi propor um modelo de recomendação de informações.

Apesar dos temas serem bem distintos, e das áreas serem diferentes, a metodologia utilizada nos dois casos são semelhantes. Foram utilizadas as redes bayesianas para propor possíveis soluções para os problemas abordados. As redes bayesianas são modelos gráficos e são utilizadas em problemas nos quais não há todas as informações precisas. Essas situações são chamadas de incerteza.

Um sistema que possa atuar em situações de incerteza deve ser capaz de atribuir níveis de confiabilidade para todas as sentenças em sua base de conhecimento, e ainda, estabelecer relações entre as sentenças (RUSSEL; NORVIG, 2004). Redes bayesianas oferecem uma abordagem para o raciocínio probabilístico que engloba teoria dos grafos, para o estabelecimento das relações entre sentenças, e ainda, teoria da probabilidade, para atribuição de níveis de confiabilidade, contemplando as necessidades de se tratar a incerteza (VALENTIM, 2007).

Segundo (HECKERMAN, 1995), redes bayesianas, quando utilizadas em conjunto com as técnicas estatísticas, apresenta várias vantagens para realizar a análise dos dados. Primeiro, devido a representação de dependências entre todas as variáveis, e por lidar facilmente com situações em que alguns dados não estão disponibilizados. Segundo, as redes podem ser utilizadas para o aprendizado (de relações entre as variáveis) compreendendo o domínio do problema. Neste caso, há dois tipos de aprendizagem: a aprendizagem da topologia da rede e a dos parâmetros numéricos.

1.2 Motivação

1.2.1 Teste de Paternidade

O primeiro problema que é tratado neste trabalho, denominado Teste de Paternidade, possui um contexto que justifica sua abordagem.

De acordo com Estatísticas do Registro Civil, cerca de 30% das crianças nascidas no Brasil não têm pai declarado, o que freqüentemente representa um sério problema emocional, econômico e social(SILVA, 2001). É grande, portanto, a necessidade de determinar paternidade com absoluta confiabilidade em diversas situações da vida contemporânea. Esta necessidade surge, por exemplo, em casos amigáveis de confirmação de paternidade, disputas legais para fins de pensão alimentícia e herança, casos criminais envolvendo estupro, rapto, troca ou abandono de crianças e casos-médicos de diagnóstico pré-natal e aconselhamento genético (SILVA, 2001).

Teste de paternidade é a resolução de um problema que pode ser definido da seguinte forma: como afirmar se um determinado indivíduo é realmente pai de um filho que reclama ser filho dele.

O teste mais simples, envolvendo o trio: criança (reclamante), suposto pai (demandado) e a mãe, pode ser resolvido com facilidade. No entanto, frequentemente, aparecem questões a serem resolvidas relacionadas a paternidade após a morte do pai. Nestas situações o perfil do demandado

não está disponível e a metodologia básica que deve ser empregada é fazer a reconstituição do perfil genético do possível pai, a partir de seus familiares vivos.

Na investigação de paternidade, testes de vínculo genético são feitos a partir da utilização de marcadores genéticos presentes na criança, na mãe e no suposto pai. Os referidos marcadores são compostos de DNA (ácido desoxiribonucléico) encontrados no núcleo das células do corpo.

1.2.2 Sistemas de Recomendação

Em agosto de 2008, a quantidade de websites encontrados na internet é de aproximadamente 178 milhões (domínios registrados)[1]. Em meio à profusão de informações e serviços on-line estão os usuários da Internet. Cerca de um bilhão e meio de pessoas no mundo já utilizam a rede mundial de computadores[2].

À medida que cresce o volume de informações na web, aumenta a dificuldade dos usuários de encontrar a ”informação certa” no ”tempo certo” (O'DONOVAN; SMYTH, 2005). A grande diversidade de conteúdo disponível gera sobrecarga de informação ao usuário. Os sistemas de recomendação vêm sendo apontados como uma importante ferramenta para sanar essa dificuldade (O'DONOVAN; SMYTH, 2005).

Muitas vezes um indivíduo possui muito pouca ou quase nenhuma experiência pessoal para realizar escolhas entre as várias alternativas que lhe são apresentadas. (REATEGUI; CAZELLA, 2005). Os sistemas de recomendação auxiliam no aumento da capacidade e eficácia deste processo de indicação já bastante conhecida na relação social entre seres humanos (RESNICK; VARIAN, 1997). Em um sistema típico as pessoas fornecem recomendações como entradas e o sistema agrega e direciona para os indivíduos considerados potenciais interessados neste tipo de recomendação. Um dos grandes desafios deste tipo de sistema é realizar o casamento correto entre os que estão recomendando e aqueles que estão recebendo a recomendação, ou seja, definir e descobrir este relacionamento de interesses.

Para (ADOMAVICIUS; TUZHILIN, 2005), o interesse em tal área é grande porque constitui um rico problema de pesquisa, além de ter inúmeras aplicações práticas que podem ajudar os usuários

[1]Dados do Web Server Survey Archives – Netcraft. Corresponde ao número de domínios registrados que possuem conteúdo publicado. Disponível em <http://www.netcraft.com/survey/>. Acesso em 15 de setembro 2008.

[2]Dados do World Internet Usage Statistics News and World Population. Disponível em <http://www.internetworldstats.com/stats.htm>. Acesso em 15 de setembro de 2008.

a lidar com a sobrecarga de informação, uma vez que tais sistemas oferecem recomendações personalizadas de conteúdos e serviços. (REATEGUI; CAZELLA, 2005) pontuam que "os websites de comércio eletrônico são atualmente o maior foco de utilização dos sistemas de recomendação, empregando diferentes técnicas para encontrar os produtos mais adequados para seus clientes e aumentar deste modo sua lucratividade".

Introduzido em julho de 1996 o My Yahoo foi o primeiro website a utilizar os sistemas de recomendação em grandes proporções, utilizando a estratégia de customização (MANBER; PATEL; ROBISON, 2000). Hoje em dia, um grande número de websites emprega os sistemas de recomendação para levar aos usuários diferentes tipos de sugestões, como ofertas casadas ("clientes que compraram item X também compraram item Y"), itens de sua preferência, itens mais vendidos nas suas categorias favoritas (REATEGUI; CAZELLA, 2005).

1.3 Objetivos gerais do Trabalho

O objetivo do trabalho é compreender a teoria de probabilidades, incluindo as redes bayesianas e as regras geradas a partir dos axiomas de probabilidade, e aplicá-la aos problemas de teste de paternidade e sistemas de recomendação.

Além disso, o objetivo é analisar e sugerir possíveis modelos e metodologias de implementação das redes bayesianas para estes problemas.

No caso de teste de paternidade, o objetivo foi de elaborar um modelo que atenda as regras de hereditariedade, e que resulte em uma aplicação (um programa) para utilizá-lo de maneira prática.

Em relação aos sistemas de recomendação, o objetivo principal foi de compreender os conceitos envolvidos, verificar como os sistemas já existentes funcionam e propor um modelo que seja utilizado junto a recuperação de informação.

Capítulo 2

PROBABILIDADE

2.1 Axiomas de Probabilidade

Probabilidade é um tema abordado por cientistas desde o tempo antes de Cristo. No entanto um matemático do século passado, o russo Andrei Kolmogorov, formalizou a teoria a partir de seus princípios básicos expressos em (KOLMOGOROV, 1933). Os axiomas de probabilidade, que serão vistos logo abaixo, são chamados muitas vezes de axiomas de Kolmogorov.

De acordo a formalização dos axiomas em (KOLMOGOROV, 1933), o conjunto de todos os eventos possíveis em um experimento é chamado espaço amostral $S = \{A_1, A_2, A_3, ...\}$. S pode ser por exemplo o resultado de lançamentos de um dado (qual a face que estará para cima do conjunto de faces do dado).

Espaço amostral (S) é o conjunto de todos os resultados possíveis de um experimento aleatório (LACERDA; BRAGA, 1998). Probabilidades devem obedecer a três condições (ou *axiomas*):

1. Probabilidade é um número não negativo que deve estar no intervalo entre 0 e 1. Ou seja, $0 \leq P(Ai) \leq 1, \forall\, A_i \in S$. Considera-se A_i um evento e a probabilidade da variável aleatória a, assumir o valor de A_i, um valor maior ou igual a 0 e menor ou igual a 1.

   Exemplo: No lançamento de uma moeda, os eventos são $A_1 = cara$, e $A_2 = coroa$. Assim a probabilidade de a (a é o lançamento feito por alguém) ser igual a A_1 é igual a probabilidade de no lançamento obter-se cara, e esta probabilidade está entre 0 e 1 ($0 \leq P(A_1) \leq 1$).

2. $P(S) = 1$. A probabilidade de ocorrência de algum evento é um, pois S é a união de todos os eventos, e com certeza um deles ocorre.

3. A probabilidade da união de eventos disjuntos é a soma das probabilidades dos eventos, ou $P(\cup i\, A_i) = \sum_i (P(A_i))$.

Da definição, seguem algumas propriedades importantes vistas em (NAKANO, 2007). Estão expressas abaixo:

1. Dado um conjunto de eventos, B = $\cup A_i$, com probabilidade $P(B)$, a probabilidade do conjunto complementar, ou seja, o conjunto de todos os elementos que estão em S mas não estão em B, B^c (B complementar) vale $P(B^c) = 1 - P(B)$.

2. Se $B \subset C$ então $P(B) \leq P(C)$. Em especial, se $C = S$, C é o conjunto de todos os eventos possíveis, $P(C) = 1$.

3. Para dois conjuntos de eventos B e C quaisquer $P(B \cup C) = P(B) + P(C) - P(B \cap C)$.

2.2 Variáveis Aleatórias

A definição em (RUSSEL; NORVIG, 2004) de variável aleatória diz que ela pode ser imaginada como algo que se refere a uma "parte" do mundo cujo "status" é inicialmente desconhecido. Por exemplo, *Carie* poderia se referir a um determinado dente de siso inferior esquerdo ter uma cárie.

Uma variável aleatória pode ser considerada uma função que mapeia todos os elementos do espaço amostral (coisas) nos pontos da linha real (números) ou alguma parte dela (LACERDA; BRAGA, 1998) citado em (VALENTIM, 2007). (RUSSEL; NORVIG, 2004) divide as variáveis aleatórias em algumas espécies de acordo com seu domínio:

- Variáveis aleatórias booleanas: quando podem assumir apenas os valores de $< verdadeiro, falso >$. Por exemplo, o domínio de *Carie* poderia assumir tanto o valor verdadeiro (se o dente contém cárie) ou falso caso contrário.

- Variáveis aleatórias discretas: que incluem variáveis aleatórias booleanas como um caso especial, admitem valores de um domínio enumerável. Por exemplo o domínio lançamento

de dado poderia ser $< 1,2,3,4,5,6 >$. Ou ainda, o lançamento de uma moeda poderia ser $< cara, coroa >$. Os valores no domínio devem ser mutuamente exclusivos e exaustivos.

- Variáveis aleatórias contínuas: que assumem valores a partir dos números reais. O domínio pode ser a linha real inteira ou algum subconjunto como o intervalo $[0,1]$. Um exemplo de um valor para uma variável dessa poderia ser $X = 4,02$. As proposições relativas a variáveis aleatórias contínuas também podem ser desigualdades, como $X \leq 8,71$.

É natural tratar as variáveis aleatórias com letras maiúsculas, e seus valores em minúsculos. Será utilizado este padrão para o trabalho.

2.3 Probabilidade a *priori* ou incondicional

A probabilidade a *priori* associada a uma proposição a é o grau de crença acordado para a proposição na ausência de quaisquer outras informações; ela é representada por $P(a)$. Por exemplo, se a probabilidade a *priori* de se ter cárie em um determinado indivíduo é de $0,1$, então a representação dessa seria:

$P(Carie = verdadeiro) = 0,1$ ou $P(carie) = 0,1$

É importante lembrar que $P(a)$ deve ser utilizada apenas na ausência de outras informações. Ou seja, assim que algumas informações novas serem conhecidas, deve-se raciocionar com a probabilidade condicional de a, dadas as novas informações.

Outro exemplo é a exploração de uma região, em busca de pedras preciosas. Sabe-se que elas estão em cerca de 20% de todo o terreno. Logo a probabilidade a *priori* de encontrar as pedras preciosas no início da exploração é de $0,2$. Pode ser expressa por $P(EncontrarPedras = 0,2)$. Supondo que fosse explorado 40% do terreno sem encontrar pedras, a nova probabilidade de encontrá-las é de:

$\frac{20}{60} = \frac{1}{3}$ ou $33,33\%$

Esta é uma nova probabilidade dado o evento que ocorreu. Esta nova probabilidade é dita probabilidade condicional a um determinado evento.

2.4 Distribuição conjunta de probabilidade e distribuição de probabilidade conjunta total

Conforme em (RUSSEL; NORVIG, 2004), dada uma variável aleatória, por exemplo o tempo, esta pode assumir alguns estados, como por exemplo: $ensolarado, chuvoso, nublado$ e $nevoento$.

As probabilidades a priori para essa variável poderiam ser:

$P(Tempo = ensolarado) = 0,7$
$P(Tempo = chuvoso) = 0,2$
$P(Tempo = nublado) = 0,08$
$P(Tempo = nevoento) = 0,02$

Poderia ser expresso por:
$P(Tempo) =< 0,7;\ 0,2;\ 0,08;\ 0,02 >$

Esta declaração define a distribuição de probabilidade a *priori* para a variável $Tempo$. Considerando $Carie$ uma variável aleatória, pode-se utilizar também a denotação $P(Carie, Tempo)$ para denotar as probabilidades de todas as combinações de valores de um conjunto de variáveis aleatórias. $P(Carie, Tempo)$ pode ser representada por uma tabela 4 x 2 de probabilidades, e isto poderia ser chamado de distribuição de probabilidade conjunta de $Tempo$ e $Carie$.

As vezes é necessário expressar um conjunto completo de variáveis aleatórias usadas para descrever o mundo. Uma distribuição de probabilidade conjunta que abrange esse conjunto completo é chamada distribuição de probabilidade conjunta total (RUSSEL; NORVIG, 2004).

Por exemplo, se o mundo consistir apenas nas variáveis $Carie$, $DorDeDente$, e $Tempo$, a distribuição conjunta total será dada por:

$P(Carie, DorDeDente, Tempo)$.

Essa distribuição conjunta seria uma tabela de 2 x 2 x 4 com 16 entradas. Uma distribuição conjunta total especifica a probabilidade de todo evento atômico e é, portanto, uma especificação completa da incerteza sobre o mundo em questão.

No caso de variáveis contínuas, não é possível representar a distribuição inteira como uma tabela, porque existem infinitamente muitos valores. Em vez disso, em geral se define a probabilidade de uma

variável aleatória assumir algum valor x como uma função parametrizada de x. Por exemplo, seja a variável aleatória X que denota a temperatura mínima de uma determinada cidade. Então a sentença:

$P(X = x) = \cup[18,26](x)$

expressa a crença de que X está distribuída uniformemente entre 18 e 26 graus Celsius. As distribuições de probabilidade para variáveis contínuas são chamadas de *funções de densidade de probabilidade* (RUSSEL; NORVIG, 2004).

2.5 Probabilidade condicional

Uma vez que é obtida uma evidência relativa as variáveis aleatórias anteriormente deconhecidas que constituem o domínio, as probabilidades a *priori* não são mais aplicáveis. Em vez disso é utilizado probabilidades condicionais ou *posteriori* (RUSSEL; NORVIG, 2004).

A notação utilizada para probabilidade condicional é $P(a|b)$, e deve ser lida: "probabilidade de a dado que ocorreu b".

Exempo de probabilidade condicional:

$P(Carie|DorDeDente) = 0,8$.

Nem todas as pessoas que tem dor de dente, possuem cárie. Assim, a sentença acima diz que se a pessoa está com dor de dente, é possível afirmar que a chance de ela ter cárie é de 80%.

Em (VALENTIM, 2007) há um exemplo de probabilidade condicional, expresso pela probabilidade de um indíviduo analisado ter câncer, considerando se ele é ou não fumante:

Para todo valor de x (*presente* ou *ausente*) de uma variável aleatória *Cancer*, $P(Cancer = x|Fumante = y) = P(X|y)$ significa a distribuição de probabilidade de *Cancer* condicionada a $Y = y$ (y pode ser *sim* ou *nao*), onde as demais informações conhecidas são irrelevantes para X. Assim, se houvesse uma probabilidade $P(Cancer|Fumante = nao) = 0,45$, isto informaria que dado que uma pessoa não é fumante, a chance de ela ter a doença é de 45%.

Em (RUSSEL; NORVIG, 2004) as probabilidades condicionais podem ser definidas em termos de probabilidades incondicionais como na equação 2.1 que é válida sempre para $P(b) > 0$.

$$P(a|b) = \frac{P(a \cap b)}{P(b)} \tag{2.1}$$

Esta equação também pode ser escrita como na equação 2.2 que denomina-se *regra do produto*. A regra do produto vem do fato de que, para a e b serem verdadeiros, é necessário que b seja verdadeiro, e que a seja também verdadeiro, dado b.

$$P(a \cap b) = P(a|b)P(b) \tag{2.2}$$

Isto também pode ser visto no contrário: $P(a \cap b) = P(b|a)P(a)$.

Conforme (RUSSEL; NORVIG, 2004), a notação $P(X|Y)$ fornece os valores de $P(X = x_i|Y = y_i)$ para cada i e j possível. Para exemplificar como isso torna a notação mais concisa, é considerado a aplicação da regra do produto a cada caso em que as proposições a e b afirmam valores específicos de X e Y, respectivamente. Obtemos as equações a seguir:

$P(X = x_1 \cap Y = y_1) = P(X = x_1|Y = y_1)P(Y = y_1)$.
$P(X = x_1 \cap Y = y_2) = P(X = x_1|Y = y_2)P(Y = y_2)$.
$P(X = x_1 \cap Y = y_3) = P(X = x_1|Y = y_3)P(Y = y_3)$.
...

Assim é combinado tudo isso na equação 2.3 (VALENTIM, 2007) afirma que a regra do produto pode ser generalizada a fim de se obter a fórmula da regra da cadeia, que corresponde ao que é visto na equação 2.4 com o produtório indo de 1 até n e assumindo que $X = X_1, X_2, ..., X_n$ é um conjunto de variáveis aleatórias, esta equação fornece a distribuição conjunta de X.

$$P(X,Y) = P(X|Y)P(Y) \tag{2.3}$$

$$P(X_1, X_2, ..., X_n) = \prod_i^n P(X_i | X_i - 1, ..., X_1) \tag{2.4}$$

Em (RUSSEL; NORVIG, 2004) são feitas algumas observações importantes sobre probabilidades condicionais. Dada a probabilidade condicional $P(a|b) = 0,8$, não pode ser interpretado que "sempre que b for válida, $P(a) = 0,8$". Tal afirmação tem 2 erros: Primeiro, que $P(a)$ sempre denota a probabilidade a *priori* e não a probabilidade posteriori dado alguma evidência. E segundo, a declaração $P(a|b) = 0,8$ é imeditamente relevante apenas quando b é a única evidência disponível. Quando uma informação adicional c está disponível, o grau de crença em a é $P(a|b \cap c)$ pode ter pouca relação com $P(a|b)$. Por exemplo c poderia nos informar diretamente se a é *verdadeira* ou $falsa$.

2.6 Probabilidade Marginal

Uma tarefa particularmente comum é extrair a distribuição sobre algum subconjunto de variáveis ou sobre uma única variável.

A tabela 2.1 consiste em um domínio que contém apenas 3 variáveis booleanas: *DorDeDente*, *Carie* e *Boticao*. A distribuição conjunta total é expressa pela tabela que tem dimensões 2 x 2 x 2 (RUSSEL; NORVIG, 2004). Nota-se que a soma das distribuições tem valor igual a 1.

Tabela 2.1: Distribuição conjunta total para o mundo de DorDeDente, Cárie e Boticão

	dorDeDente		*¬DorDeDente*	
	Boticao	*¬Boticao*	*Boticao*	*¬Boticao*
carie	0,108	0,012	0,072	0,08
¬carie	0,016	0,064	0,144	0,576

Um exemplo simples de consulta a esses dados, é a adição das entradas da primeira linha da tabela 2.1 que produz a probabilidade incondicional ou probabilidade marginal de *carie*:

$P(carie) = 0,108 + 0,012 + 0,072 + 0,08 = 0,2.$

Este processo é chamado de Marginalização porque as variáveis, com exceção de *Carie*, são totalizadas. Pode-se escrever a regra geral da marginalização como segue abaixo para qualquer conjunto de variáveis Y e Z (RUSSEL; NORVIG, 2004).

$$P(Y) = \sum_z P(Y,z)$$

De acordo com (RUSSEL; NORVIG, 2004) poderia ser calculado alguma probabilidade condicional, utilizando a equação 2.1, a partir da distribuição conjunta total. Por exemplo, obter a probabilidade de uma cárie dado a evidência de uma dor de dente:

$$P(carie|DorDeDente) = \frac{P(carie \cap DorDeDente)}{P(DorDeDente)} = \frac{0{,}108+0{,}012}{0{,}108+0{,}012+0{,}016+0{,}064} = 0,6$$

Para conferência poderia ser calculado a probabilidade de não haver cárie, dado uma dor de dente:

$$P(\neg\, carie|DorDeDente) = \frac{P(\neg\, carie \cap DorDeDente)}{P(DorDeDente)} = \frac{0{,}016+0{,}064}{0{,}108+0{,}012+0{,}016+0{,}064} = 0,4$$

(VALENTIM, 2007) mostra um exemplo envolvendo apenas duas variáveis booleanas, que pode ser apresentado montando uma tabela que relaciona as probabilidades de uma pessoa ter Câncer(variável aleatória *Cancer*) condicionada à evidência de uma pessoa fumar ou não (variável aleatória $Fumante$) (ver tabela 2.2).

Tabela 2.2: Distribuição conjunta de *Cancer* e $Fumante$

	Cancer		Marginal de $Fumante$
$Fumante$	*presente*	*ausente*	
sim	0,3	0,15	0,45
nao	0,1	0,45	0,55
Marginal de *Cancer*	0,4	0,6	1,00

Pode-se observar na Tabela 2.2 as probabilidades conjuntas:

$P(Cancer = presente, Fumante = sim) = 0,30,$

$P(Cancer = presente, Fumante = nao) = 0,10,$

$P(Cancer = ausente, Fumante = sim) = 0,15,$

$P(Cancer = ausente, Fumante = nao) = 0,45.$

A distribuição marginal da variável $Fumante$ é dada pela última coluna da Tabela 2.2, enquanto que a distribuição marginal de *Cancer* é fornecida pela última linha. Portanto, em relação a *Cancer* tem-se:

$P(Cancer = presente) = 0,40$ e $P(Cancer = ausente) = 0,60$

enquanto que em relação a Fumante tem-se:

$P(Fumante = sim) = 0,45$ e $P(Fumante = nao) = 0,55$.

Pode-se notar mais uma vez, que o somatório das probabilidades conjuntas, bem como o somatório das probabilidades marginais de cada variável, resulta em 1.

2.7 Tabela de Probabilidade Condicional

(VALENTIM, 2007) demonstra como pode-se montar uma tabela de probabilidade conjunta, ou TPC, formada pelas probabilidades condicionais da variável *Cancer*:

$$P(Cancer = presente, Fumante = sim) = \frac{P(Cancer=presente|Fumante=sim)}{P(Fumante=sim)} = \frac{0,3}{0,45} = 0,667$$

$$P(Cancer = ausente, Fumante = sim) = \frac{P(Cancer=ausente|Fumante=sim)}{P(Fumante=sim)} = \frac{0,15}{0,45} = 0,333$$

$$P(Cancer = presente, Fumante = nao) = \frac{P(Cancer=presente|Fumante=nao)}{P(Fumante=nao)} = \frac{0,10}{0,55} = 0,1818$$

$$P(Cancer = ausente, Fumante = nao) = \frac{P(Cancer=ausente|Fumante=nao)}{P(Fumante=nao)} = \frac{0,45}{0,55} = 0,8182$$

A Tabela 2.3 representa a TPC de *Cancer* que é condicionada à variável aleatória $Fumante$ expressa em (VALENTIM, 2007):

Tabela 2.3: Tabela de Probabilidade Conjunta Total de *Cancer*

	Cancer	
Fumante	*presente*	*ausente*
sim	0,667	0,333
nao	0,1818	0,8182

2.8 Independência Condicional

Conforme (VALENTIM, 2007), diz-se que duas variáveis X e Y são independentes se:

$$P(x|y) = P(x)$$

sempre que $P(y) > 0, \forall x \in D_x$ e $y \in D_y$, onde D_x e D_y denotam os domínios de X e Y, respectivamente. Se X e Y são independentes, então Y não é informativa para X. Significa que conhecer Y não altera a probabilidade de X.

Dados dois conjuntos disjuntos de variáveis, V_i e V_j, e uma variável V, diz-se que a variável V é condicionalmente independente do conjunto V_i, dado V_j, se: $P(V|V_i,V_j) = P(V|V_j)$

Neste caso, a notação $I(V,V_i|V_j)$ é empregada para indicar este fato. A interpretação para a independência condicional é que se $I(V,V_i|V_j)$, então V_i não acrescenta informação relevante para V, quando já se dispõe de V_j (VALENTIM, 2007).

Agora, considere um caso com duas variáveis aleatórias, X e Y, e seus respectivos domínios, D_X e D_Y. Se $P(X|Y) = P(X)$ sempre que $P(y) > 0, \forall X \in D_X$ e $Y \in D_Y$, então diz-se que X e Y são condicionalmente independentes. É possível representar esta independência através da equação $P(X,Y) = P(X)P(Y)$, que origina-se da equação 2.1.

(RUSSEL; NORVIG, 2004) demonstram de maneira prática o conceito de independência condicional conforme a seguir. Adiciona-se uma variável ao domínio exposto na seção *Probabilidade Marginal* pela tabela 2.1. Esta variável é denominada *Tempo*.

Logo a distribuição conjunta total se torna $P(DorDeDente, Boticao, Carie, Tempo)$ que tem 32 entradas porque tempo tem quatro valores $(nublado, nevoento, chuvoso$ e $ensolarado)$.

Uma consulta poderia ser a que consta abaixo:

$P(dordedente, boticao, carie, Tempo = Nublado) =$
$P(Tempo = nublado|dordedente, boticao, carie)P(dordedente, boticao, carie)$

Considerando a independência condicional tem-se que:

$P(Tempo = nublado|dordedente, boticao, carie) = P(Tempo = dordedente)$

Assim deduz-se que:

$P(DorDeDente, Boticao, Carie, Tempo) =$
$P(DorDeDente, Boticao, Carie)P(Tempo)$

Desse modo, a tabela de 32 elementos (entradas) para 4 variáveis pode ser construída a partir de uma tabela de 8 elementos e uma de 4 elementos. Essa decomposição é ilustrada esquematicamente na Figura 2.1.

Formalmente (RUSSEL; NORVIG, 2004) expressam então a relação de independência através das equações sobre as variáveis aleatórias X e Y abaixo (são equivalentes).

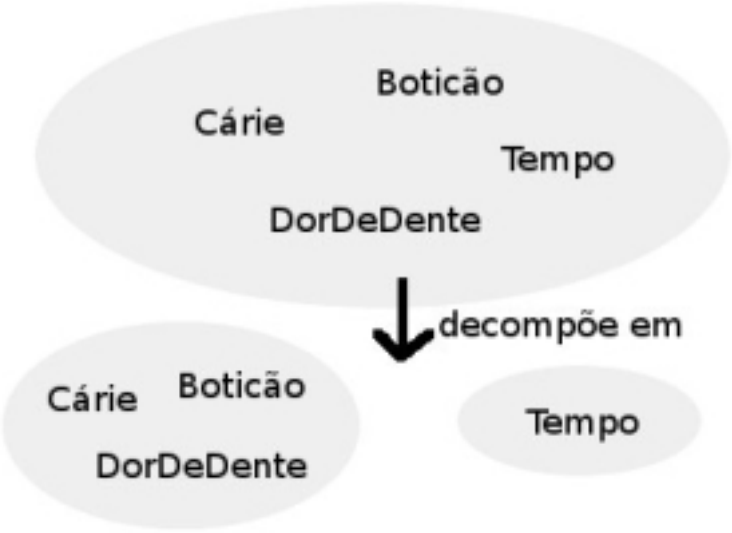

Figura 2.1: Independência entre as condições de tempo e os problemas dentários. Fonte: (RUSSEL; NORVIG, 2004)

$$P(X|Y) = P(X) \; ou \; P(Y|X) = P(Y) \; ou \; P(X \cap Y) = P(X) * P(Y) \qquad (2.5)$$

Capítulo 3

RACIOCÍNIO SOBRE INCERTEZA E INFERÊNCIA PROBABILÍSTICA

A rede bayesiana é baseada no raciocínio sobre incerteza, pois ela representa graficamente um conjunto de variáveis e também a distribuição de probabilidade entre elas, expressando também o grau de crença de uma variável assumir determinado valor.

É importante observar que um grau de crença é diferente de um grau de verdade. Uma probabilidade de 0,8 não significa "80%" verdadeira, mas sim um grau de crença igual a 80%, isto é, uma expectativa forte (RUSSEL; NORVIG, 2004). Grau de verdade, em oposição ao grau de crença, é abordado com a lógica fuzzy em (KLIR; YUAN, 1995).

3.1 Redes bayesianas

As redes bayesianas tem sido utilizadas em diversas áreas, como em reconhecimento de spam (REAL, 2003), reconhecimento de voz (ZWEIG, 1998), sistema de exploração e aquisição de conhecimento espacial (MORRIS, 2003), diagnósticos médicos (SAHEKI, 2005), robótica, entre outros.

Raciocínio bayesiano é explicado com um exemplo médico extraído de (YUDKOWSKY, 2008) citado por (VALENTIM, 2007). Seja o problema:

"1% das mulheres com mais de 40 anos que participam de exames de rotina são portadoras de câncer de mama. 80% das mulheres com câncer terão resultados positivos de mamografias. 9,6%

Tabela 3.1: Probabilidade *a priori* de *Cancer*. Fonte: (VALENTIM, 2007)

	$P(Cancer)$	
	$presente$	$ausente$
Probabilidade *a priori*	0,01	0,99

das mulheres sem a doença também terão resultado positivo nas mamografias. Uma mulher dessa idade se depara com um resultado positivo de mamografia, qual a probabilidade dela portar câncer de mama?"

Segundo (YUDKOWSKY, 2008), a maioria dos médicos estimaria que a probabilidade da mulher em questão, ter câncer de mama estaria entre 70% e 80%. Montaremos o exemplo de maneira bayesiana para chegar ao resultado correto. Em primeiro lugar, em uma mulher com mais de 40 anos o câncer de mama (*Cancer*) pode estar presente ou ausente. Essas alternativas, mutuamente excludentes, podem ser colocadas em uma tabela, como na Tabela 3.1. Podemos iniciar o raciocínio pela probabilidade de cada alternativa 'antes de fazer qualquer teste'. É a chamada probabilidade a *priori* - $Cancer = presente$ ou $Cancer = ausente$. Como 1% das mulheres com mais de 40 anos têm câncer de mama, a probabilidade a *priori* de *Cancer* estar presente é de 0,01 e de estar ausente é de 0,99.

Agora é incorporado o resultado da mamografia. Se Câncer está presente, a probabilidade condicional de Mamografia ser positiva é de 0,80 (80%), e se Câncer está ausente esta probabilidade é de 0,096 (9,6%). Pode-se reunir essas informações em uma tabela de probabilidade condicional (TPC) de Mamografia, como na Tabela 3.2.

Tabela 3.2: TPC de $Mamografia$ - $P(Mamografia|Cancer)$ Fonte: (VALENTIM, 2007)

	$Mamografia$	
$Cancer$	$positiva$	$negativa$
$presente$	0,8	0,2
$ausente$	0,096	0,904

Conforme (VALENTIM, 2007), multiplica-se a probabilidade a priori pela condicional e obtém-se a probabilidade conjunta de Câncer e Mamografia, conforme 3.3.

Para fazer com que a soma de cada linha da probabilidade conjunta se torne 1, é preciso usar uma normalização: multiplicando cada probabilidade pela constante de normalização, que é dada por 1

Tabela 3.3: Probabilidade conjunta de *Cancer* e $Mamografia$ - $P(Mamografia, Cancer)$. Fonte: (VALENTIM, 2007)

	$Mamografia$	
Cancer	*positiva*	*negativa*
presente	$0,01 * 0,8 = 0,008$	$0,01 * 0,2 = 0,002$
ausente	$0,99 * 0,096 = 0,09504$	$0,99 * 0,904 = 0,89496$

dividido pelo somatório de cada linha da tabela de probabilidade conjunta. Obtendo assim a chamada probabilidade a posteriori, mostrada na Tabela 3.4.

Portanto, com o raciocínio bayesiano conclui-se que a probabilidade a posteriori após os testes, de uma mulher com mais de 40 anos, de posse de um exame de mamografia cujo resultado é positivo, ter câncer de mama é de 0,07764 (7,764%). A representação pode ser feita com $P(Cancer = presente|Mamografia = positiva) = 0,07764(7,764\%)$.

Tabela 3.4: Probabilidades a posteriori - $P(Cancer|Mamografia)$. Fonte: (VALENTIM, 2007)

	$Mamografia$	
Cancer	*positiva*	*negativa*
presente	$\frac{0,008}{(0.008+0.09504)} = 0,07764$	$0,92236$
ausente	$0,00223$	$0,99777$

(PENA, 2006) relata que quando esse problema foi apresentado a vários médicos e estudantes de medicina, observou-se uma tendência a superestimar a probabilidade a posteriori da doença, e segundo (YUDKOWSKY, 2008), apenas 46% dos entrevistados estimaram uma probabilidade condizente com a resposta correta. Isso revela que o raciocínio bayesiano não é intuitivo. Parece haver uma tendência geral de ignorar o fato de que a probabilidade a priori de doença é pequena.

No exemplo acima, o raciocínio bayesiano permitiu quantificar o grau em que o resultado positivo de mamografia ajustou uma estimativa inicial da chance de uma mulher ter câncer de mama. Sob esse ponto de vista, um teste médico (ou evidência) funciona como um 'modificador de opinião', atualizando uma hipótese inicial (probabilidade a priori) para gerar outra (probabilidade a posteriori). Essa última engloba tanto a crença anterior (probabilidade a priori) quanto o resultado do teste. A probabilidade a posteriori torna-se automaticamente a probabilidade a priori para testes subseqüentes (PENA, 2006).

A principal vantagem de um raciocínio probabilístico se comparado com um agente lógico, é que os agentes (agindo sobre um sistema de incerteza) poderão tomar decisões mesmo que não haja informações suficientes para se provar que aquela ação irá funcionar (RUSSEL; NORVIG, 2004).

Com ajuda de um especialista num certo domínio de dados, define-se um modelo de redes bayesianas: determina-se a estrutura e suas probabilidades condicionais associadas. Mas, em situações em que o especialista não está disponível, ou no caso de um grande domínio de dados em que fica difícil se especializar, são úteis métodos automáticos para aprendizado de estruturas e probabilidades (também denominados parâmetros) a partir de dados disponíveis (HECKERMAN, 1995).

Uma rede bayesiana fornece uma descrição completa do domínio. Toda entrada na distribuição de probabilidade conjunta total pode ser calculada a partir de informações armazenadas na rede. Uma entrada genérica na distribuição conjunta é a probabilidade de uma conjunção de atribuições específicas a cada variável, tal como $P(X_1 = x_1 \cap ... \cap X_n = x_n)$ (RUSSEL; NORVIG, 2004).

O valor dessa entrada é dado pela equação exposta abaixo:

$$Pr(U) = \prod_i (Pr(A_i) \mid pa(A_i)) \tag{3.1}$$

Uma rede bayesiana é um grafo, composto por um conjunto de vértices (representados graficamente por círculos) e um conjunto de arestas (ligações entre vértices). A rede bayesiana é um grafo direcionado. O que corresponde a toda aresta conter direção de um nó a outro. Dado que um nó X tem uma aresta direcionada a Y, diz-se que X é pai de Y dentro da rede (JENSEN, 2001).

Uma rede bayesiana busca representar completamente o domínio de um problema. Assim, toda entrada na distribuição de probabilidade conjunta total pode ser calculada a partir das informações armazenadas na rede. Uma entrada genérica na distribuição conjunta é a probabilidade de uma conjunção de atribuições específicas a cada variável, tal como $P(X_1 = x_1 \cap ... \cap X_n = x_n)$, (RUSSEL; NORVIG, 2004).

Ou seja, as redes bayesianas é uma forma reduzida da tabela de distribuição conjunta total, mas que fornece qualquer probabilidade (*a priori*) a ser consultada. Além de ser uma representação completa e não-redundante do domínio, uma rede bayesiana freqüentemente pode ser muito mais compacta que a distribuição conjunta total. Essa propriedade é o que torna viável manipular domínios com

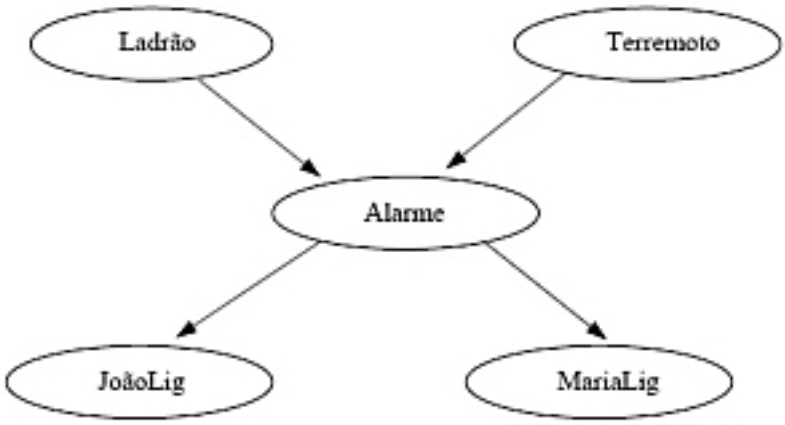

Figura 3.1: Rede bayesiana para problema do Alarme encontrado em (RUSSEL; NORVIG, 2004)

muitas variáveis. A densidade das redes bayesianas é um exemplo de uma propriedade muito geral de sistemas localmente estruturados (também chamados de sistemas esparsos). Em um sistema localmente estruturado, cada subcomponente interage diretamente apenas com um número limitado de outros componentes, não importando o número total de componentes. A estrutura local normalmente está associada ao crescimento linear, e não ao crescimento exponencial da complexidade, (RUSSEL; NORVIG, 2004).

Considere o domínio, extraído de (RUSSEL; NORVIG, 2004), como exemplo:

"Você possui um novo alarme contra ladrões em casa. Este alarme é muito confiável na detecção de ladrões, entretanto, ele também pode disparar caso ocorra um terremoto. Você tem dois vizinhos, João e Maria, os quais prometeram telefonar-lhe no trabalho caso o alarme dispare. João sempre liga quando ouve o alarme, entretanto, algumas vezes confunde o alarme com o telefone e também liga nestes casos. Maria, por outro lado, gosta de ouvir música alta e às vezes não escuta o alarme." Este domínio pode ser representado como apresenta a Figura 3.1.

Segue tabelas de probabilidades para as variáveis acima:

Tabela 3.5: Probabilidade *a priori* de Roubo

$P(Roubo)$
0,01

As tabelas 3.5, 3.6, 3.7, 3.8 e 3.9 expressam as tabelas de probabilidade condicional (TBC) para a rede bayesiana da figura 3.1.

Tabela 3.6: Probabilidade *a priori* de Terremoto

$P(Terremoto)$
0,002

Tabela 3.7: Probabilidade *a priori* de Alarme dado os estados de Roubo e Terremoto

$Roubo$	$Terremoto$	$P(Alarme)$
v	v	0,95
v	f	0,94
f	v	0,29
f	f	0,001

Para ilustrar o que foi visto na equação 3.1, pode-se calcular a probabilidade de que o alarme tenha soado, mas não tenha ocorrido nenhum roubo nem um terremoto, e que tanto João quanto Maria tenham ligado. Este exemplo está em (RUSSEL; NORVIG, 2004).

$P(j \cap m \cap a \cap (\neg b) \cap (\neg e)) = P(j|a) * P(m|a) * P(a|(\neg b) \cap (\neg e)) * P(\neg b) * P(\neg e) = 0,90 * 0,70 * 0,001 * 0,999 * 0,998 = 0,00062$

As letras j, m, a, b, e, expressam respectivamente $JoaoLiga, MariaLiga, Alarme, Roubo$ e $Terremoto$. E os valores utilizados são os contidos nas tabelas de distribuição de probabilidade.

Portanto, a distribuição conjunta total pode ser utilizada para responder a qualquer consulta sobre o domínio. Se uma rede bayesiana estiver representando a distribuição conjunta, ela poderá ser utilizada para responder as questões propostas, através das inferências probabilísticas conforme a seção Inferência em Redes Bayesianas.

Tabela 3.8: Probabilidade *a priori* de JoaoLiga dado o estado Alarme

$Alarme$	$P(JoaoLiga)$
v	0,90
f	0,05

Tabela 3.9: Probabilidade *a priori* de MariaLiga dado o estado Alarme

$Alarme$	$P(MariaLiga)$
v	0,70
f	0,01

3.2 Inferência em Redes Bayesianas

De acordo com (VALENTIM, 2007), a tarefa básica de um sistema de redes bayesianas é computar a distribuição da probabilidade condicional para um conjunto de variáveis de consulta, dado os valores de um conjunto de variáveis de evidência, ou seja, computar a $P(varivel\_consulta|variveis\_evidencia)$.

Essa tarefa é denominada inferência bayesiana e permite responder a uma série de ”consultas” sobre um domínio de dados. Por exemplo, na área médica, a principal tarefa consiste em obter um diagnóstico para um determinado paciente apresentando certos sintomas (evidências). Esta tarefa consiste em atualizar as probabilidades das variáveis em função das evidências. No caso do diagnóstico médico, tenta-se conhecer as probabilidades de cada uma das possíveis doenças, dados os sintomas observados no paciente. Essas são probabilidades a posteriori (VALENTIM, 2007).

De acordo com (LUNA, 2004), uma vez construída uma representação probabilística através do modelo de RBs, para a incerteza presente no relacionamento entre variáveis de um domínio de dados, uma das tarefas mais importantes consiste em obter estimativas de probabilidades de eventos relacionados aos dados, a medida que novas informações ou evidências sejam conhecidas. E este processo é denominado inferência bayesiana.

Segundo (CASTILLO; GUTIERREZ; HADI, 1996) há três tipos distintos de algoritmos de inferência: exatos, aproximados e simbólicos. Um algoritmo de inferência denomina-se exato se as probabilidades dos nós são calculadas sem outro erro senão o de arredondamento, inerente a limitações de cálculo dos computadores. Os algoritmos aproximados utilizam distintas técnicas de simulação para obter valores aproximados das probabilidades. Em geral, estes algoritmos são utilizados em casos em que os algoritmos exatos não são aplicáveis, ou o custo computacional é elevado. Já os algoritmos simbólicos podem operar tanto com parâmetros numéricos quanto com parâmetros simbólicos, obtendo probabilidades na forma simbólica, em função dos parâmetros.

Abaixo segue as características, obtidas em (VALENTIM, 2007)(RUSSEL; NORVIG, 2004), de alguns algoritmos de inferência.

3.2.1 Algoritmos exatos

Algoritmo de Pearl (aplicável em *Poliarvores*)

- Aplicável apenas a redes sem ciclos não direcionados (*polytrees*).
- Justificativa complicada para as equações usadas no algoritmo.
- Algoritmo linear e simples, com equações recursivas.
- Usa programação dinâmica, para evitar calcular várias vezes um mesmo fator.

Algoritmo de Enumeração

- Complexidade de espaço do algoritmo é linear em relação ao número de variáveis.
- Complexidade de tempo para uma rede com n variáveis booleanas é $O(2^n)$.
- Permite responder a qualquer consulta $P(X|e)$ a partir da distribuição conjunta total da rede bayesiana já que corresponde a avaliar a equação de consulta a Tabela de Probabilidade Conjunta Total.

3.2.2 Algoritmos Aproximados

Os denominados algoritmos aproximados são utilizados quando a inferência exata consome um tempo inviável para execução em uma rede.

Forward Sampling

- Gera muitos casos que são descartados (prejuízo de processamento).
- Se for um problema de muitas evidências, terá um número expressivo de configurações descartadas.
- A fração de amostras consistentes com a evidência cai exponencialmente, conforme o número de variáveis de evidência cresce.
- O desvio-padrão do erro em cada probabilidade será proporcional a $\frac{1}{\sqrt{n}}$ onde n é o número de amostras usadas na estimativa.

Likelihood Sampling

- Fácil implementação.
- Rápido tempo de convergência comparado com o algoritmo Forward Sampling.
- Utiliza todas as amostras geradas (não perde o processamento)

Capítulo 4

SISTEMAS DE RECOMENDAÇÃO

(BURKE, 2002) define sistemas de recomendação como qualquer sistema que produza recomendações individualizadas como saída, ou que tenha o efeito de guiar o usuário de forma personalizada a objetos interessantes e úteis, diante de uma grande variedade de opções.

Os proponentes do primeiro sistema de recomendação denominado Tapestry (GOLDBERG *et al.*, 1992)(RESNICK; VARIAN, 1997), criaram a expressão ”filtragem colaborativa”, visando designar um tipo de sistema específico no qual a filtragem de informação era realizada com o auxilio humano, ou seja, pela colaboração entre os grupos de interessados. Os autores preferem utilizar a expressão sistemas de recomendação, por ser um termo genérico e defendem este posicionamento por dois motivos: primeiro porque os recomendadores podem não explicitar colaboração com os que as recebem, pois um pode não conhecer o outro, e por último os recomendadores podem sugerir itens de interesse particular, incluindo aqueles que poderiam ser desconsiderados (REATEGUI; CAZELLA, 2005).

(BURKE, 2002) afirma que os sistemas de recomendação são formados por:

- *background data*: os dados prévios armazenados, corresponde as informações que o sistema utiliza antes do processo de recomendação,
- *Input data*: informações que o usuário deve comunicar com o sistema para gerar a recomendação,
- *Algorithm*: o terceiro componente do sistema de recomendação é um algoritmo que estabelece a relação entre os outros dois. Trata-se de um algoritmo que combina os dados prévios

(*background data*) e as informações de entrada do usuário (*Input data*). É o processamento das informações para gerar recomendações adequadas.

Os dados prévios compreendem o conjunto de itens que poderão ser recomendados, tais como listas de produtos, de documentos, de filmes, de páginas web etc.

4.1 Classificação dos sistemas de recomendação

No que diz respeito a classificação dos sistemas de recomendação, são discutidos as fontes de dados do sistema (dados prévios e de entrada) e a maneira como esses dados são utilizados para fornecer as recomendações (BURKE, 2002). Os sistemas podem ser classificados nas seguintes categorias (ADOMAVICIUS; TUZHILIN, 2005)(BURKE, 2002):

- recomendação baseada em conteúdo ou filtragem baseada em conteúdo;
- recomendação colaborativa ou filtragem colaborativa;
- recomendação demográfica;
- recomendação baseada em utilidade;
- recomendação baseada em conhecimento;
- abordagem híbrida ou filtragem híbrida.

A tabela 4.1, obtida em (BURKE, 2002) compara as várias formas de recomendações existentes. Na tabela, há os seguintes elementos:

- I: corresponde aos itens nos quais as recomendações podem ser feitas.
- U: corresponde ao conjunto de usuários, nos quais suas preferências são conhecidas.
- u: o usuário que necessita de uma recomendação a ser gerada.
- i: um item no qual nós gostaríamos de predizer para as preferências do usuário u.

Tabela 4.1: Comparação das várias formas de sistemas de recomendações

Técnica	Dados prévios (*Background data*)	Entrada de informação (*Input data*)	Processo (*Algorithm*)
Colaboração (*Collaborative*)	Avaliações dos usuários U em relação aos itens I	Avaliação do usuário u em relação ao conjunto de itens I.	Identificar usuários similares ao usuário u no conjunto U, e verificar as preferências destes usuários similares em relação a um item i
Baseada em Conteúdo (*Content-based*)	Características dos itens I	Avaliações do usuário u, em relação aos itens do conjunto I	Gera uma classificação do usuário u, em relação ao seu comportamento, e utiliza-a para recomendar algum i
Demográfica (*Demographic*)	Informações demográficas sobre o conjunto de usuários U, e suas avaliações sobre os itens I	Informação demográfica do usuário u	Identificar usuários que são similares ao usuário u, nos aspectos demográficos. E utilizar destes usuários suas avaliações de um item i
Baseada em utilidade (*Utility-based*)	Características dos itens em I	A função utilidade sobre os itens em I, que descreve as preferências do usuário u	Aplica a função utilidade para os itens I e determina um ranking entre eles
Baseada em conhecimento (*Knowledge-based KB*)	Características dos itens do conjunto I. Conhecimento de como estes itens podem atender as necessidades dos usuários	Uma descrição das necessidades ou dos interesses de um usuário u	Inferir a relação entre um item i e as necessidades do usuário u

4.1.1 Recomendação baseada em conteúdo

Os sistemas de recomendação baseados em conteúdo têm origem nas técnicas de recuperação de informação (BALABANOVIC; SHOHAM, 1997) e nas pesquisas de filtragem de informação (ADOMAVICIUS; TUZHILIN, 2005)(BURKE, 2002).

Em um sistema de recomendação baseado em conteúdo, cada item de um conjunto I é definido por características associadas a ele (BURKE, 2002). Por exemplo, um texto pode conter palavras chaves que pode ser considerado suas características. Já um filme pode ter como características: título, diretor, elenco, gênero etc. É com base nessas características que os itens podem ser comparados e a semelhança entre eles estabelecida.

Conhecendo o histórico de preferência do usuário por itens do conjunto I, e a semelhança entre esses itens, o sistema de recomendação baseado em conteúdo é capaz de recomendar para este usuário outros itens que possam ser de seu interesse (REATEGUI; CAZELLA, 2005).

De acordo com (ADOMAVICIUS; TUZHILIN, 2005), em virtude dos significativos avanços na área de recuperação de informação e filtragem, e dada à importância das inúmeras aplicações baseadas em texto, muitos dos sistemas baseados em conteúdo focaram na recomendação de itens que contém informação textual, como documentos, websites e notícias. Para esses autores, a melhoria trazida pelos sistemas baseados em conteúdo, em relação à tradicional recuperação de informação, vem da utilização de perfis de usuários que contém dados sobre seus gostos, preferências e necessidades.

De maneira geral, os sistemas baseados em conteúdo mantêm perfis de longo prazo: o modelo de usuário se baseia nas características dos itens avaliados por ele, sendo atualizado à medida que outros itens vão sendo avaliados (BURKE, 2002).

De acordo com (ADOMAVICIUS; TUZHILIN, 2005)(BALABANOVIC; SHOHAM, 1997)(BURKE, 2002) os sistemas baseados em conteúdo possuem as seguintes limitações:

- *Análise de conteúdo limitada*: As técnicas baseadas em conteúdo são limitadas pelas características que explicitamente devem estar associadas aos itens que serão recomendados. Isso significa que as características do conteúdo devem estar estruturadas de forma que possam ser analisadas automaticamente pelo computador. A extração automática de características é de difícil aplicação a alguns domínios - como dados multimídia (por exemplo, vídeo e som) - e atribuí-las manualmente pode ser inviável, em alguns casos, devido a limitações de recursos;

- *Superespecialização*: Em virtude de o sistema se basear nos itens já avaliados pelo usuário para fazer suas recomendações, ele acaba limitando a recomendação a itens de grande semelhança àqueles já conhecidos pelo usuário. Exemplificando: uma pessoa que nunca experimentou comida árabe jamais receberia uma recomendação para ir ao melhor restaurante de comida árabe da cidade. A diversidade de recomendações é, freqüentemente, uma característica desejável em sistemas de recomendação. Outra conseqüência da superespecialização é que, em alguns casos, certos itens não deveriam ser recomendados justamente por serem muito semelhantes aos já avaliados pelo usuário, como notícias que descrevem o mesmo fato. O sistema DailyLearner (BILLSUS; PAZZANI, 2000) por exemplo, emprega um limite de similaridade para excluir as notícias muito semelhantes às já lidas pelo usuário;

- O problema do *Novo usuário*: Para que o sistema de recomendação baseado em conteúdo possa realmente compreender as preferências e apresentar recomendações confiáveis, o usuário tem que avaliar um número suficiente de itens. Um novo usuário, que avaliou poucos itens, conseqüentemente, receberá recomendações pouco precisas.

- *Elasticidade versus plasticidade*: Ao contrário do problema do novo usuário, o problema conhecido como "estabilidade versus plasticidade" (BURKE, 2002) surge quando uma quantidade substancial de avaliações é feita e o perfil do usuário é consolidado. Com isso, torna-se difícil alterar suas preferências. Caso um pesquisador resolva atuar em uma nova área da ciência, ele continuará recebendo recomendações de sua área de pesquisa anterior por um bom tempo, até que suas novas avaliações sejam suficientes para alterar seu perfil. A fim de contornar esse problema, alguns sistemas procuram incorporar algum tipo de variável temporal que aumente a relevância das avaliações mais recentes. Em contrapartida, assumem o risco de perder informações sobre interesses de longo prazo que são manifestados esporadicamente (BURKE, 2002).

4.1.2 Recomendação colaborativa

Segundo (BURKE, 2002), a recomendação colaborativa é provavelmente a mais familiar, a mais utilizada e a que apresenta tecnologias mais consolidadas. Os sistemas de recomendação baseados em filtragem colaborativa procuram predizer a utilidade de um item para um usuário particular com base em itens previamente avaliados por outros usuários (ADOMAVICIUS; TUZHILIN, 2005). Diferencia-se

Figura 4.1: Sistema movielens <www.movielens.org>. Avaliação de um filme que o usuário assistiu.

da filtragem baseada em conteúdo exatamente por não exigir a compreensão ou reconhecimento das características dos itens.

Nos sistemas colaborativos, a essência está na troca de experiências entre as pessoas que possuem interesses comuns (REATEGUI; CAZELLA, 2005). Um perfil típico de usuário em um sistema colaborativo consiste em um vetor de itens e suas avaliações, que aumenta continuamente à medida que o usuário interage com o sistema (BURKE, 2002).

Um exemplo de ambiente baseado em filtragem colaborativa é o sistema de recomendação de filmes MovieLens (GOOD, 1999). Nele o usuário atribuiu nota aos filmes que já assistiu e o sistema utiliza essas avaliações para encontrar pessoas com gostos similares e fazer suas predições. A figura 4.1 exibe uma avaliação do usuário a um filme.

Os sistemas colaborativos puros resolvem as principais deficiências dos sistemas baseados em conteúdo. Usando recomendações de outros usuários é possível tratar qualquer tipo de conteúdo e recomendar itens, mesmo que esses não sejam semelhantes aos itens já avaliados pelo usuário. No entanto, os sistemas colaborativos têm suas próprias limitações (ADOMAVICIUS; TUZHILIN, 2005)(BALABANOVIC; SHOHAM, 1997)(BURKE, 2002):

- *O problema do novo item*: Novos itens são regularmente adicionados aos sistemas de recomendação. Como os sistemas colaborativos dependem unicamente da avaliação dos usuários para que possam fazer recomendações, enquanto um novo item não for avaliado por um número considerável de usuários, o sistema não o recomendará.

- *Avaliações esparsas*: Caso o número de usuários seja pequeno em relação ao volume de informações no sistema, existe um grande risco das avaliações tornarem-se muito esparsas, devido ao pequeno número de avaliações, comparado ao necessário, para se fazer recomendações. Em um sistema de recomendação de filmes, só para exemplificar, aqueles que forem avaliados por poucas pessoas são raramente recomendados, mesmo que essas atribuam pontuações elevadas.

- *Similaridade (usuário incomum)*: Um usuário cujo gosto seja incomum, comparado à maioria da população, terá dificuldades para encontrar usuários com gostos semelhantes ao seu, e, por isso, suas recomendações podem se tornar pobres (REATEGUI; CAZELLA, 2005). Esse problema é também encontrado na literatura com o nome de “gray sheep”, ou “ovelha negra” (CLAYPOOL *et al.*, 1999). Para (ADOMAVICIUS; TUZHILIN, 2005), esse problema é considerado um problema de esparsidade.

(BURKE, 2002) afirma que um sistema colaborativo trabalha melhor quando o usuário pertence a um grupo com muitos ”vizinhos” de gostos similares ao seu. Assim como os sistemas baseados em conteúdo, os sistemas colaborativos apresentam problemas de ”novo usuário” e de ”elasticidade versus plasticidade”. Na literatura, os problemas de ”novo usuário” e ”novo item”, provocados pela escassez de dados no sistema, são também conhecidos como ”problemas de $ramp-up$” (KONSTAN, 1998)(BURKE, 2002).

4.1.3 Recomendação demográfica

Os sistemas de recomendação demográfica têm como objetivo classificar o usuário em função de seus atributos pessoais (idade, sexo, naturalidade, renda etc). Realizam a recomendação com base em classes demográficas (BURKE, 2002).

Para (ADOMAVICIUS; TUZHILIN, 2005), a abordagem demográfica é uma extensão dos sistemas colaborativos, uma vez que ambos se baseiam na similaridade entre perfis de usuários. A diferença está na forma como o perfil do usuário é construído: enquanto a filtragem demográfica

calcula a similaridade com base nos dados demográficos do usuário, a filtragem colaborativa utiliza o histórico de avaliações. Conseqüentemente, a recomendação demográfica apresenta as principais desvantagens encontradas no método colaborativo. Contudo, como não depende das avaliações para comparar usuários, sofre menos com o problema de *esparsidade*.

(REATEGUI; CAZELLA, 2005) afirma ainda que, nos sistemas de recomendação demográficos, dados pessoais são requisitados ao usuário, geralmente em formulários de registro, e usados como caracterização dos mesmos e de seus interesses.

Como exemplo, (MONTANER; LOPEZ; LA, 2003) cita o método denominado *LifeStyle Finder* onde é utilizado um sistema demográfico chamado *PRIZM* da Claritas Corporation. Este sistema tem o objetivo de dividir a população americana em 62 agrupamentos demográficos de acordo com seus históricos de compra, características referentes ao tipo de vida e respostas a pesquisas.

(BURKE, 2002) menciona que todas as abordagens baseadas em aprendizagem (baseada em conteúdo, colaborativa e demográfica) possuem, de uma forma ou de outra, o problema de $ramp-up$, porque dependem de um número suficiente de avaliações para aprimorar o processo de recomendação.

4.1.4 Recomendação baseada em utilidade

Os sistemas baseados em utilidade fazem sugestões considerando um cálculo de utilidade de cada item para o usuário, sem que este precise ter um perfil de longo prazo identificado pelo sistema. A questão central consiste em criar uma função de utilidade que atenda aos interesses de cada usuário individualmente (BURKE, 2002). A entrada de dados fornece informações necessárias para definir a função de utilidade, que será utilizada para encontrar a melhor opção de acordo com suas preferências.

As técnicas baseadas em utilidade requerem do sistema uma configuração que considere todas as características dos itens na criação da função de utilidade. Além das características próprias dos itens, o sistema pode incorporar outros fatores que contribuem para a análise de valor de um produto, tais como prazo de entrega e garantia. Em alguns casos, tais fatores podem ser decisivos para uma decisão de compra (BURKE, 2002).

Como os sistemas baseados em utilidade não empregam processos que levem em consideração o histórico de avaliações do usuário, pode-se concluir que eles não enfrentam problemas típicos dos sistemas colaborativos e baseados em conteúdo, tais como: ”novo usuário”, ”novo item” ou ”espar-

sidade". Em contrapartida, neles, o usuário deve construir totalmente uma função de utilidade que determine suas preferências, o que implica considerar a importância de cada uma das características possíveis. Pode-se considerar isto, uma flexibilidade do sistema, mas também, em algum grau, um inconveniente, pois exige do usuário alto nível de interação (BURKE, 2002).

4.1.5 Recomendação baseada em conhecimento

Assim como os sistemas baseados em utilidade, os sistemas de recomendação baseados em conhecimento não utilizam perfis de longo prazo, já que suas recomendações fundamentam-se na análise da correspondência entre as necessidades do usuário e o conjunto de opções disponíveis a ele (BURKE, 2002).

Nos sistemas baseados em conhecimento, ao contrário dos sistemas baseados em utilidade, não se exige que os usuários explicitem todas as suas necessidades para fazer recomendações. Na abordagem baseada em conhecimento, o sistema utiliza efetivamente o conhecimento a respeito dos usuários e produtos para fazer inferências sobre suas preferências (BURKE, 2002).

Conforme (ADOMAVICIUS; TUZHILIN, 2005), os sistemas de recomendação podem ser melhorados com técnicas baseadas em conhecimento. Entre essas técnicas está o Raciocínio Baseado em Casos (RBC) (RICCI *et al.*, 1997).

A principal vantagem desta abordagem é aumentar a precisão e evitar limitações intrínsecas aos sistemas colaborativos e baseados em conteúdo. Outra vantagem, a ser citada, é que estes sistemas são apropriados para aplicações onde os usuários são esporádicos; onde raramente consultam o sistema para ter uma necessidade específica atendida. O principal inconveniente dos sistemas baseados em conhecimento consiste, justamente, na necessidade de adquirir o conhecimento (ADOMAVICIUS; TUZHILIN, 2005).

4.1.6 Filtragem híbrida

É a combinação de dois ou mais tipos de recomendação. Vários sistemas de recomendação combinam diferentes abordagens em uma estrutura híbrida. O principal objetivo é evitar limitações apresentadas em sistemas que aplicam apenas uma abordagem (ADOMAVICIUS; TUZHILIN, 2005)(BALABANOVIC; SHOHAM, 1997)(BURKE, 2002).

De acordo com (ADOMAVICIUS; TUZHILIN, 2005), as principais formas adotadas para combinar filtragem baseada em conteúdo e colaborativa em um sistema híbrido são:

- implementar os métodos colaborativos e baseados em conteúdo separadamente e combinar suas predições: desta forma é possível combinar as avaliações obtidas individualmente em cada um dos métodos para oferecer uma recomendação final. Outra possibilidade consiste em que o próprio sistema escolha, entre dois métodos, aquele que oferece ”melhor” resultado, baseado em alguma métrica de ”qualidade”;

- incorporar algumas características baseadas em conteúdo em uma abordagem colaborativa: a exemplo do Fab System (BALABANOVIC; SHOHAM, 1997), o sistema pode manter os perfis de usuários baseados em conteúdo, comparar diretamente os perfis para determinar os usuários semelhantes e então utilizar uma recomendação colaborativa. Assim, o usuário ativo recebe não só as recomendações de itens que foram bem avaliados por usuários com perfis semelhantes, mas também itens que sejam semelhantes àqueles já avaliados positivamente por ele;

- incorporar algumas características colaborativas em uma abordagem baseada em conteúdo: o mais comum nessa categoria é a utilização de uma técnica de redução de dimensionalidade (por exemplo, lantent semantic indexing) para criar uma ”visão” colaborativa de um grupo de perfis baseados em conteúdo;

- construir um modelo unificado que incorpore características das abordagens baseada em conteúdo e colaborativa: são inúmeras as pesquisas que tem sido desenvolvidas utilizando essa abordagem, onde várias técnicas são empregadas com o objetivo de se ter recomendações mais precisas.

Capítulo 5

TESTE DE PATERNIDADE

5.1 Hereditariedade

Gregor Mendel publicou em 1865, o resultado de uma pesquisa sobre a hereditariedade que tratava da utilização das características das evilhas (MENDEL, 1865). O experimento consistia em anotar as características físicas das plantas, como: cor e rugosidade da semente e cor da flor, e verificar o que ocorria com essas características em função do cruzamento controlado entre as plantas. Ele concluiu que a característica do indivíduo é determinada pela combinação de dois fatores e que os filhos herdam um dos fatores de seu pai e um dos de sua mãe (NAKANO, 2007).

Naquele experimento, considerando a cor da semente, o indivíduo poderia ter três genótipos possíveis, dependendo da combinação dos fatores (genes alelos) presentes no *loci* que determina a característica. Os dois alelos poderiam ser V, assim assumindo o genótipo VV, resultando em um indivíduo homozigoto. Poderia também ter um alelo V e outro v, tendo dessa maneira o genótipo Vv e sendo heterozigoto. Caso os dois alelos fossem v, o genótipo resultante era vv e neste caso os indivíduos eram homozigotos.

A cor da semente era amarela se o genótipo fosse VV, ou então Vv. Caso o indivíduo fosse vv a cor era verde. Infere-se que, mesmo sabendo o fenótipo, em alguns casos não dá pra determinar qual é o genótipo do indivíduo.

No experimento, Mendel cruzou sementes verdes com amarelas, obtendo assim na primeira geração (ou F^1) somente indivíduos amarelos. No entanto Mendel verificou que os filhos não eram

Tabela 5.1: Primeiro cruzamento para obter F^1

	V	V
v	Vv	Vv
v	Vv	Vv

Tabela 5.2: Cruzamento para obter F^2

	V	v
V	VV	Vv
v	Vv	vv

iguais aos pais. Pois quando foi obter a segunda geração (F^2), obteve sementes amarelas e verdes, na proporção de 3 para 1, 3 amarelas para 1 verde. Ou seja, $\frac{3}{4}$ dos netos eram de cor amarela mas $\frac{1}{4}$ eram da cor verde.

Interpretação dos resultados

Mendel começou a explicação dos resultados admitindo a existência de fatores que vêm dos pais (sendo um fator de um pai e outro fator da mãe). Sendo assim, para a geração de F^1 ele considerou os pais com os genótipos VV e vv (amarelos e verdes, respectivamente). Logo, todos os filhos gerados receberiam um alelo V e o outro v. Portanto quando a ervilha amarela pura (VV) é cruzada com uma ervilha verde pura (vv), o híbrido F^1 recebe o fator V e o fator v, sendo portador de ambos os fatores. Assim, todos os indivíduos gerados eram Vv. Esse cruzamento é representado na tabela 5.2.

Seguindo o mesmo raciocínio para F^2, o cruzamento entre filhos Vv resultava portanto em genótipos do tipo VV ($\frac{1}{4}$), Vv ($\frac{2}{4}$ ou $\frac{1}{2}$), vv ($\frac{1}{4}$) como na tabela X.2, constatado pela contagem dos indivíduos gerados a partir do experimento.

Esse fato foi posteriormente explicado pela meiose, que ocorre durante a formação dos gametas. Mendel havia criado então a teoria sobre a hereditariedade e da segregação dos fatores, que ficou conhecida como as leis de Mendel.

5.2 Marcadores Moleculares

Mutações em genes podem gerar indivíduos inviáveis e processos seletivos como doenças, ou cruzamentos não aleatórios podem privilegiar certos genótipos. Desta forma, em identificação são utilizados ”elementos do genoma” que até onde se sabe não conferem ao indivíduo nenhuma cracterística. Estes são chamados marcadores moleculares ideais (NAKANO, 2007).

Existe um conjunto crescente de marcadores moleculares ideais utilizados em identificação e teste de paternidade, conforme (HAMMOND *et al.*, 1994) e (LINS *et al.*, 1998). ”Perfil de DNA” de um indivíduo é o nome dado ao conjunto dos genótipos do indivíduo para esses marcadores.

5.3 A Descoberta e a Estrutura do DNA

O DNA (deoxyribose nucleic acid, ou ácido desoxirribonucleico) é o material genético que contém as informações cruciais para hereditariedade. A descoberta de sua estrutura foi publicada em (WATSON; CRICK, 1953). E sua descoberta é um marco no desenvolvimento da biologia.

Foi no dia 25 de abril de 1953 que a revista inglesa Nature publicou o fato da descoberta, que juntamente com o livro de Darwin (1859) e a publicação de Mendel (1866) está entre as maiores publicações da área biológica.

O DNA é um longo polímero de unidades simples (chamados monômeros) de nucleotídeos. Este é formado por açucares e fosfato intercalados através de ligações fosfodiéster.

Do ponto de vista químico, o DNA é um longo polímero de unidades simples (monômeros) de nucleotídeos, cujo cerne é formado por açúcares e fosfato intercalados unidos por ligações fosfodiéster. Ligadas à molécula de açúcar está uma de quatro bases nitrogenadas e é a seqüência dessas bases ao longo da molécula de DNA que carrega a informação genética. A leitura destas seqüências é feita através do código genético, o qual especifica a sequência linear dos aminoácidos das proteínas. A tradução é feita por um RNA mensageiro que copia parte da cadeia de DNA por um processo chamado transcrição e posteriormente a informação contida neste é ”traduzida” em proteínas pela tradução. Embora a maioria do RNAs produzidos sejam usados na síntese de proteínas, o RNA ribossômico que faz parte da constituição dos ribossomos tem função estrutural.

O DNA pode ser definido como um composto orgânico cujas moléculas contêm as instruções genéticas que coordenam o desenvolvimento e funcionamento de todos os seres vivos e alguns vírus. O seu principal papel é armazenar as informações necessárias para a construção das proteínas e RNAs. Os segmentos de DNA que são responsáveis por carregar a informação genética são denominados genes. O restante da seqüência de DNA tem importância estrutural ou está envolvido na regulação do uso da informação genética.

A descoberta do DNA favoreceu também o melhoramento dos organismos vivos e no entendimento do processo biológico.

5.4 Estrutura Gênica e sua Representação Gráfica

A metodologia de teste de Paternidade via DNA utiliza o conjunto de moléculas de DNA que compõe os cromossomos, e que estão localizados nos núcleos das células e arranjados aos pares. A espécie humana possui 46 cromossomos, sendo uma metade deles de origem materna, e a outra, de origem paterna. Cada cromossomo é composto por moléculas de DNA dispostas em seqüência única para cada indivíduo (SILVA, 2001).

(SILVA, 2001) afirma que o DNA de cada ser humano é único e diferente dos demais, com exceção de gêmeos univitelinos. Todo ser humano possui duas formas de cada gene, uma forma recebida de sua mãe e a outra de seu pai. Embora a maioria dos genes seja essencialmente igual entre as pessoas, algumas seqüências específicas do DNA são extremamente variáveis entre indivíduos. O local onde uma destas seqüências hipervariáveis é encontrada no cromossomo é denominado loco. Cada um destes locos pode, portanto, ter várias formas diferentes denominadas alelos.

Alguns trabalhos tem tratado problemas de genética, que envolvem pedigree, com estruturas gráficas e redes bayesianas. (LAURITZEN; SHEEHAN, 2003) é um deles e demonstra uma rede alélica como exposto na figura 5.1, no qual cada nó é um alelo de um indivíduo (cada indivíduo tem dois alelos - um recebido do pai e outro da mãe).

A figura 5.2 mostra uma rede bayesiana sobre uma simples pedigree que está em (NAKANO, 2007). Ela trata cada indivíduo como um nó. O que difere do caso anterior, no qual para representar um indivíduo eram necessários 2 nós.

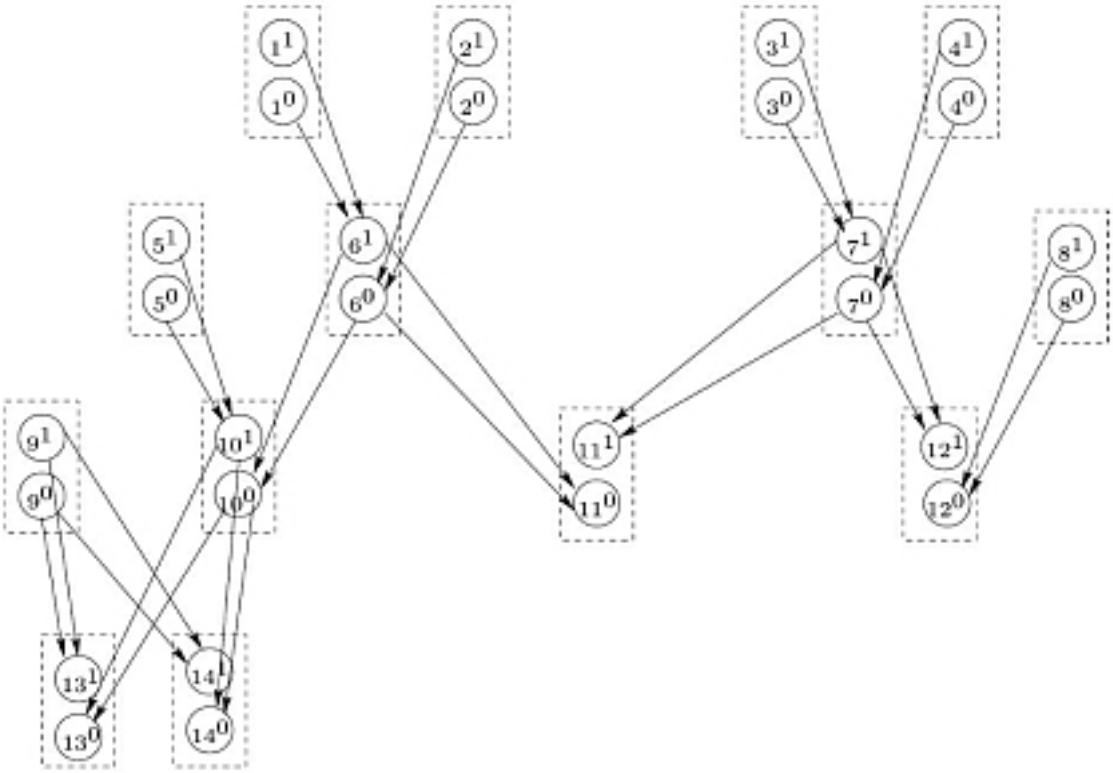

Figura 5.1: Rede alélica. Fonte: (LAURITZEN; SHEEHAN, 2003)

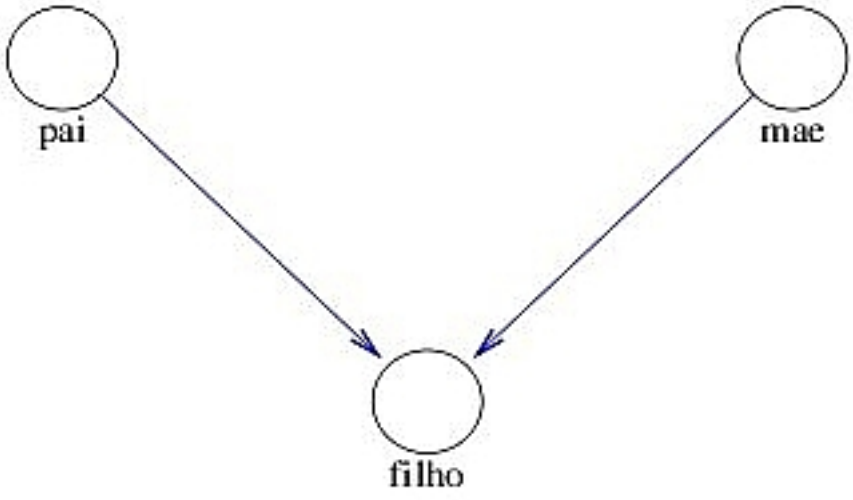

Figura 5.2: Rede Bayesiana de um Trio (pai, mãe e filho).

5.5 Complexidade dos Testes de Paternidade

O teste de paternidade tradicional faz cálculos simples. Isso ocorre devido a presença do possível pai e da mãe, além da criança. Nos casos mais complexos, para resolver esse problema, em geral, utiliza-se alguma metodologia ligada ao pedigree e que envolve probabilidades. Nestes testes tradicionais é utilizado a chamada *exclusão de paternidade* para realização do teste. A exclusão tenta detectar no filho um gene que veio do pai biológico e que não esteja no suposto pai para eliminar a possibilidade dele ser o pai.

(NAKANO, 2007) diz que os casos simples envolvem o reclamente (suposto filho), o demandado (suposto pai) e a mãe e podem ser resolvidos com uma planilha eletrônica. No entanto casos mais complexos, quando o perfil do demandado não está disponível e seus irmãos são examinados, ou quando há outros vínculos discutíveis como se um irmão do demandado não for legítimo ou há consanguineidade, a quantidade e complexidade dos cálculos cresce, o que mostra a necessidade de automação dos mesmos.

5.6 Combinação dos resultados de vários *loci*

(NAKANO, 2007) demonstra como o resultado de vários experiementos pode ser obtido a partir da aplicação básica do teorema de Bayes que corresponde ao produtório das razões de verossimilhanças.

A partir de uma hipótese H_1, dos dados ($\triangle$) e da priori π, a razão de verossimilhança pode ser obtida.

A partir do teorema de bayes obtém-se:

$$P(H_1|\triangle) = \frac{P(\triangle|H_1)}{P(H_1)} * \pi$$

que corresponde a:

$$P(H_1|\triangle) = \frac{L(H_1|\triangle)}{P(H_1)} * \pi$$

Logo:

$$\frac{P(H_1{=}Verdadeiro|\triangle)}{P(H_1{=}Falso|\triangle)} = \frac{L(H_1{=}verdadeiro|\triangle)}{L(H_1{=}falso|\triangle)} * \frac{\pi}{1-\pi}$$

$\frac{L(H_1{=}verdadeiro|\triangle)}{L(H_1{=}falso|\triangle)}$ é chamada de razão de verossimilhança (ou *likelihood ratio*).

Para aplicar a razão de verossimilhança ao teste de paternidade, em cada *loci* obtém-se esta razão de verossimilhança.

Após o teste dos vários *locis*, multiplica-se todas as razões de verossimilhança, obtendo uma global. Após isto, realiza o processo de verificar qual a probabilidade do indivíduo ser o pai.

Capítulo 6

METODOLOGIA

Este capítulo, descreve como a pesquisa foi feita para elaboração do trabalho. Esclarece como as informações e conceitos estabelecidos citados foram obtidos.

6.1 Tipo de Pesquisa

A natureza da pesquisa do trabalho é aplicada ou tecnológica.

Considerando os objetivos a pesquisa é exploratória. A partir de dados, e de uma experimentação, o objetivo foi elaborar modelos que sejam satisfatórios para resolução dos problemas em questão. Espera-se que estes modelos sejam inovadores e/ou explicativos.

Quanto aos procedimentos, o tipo de pesquisa é experimental. Neste tipo de pesquisa o investigador analisa o problema, constrói suas hipóteses e trabalha manipulando os possíveis fatores e as variáveis que se referem ao fenômeno observado.

Em termos do local, a pesquisa não necessita ser feita em campo, portanto pode ser considerada de laboratório. No caso da base teórica, a pesquisa é bibliográfica, e quanto ao tempo o estudo é transversal.

6.2 Procedimentos metodológicos

A pesquisa foi realizada no período de fevereiro a novembro de 2008.

Inicialmente, foi feita uma revisão bibliográfica sobre princípios das redes bayesianas, estrutura e formatos utilizados em sistemas de redes bayesianas e os algoritmos de inferência.

A partir da base teórica e prática de redes bayesianas, o estudo consistiu em avaliar os problemas, e para tal foram estudados os conceitos vindos de outras áreas como sistemas de recomendação e genética.

Em genética, foi estudado mais especificamente genética de populações, genótipos, genes, alelos, locus, pedigrees, meiose, mitose, cromossomos, dentre outros conceitos que envolvem a árvore genealógica, e as propriedades de transmissão entre gerações humanas dos genes, e do DNA, enfim regras de hereditariedade.

No caso de sistemas de recomendação vários tópicos foram abordados. Estruturas de sistemas de recomendação, formas de processamento da informação e exemplos de como eles podem contribuir para facilitar o usuário a ter acesso a informação que ele deseja. Forma vistos também conceitos de recuperação de informação devido a esta área estar muito próxima de sistemas de recomendação.

O enfoque do estudo em sistemas de recomendação foi em torno de arquiteturas desses sistemas, para assim, obter conhecimento de como elaborar um modelo que utilizasse as redes bayesianas e que tivesse a menor quantidade de falhas dentro de sua estrutura.

A consulta foi realizada a partir de livros, monografias, teses e dissertações disponibilizadas na internet e na literatura em geral.

Capítulo 7

RESULTADOS E DISCUSSÕES

7.1 Teste de Paternidade

Teste de Hipótese

O modelo descrito abaixo, através do teste de hipótese, está em (NAKANO, 2007). O teste consiste em utilizar uma medida probabilística que indica se a hipótese testada é ou não verdadeira. O teste é baseado em afirmar se o reclamante é filho legítimo do suposto pai.

H_0: o reclamante é filho

H_1: o reclamante não é filho

O teste pode ser resumido em ”determinar a probabilidade de que o pai presumido seja o pai verdadeiro”. Este teste pode ser feito com a utilização de redes bayesianas, pois, as variáveis ”genótipo do pai verdadeiro”, ”genótipo do pai presumido”, e ”pai verdadeiro é igual a pai presumido” são discretas e enumeráveis.

Considera-se que o pai presumido é um indivíduo da população, a seu genótipo pode ser atribuído uma probabilidade a *priori* da própria freqüência populacional explícita na tabela 7.1 abaixo.

Na tabela 7.2, há a probabilidade condicional de *pai_verdadeiro* dado o *pai_presumido* e a variável *eh_igual* (que corresponde aos indivíduos serem os mesmos).

O funcionamento do teste pode ser obtido através da identificação de dois eventos ditos:

A: os genótipos do pais presumido é igual ao do pai verdadeiro $(G_1 = G_2)$

	$P(pai\_presumido)$
A_1A_1	0.3
A_1A_2	0.4
A_2A_2	0.3

Tabela 7.1: Probabilidade a *priori* de pai_presumido

Genótipo *pai_verdadeiro*	A_1A_1		A_1A_2		A_2A_2	
eh_igual	V	F	V	F	V	F
A_1A_1	1	0.3	0	0.3	0	0.3
A_1A_2	0	0.4	1	0.4	0	0.4
A_2A_2	0	0.3	0	0.3	1	0.3

Tabela 7.2: $P(pai\_verdadeiro|pai\_presumido, eh\_igual)$

B: os indivíduos, pai presumido e pai verdadeiro, são iguais $(i_1 = i_2)$, neste caso o pai presumido é o pai verdadeiro.

Assim, B^c é o evento complementar (os indivíduos são diferentes).

Utiliza-se a regra de Bayes:

$$P(B|A) = \frac{P(A|B)}{P(A)} * P(B)$$

$$P(B^c|A) = \frac{P(A|B^c)}{P(A)} * P(B^c)$$

Considera-se que a *priori* de nenhuma das alternativas deve ser favorecida, então $P(B) = P(B^c)$. Também sabe-se que $P(B^c|A) = 1 - P(B|A)$. Com isso, é feita a seguinte manipulação:

$$P(B|A) = \frac{P(A|B)*P(B)}{P(A)} = \frac{P(A|B)*P(B^c)}{P(A)[1+P(B^c|A)-P(B^c|A)]} = \frac{P(A|B)*P(B^c)}{P(A)*[P(B^c|A)+P(B|A)]} =$$

$$\frac{P(A|B)*P(B^c)}{P(A)*P(B^c|A)+P(A)*P(B|A)} = \frac{P(A|B)*P(B^c)}{P(A)*P(B^c|A)+\left[\frac{P(B|A)*P(A)}{P(B)}\right]*P(B)} = \frac{P(A|B)*P(B^c)}{P(A)*P(B^c|A)+P(A|B)*P(B)} =$$

$$\frac{P(A|B)}{\frac{P(A)*P(B^c|A)}{P(B^c)}+\frac{P(A|B)*P(B^c)}{P(B^c)}} = \frac{P(A|B)}{P(A|B^c)+P(A|B)}$$

$P(A|B) = 1$, pois dado que os indivíduos são os mesmos, a probabilidade de o genótipo ser igual também é 1.

Logo chega-se a:

$$\frac{1}{P(A|B^c)+1} \tag{7.1}$$

Chega-se a seguinte conclusão: Dado os genótipos iguais do pai presumido e do pai verdadeiro, a possibilidade do pai presumido ser o pai verdadeiro é inversalmente proporcional a freqüência populacional do genótipo avaliado.

Pois, a $P(A|B^c)$ refere-se justamente a probabilidade de os indivíduos terem o mesmo genótipo, dado que eles não são a mesma pessoa. Ou seja, a probabilidade de encontrar dois indivíduos diferentes com o mesmo genótipo. Logo, quanto mais raro for o genótipo maior a probabilidade de que o demandado seja o pai biológico, pois, é menos provável que um indivíduo tomado ao acaso tenha esse genótipo.

Teste para o modelo de referência

Com a utilização da metodologia acima proposta, pode-se verificar como é feito um teste no qual tem-se disponível o perfil de DNA do suposto pai. A família em estudo pode ser vista na figura 7.1.

Na família, nota-se que os filhos são $S10/S10$ e $S10/SX$. Com o primeiro filho, dá pra inferir que o pai possui um gene $S10$. Pois pelo menos um gene foi herdado do pai. Com o segundo filho, verifica-se que o outro gene do pai verdadeiro só pode ser $S10$. Logo seu genótipo é $S10/SX$ como exposto na figura.

Considera-se um suposto pai com esse genótipo, $S10/SX$, pode-se obter a probabilidade dele ser o verdadeiro pai. São considerados 7 testes (ou casos). Os casos são denominados $C1, C2, C3, C4, C5, C6$ e $C7$.

A contagem dos indivíduos da população é dado na tabela 7.3. Em cada caso há um número de indivíduos para cada genótipo. Esse número será a base para obter a freqüência genotípica que foi contabilizada na tabela 7.4.

O resultado do teste utiliza a equação 7.1 e pode ser visto na tabela 7.5.

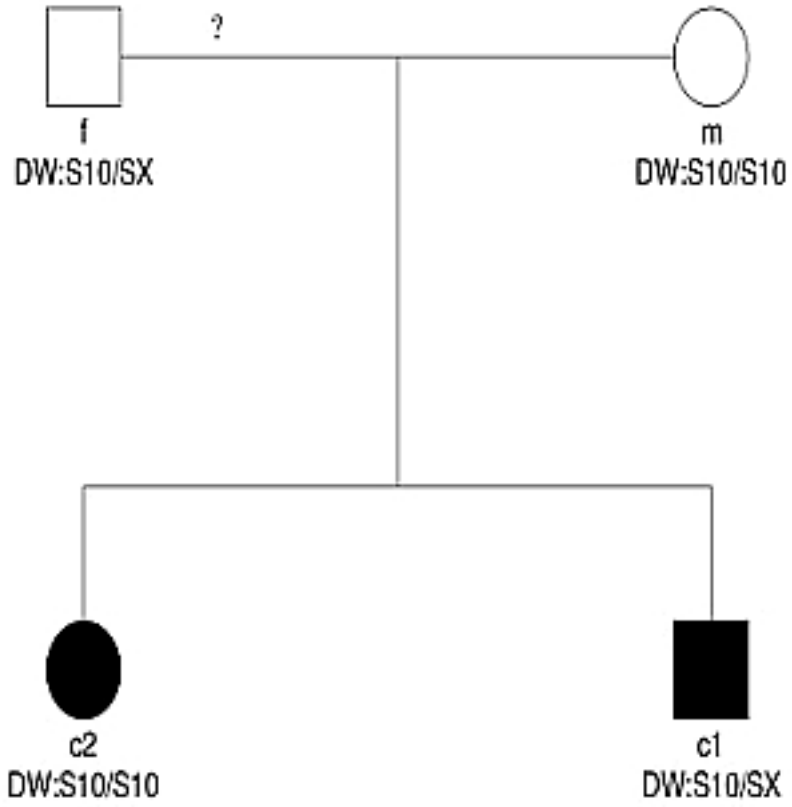

Figura 7.1: Família a ser estudada. Fonte: (NAKANO, 2007)

	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$
$S10/S10$	49	45	25	15	10	5	1
$S10/SX$	2	10	50	70	80	90	98
SX/SX	49	45	25	15	10	5	1

Tabela 7.3: Contagem dos indivíduos da população nos 7 casos.

	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$
$S10/S10$	0.49	0.45	0.25	0.15	0.1	0.05	0.01
$S10/SX$	0.02	0.1	0.5	0.7	0.8	0.9	0.98
SX/SX	0.49	0.45	0.25	0.15	0.1	0.05	0.01

Tabela 7.4: Frequencias genotípicas da população.

	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$
$fS10/SX$	0.02	0.1	0.5	0.7	0.8	0.9	0.98
$P(fserpaiverdadeiro)$	0.98	0.909	0.667	0.588	0.555	0.526	0.505

Tabela 7.5: Resultado do teste para os 7 casos demonstrados.

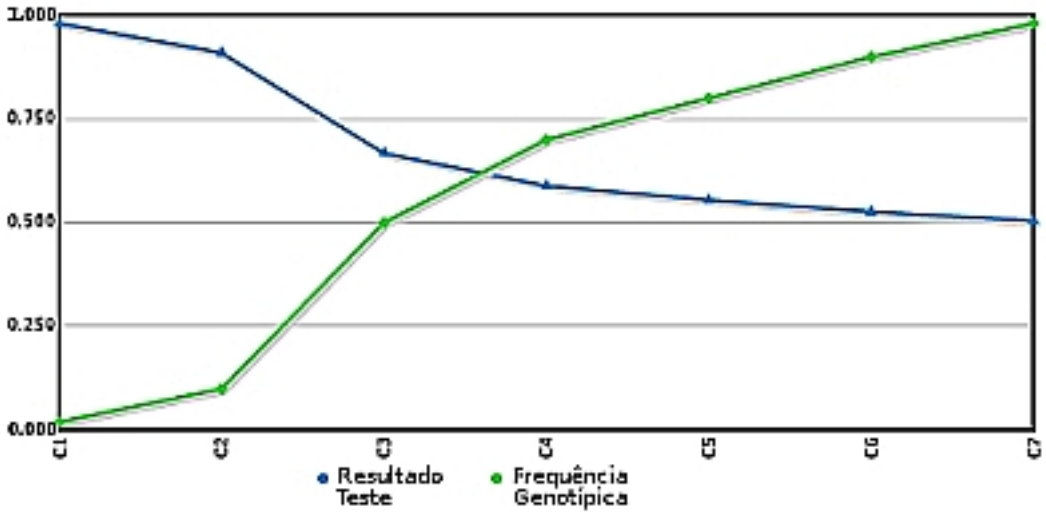

Figura 7.2: Gráfico: Frequencia do genótipo do suposto pai na população e o Resultado do Teste

Loci 1			*Loci* 2			*Loci* 3			*Loci* 4		
AA	*Aa*	*aa*	*BB*	*Bb*	*bb*	*XX*	*Xx*	*xx*	*MM*	*MN*	*NN*
442	247	311	485	265	250	185	465	350	340	350	310

Tabela 7.6: Distribuição para cada genótipo - numa população de 1000 indivíduos

Como visto na figura 7.2, a probabilidade do suposto pai ser o pai verdadeiro vai depender da freqüência do genótipo na população. Logo no gráfico abaixo é demonstrado que a probabilidade de ser pai, é maior no caso do genótipo ter freqüência genotípica menor.

Aplicação do Modelo Proposto

O modelo proposto é baseado na probabilidade dos indivíduos pai biológico e pai presumido serem iguais. Mas, ao invés de ele tratar apenas os casos nos quais o perfil do suposto pai foi totalmente reconstituído o modelo se aplica a qualquer realidade de *pedigree*.

Além disso o modelo utiliza a razão de verossimilhança para combinação dos resultados de vários *loci*, método que pode ser visto em (MARSHALL *et al.*, 1998).

Os 2 exemplos expostos abaixo foram baseados em 4 *loci* (representados pelos números 1, 2, 3 e 4). As freqüencias dos genótipos estão expressos na tabela 7.7 baseado na tabela 7.6 que corresponde a contagem de indivíduos de uma população de 1000 pessoas.

Para a prática do teste, foi elaborado um aplicativo baseado em linguagem de programação PHP. Foram implementados os algoritmos inferência por enumeração (exato) e o Likelihood Sampling

Loci 1			*Loci* 2			*Loci* 3			*Loci* 4		
AA	*Aa*	*aa*	*BB*	*Bb*	*bb*	*XX*	*Xx*	*xx*	*MM*	*MN*	*NN*
0.442	0.247	0.311	0.485	0.265	0.250	0.185	0.465	0.350	0.340	0.350	0.310

Tabela 7.7: Frequências Genotípicas para os 1000 indivíduos da tabela 7.6

(aproximado). No entanto para os exemplos expostos abaixo, o algoritmo exato demonstrou que pode ser utilizado. Em algum caso no qual for preciso introduzir parentes distantes do suposto pai, pode ser que o algoritmo aproximado seja mais eficiente, devido a várias variáveis que deverão ser inseridas.

Para execução do programa implementado é necessário realizar upload de arquivos textos que compreendem a freqüencia genotípica (probabilidade a *priori*), tabelas de probabilidades condicionais e estrutura familiar. Em todos os arquivos, as linhas iniciadas com # são consideradas comentários e são ignorados para geração dos dados.

Além disso, a cada arquivo texto enviado, o programa gera um outro em formato XML. Isto facilita o aplicativo a gerenciar os dados que foram inseridos pelo usuário do sistema. Um trecho de um arquivo gerado em XML pode ser visto em C.5.

Os arquivos que possuem a probabilidade a *priori* (freqüencias alélicas) e as tabelas de probabilidade condicional são genéricos para todos os exemplos. Eles estão expressos nos códigos dos apêndices C.1 e C.2.

O primeiro exemplo de teste de paternidade aborda uma família como expresso na figura 7.3. A rede corresponde a uma situação no qual o suposto pai é desconhecido. No entanto, tem-se a viúva, e 3 filhos legítimos. Neste caso o indivíduo 5 é o reclamante que diz ser filho de 3 (suposto pai). Sua mãe está representada pelo indivíduo 1. 4 é a esposa do demandado (indivíduo 3). Este último não tem seu perfil genético. O arquivo texto que corresponde a essa família está em C.3. A tabela 7.8 explica cada indivíduo da rede, e expõe seu genótipo para 4 *loci*.

Pode-se notar que no arquivo texto de entrada, não há referência ao vértice *NohTeste* que está na figura 7.3. O programa gera-o automaticamente para que ele seja utilizado como vértice de consulta, pois quando o algoritmo de inferência é executado, calcula-se a probabilidade desse vértice assumir verdadeiro. Busca-se, dessa forma, encontrar a probabilidade de o pai biológico ser o suposto pai (de eles serem iguais). Portanto o *NohTeste* calcula, neste caso, a probabilidade de 2 e 3 serem iguais.

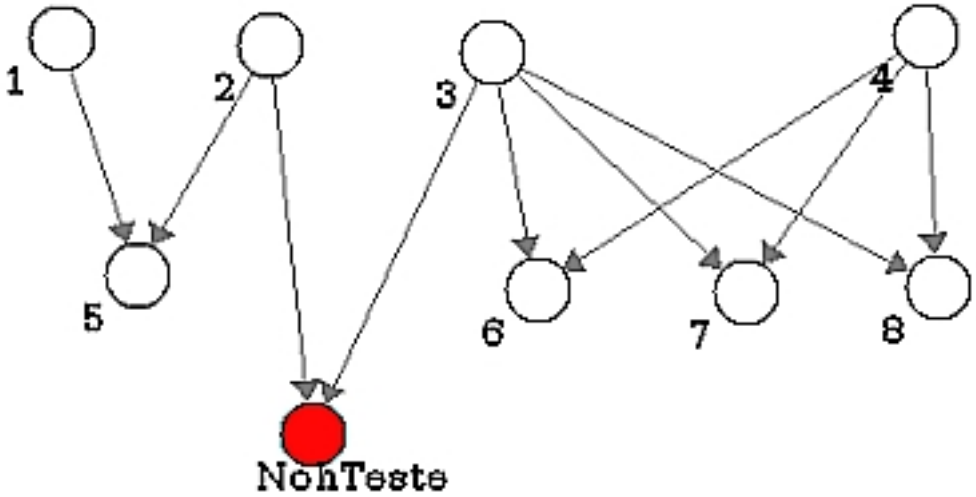

Figura 7.3: Rede Bayesiana para o primeiro exemplo de teste de paternidade

Identificador do Noh	O que representa na rede	Genótipo identificado
1	Mãe do reclamante	*aaBBxxMN*
2	Pai biológico (desconhecido) de 5	-
3	Suposto pai	-
4	Exposa de 3	*aaBbXxMN*
5	Indivíduo que reclama ser filho de 3	*AaBBXxMM*
6	Filho legítimo de 3 e 4	*AaBBxxMM*
7	Filho legítimo de 3 e 4	*AaBbXXMM*
8	Filho legítimo de 3 e 4	*AaBBxxMN*
NohTeste	Nó que é gerado pelo aplicativo para o teste	

Tabela 7.8: Legenda dos vértices da rede da figura 7.3

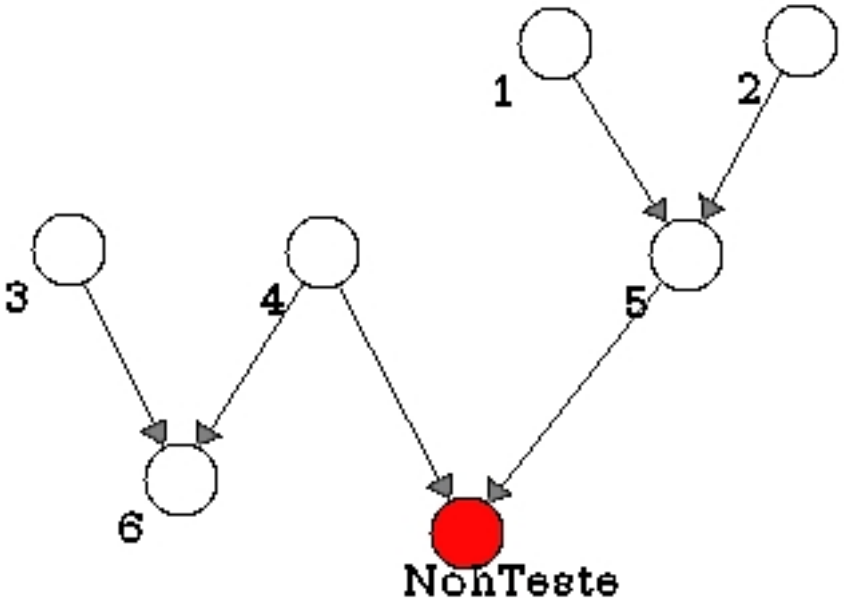

Figura 7.4: Rede Bayesiana para o segundo exemplo de teste de paternidade

O segundo exemplo de teste de paternidade aborda uma família como expresso na figura 7.4. Também neste caso, é avaliado uma situação no qual o demandado não tem seu perfil genético conhecido. No entanto os pais dele (supostos avós) estão presentes e representados pelos indivíduos 1 e 2 na rede. O arquivo texto de entrada do programa que deve ser construído para esta família está em C.4. A tabela 7.9 explica cada indivíduo da rede, e expõe seu genótipo para 4 *loci*.

Para o primeiro exemplo, o resultado da execução do programa está exposto na figura 7.6. Nota-se um resultado de cerca de 93% de o demandado ser o pai biológico. A probabilidade recomendada

Identificador do Noh	O que representa na rede	Genótipo identificado
1	Suposto avô do reclamante	*AAbbXXMN*
2	Suposta avó do reclamante	*aabbXxMN*
3	Mãe legítima do reclamante	*aaBBxxNN*
4	Pai biológico (desconhecido)	-
5	Suposto pai de 5	-
6	Indivíduo que reclama ser filho de 5	*AaBbXxMN*
NohTeste	Nó que é gerado pelo aplicativo para o teste	

Tabela 7.9: Legenda dos vértices da rede da figura 7.4

```
Teste para o loci 1.
Inferência em favor de true: 0.0192646046621
Inferência em favor de false: 0.00659963653375
Resultado do Teste de Paternidade para loci 1: 0.744835486037

Teste para o loci 2.
Inferência em favor de true: 0.00392006270508
Inferência em favor de false: 0.00154862583984
Resultado do Teste de Paternidade para loci 2: 0.716819521331

Teste para o loci 3.
Inferência em favor de true: 0.000274926708984
Inferência em favor de false: 0.000218758886719
Resultado do Teste de Paternidade para loci 3: 0.556886227545

Teste para o loci 4.
Inferência em favor de true: 0.00100229882812
Inferência em favor de false: 0.0006833203125
Resultado do Teste de Paternidade para loci 4: 0.594617611991

Likelihood: 13.6210540479
Resultado combinado do Teste de Paternidade: 0.931605478187 (93,1605478187%)
Tempo gasto: 0.066642 segundos.
```

Figura 7.5: Resultado para execução do programa com o primeiro exemplo

em teste reais é acima de 99,9999%. Com mais *loci* poderia haver uma tendência a esse valor, se o suposto pai ser realmente o pai biológico.

Para o segundo exemplo, o resultado ao executar o aplicativo está na figura 7.6. Ele demonstra uma probabilidade muito baixa no teste de paternidade (cerca de 1,2%). Provavelmente, com um teste com mais *loci*, o resultado deve tender ainda mais a 0.

7.2 Sistemas de Recomendação

Para atingir o objetivo de propor um modelo de sistemas de recomendação baseado em redes bayesianas, e aplicá-lo à recuperação de informação, foi necessário estudar e compreender alguns sistemas, e principalmente características, classificações e problemas.

```
Teste para o loci 1.
Inferência em favor de true: 0.0026468449605
Inferência em favor de false: 0.005597270705
Resultado do Teste de Paternidade para loci 1: 0.321058688147

Teste para o loci 2.
Inferência em favor de true: 0.00438511234375
Inferência em favor de false: 0.00864250296875
Resultado do Teste de Paternidade para loci 2: 0.33660130719

Teste para o loci 3.
Inferência em favor de true: 0.00232509445313
Inferência em favor de false: 0.00932537882812
Resultado do Teste de Paternidade para loci 3: 0.199570815451

Teste para o loci 4.
Inferência em favor de true: 0.0029430625
Inferência em favor de false: 0.0143789625
Resultado do Teste de Paternidade para loci 4: 0.169902912621

Likelihood: 0.0122444541446
Resultado combinado do Teste de Paternidade: 0.0120963410512 (1,20963410512%)
Tempo gasto: 0.041487 segundos.
```

Figura 7.6: Resultado para execução do programa com o segundo exemplo

Considera-se neste trabalho a criação de um modelo que fosse capaz de ser aplicado em um sistema de recuperação de informação. Nestes sistemas o usuário insere termos e obtém resultados de acordo com as palavras digitas e os documentos. Um modelo de recomendação aplicado a um sistema como esse leva em conta também o perfil do usuário (todo conhecimento extraído dele até aquele momento).

Dado as formas de sistemas de recomendação, e as arquiteturas existentes, o modelo proposto pode ser considerado híbrido e contém as seguintes classificações: sistema baseado em conteúdo, sistema demográfico e sistema baseado em conhecimento. Foi preciso incorporar o aspecto de *clustering* para verificar usuários semelhantes em relação a um assunto.

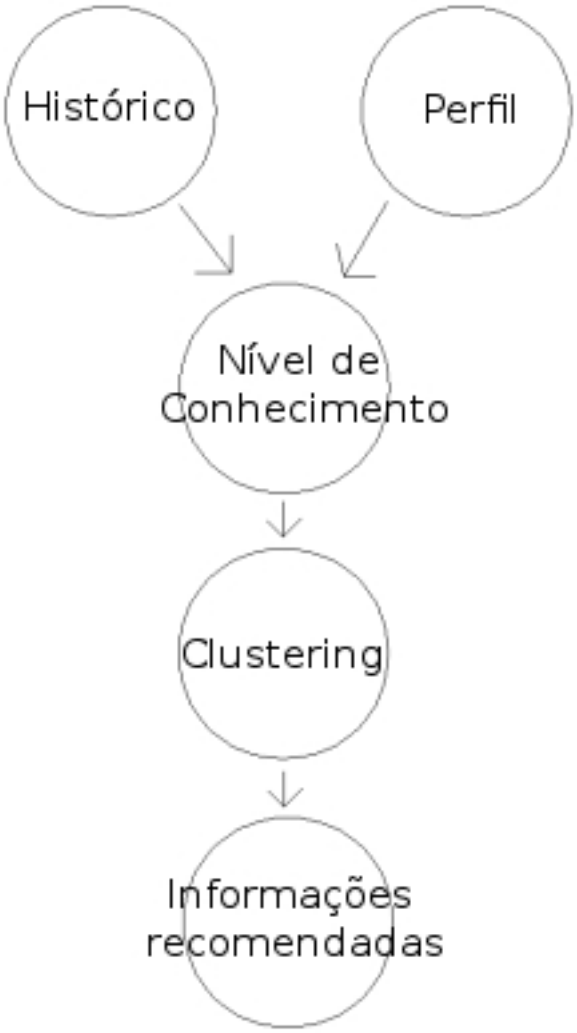

Figura 7.7: Modelo proposto através de Rede bayesiana (com utilização de Clustering) para arquitetura de sistemas de recomendação

O modelo foi elaborado com o intuito de eliminição, ou pelo menos minimização dos problemas que os sistemas de recomendação possuem. O modelo está exposto na figura 7.7.

7.2.1 Nível de Conhecimento

O núcleo do modelo está na variável nível de conhecimento. Apesar de ser difícil avaliar o quanto um usuário conhece de um assunto, se o sistema conseguir constatar algo próximo a esse dado, ele pode recomendar conteúdos (informações) mais adequadas. Um exemplo disso: Um produtor rural sem conhecimento aprofundado de ciência do solo e um doutor da área de solos pesquisam um mesmo termo relacionado a solos. Provavelmente, não seria interessante que um sistema de informação

indicasse aos dois as mesmas informações. De acordo com o perfil de um e de outro, o sistema deveria indicar a melhor informação para cada um.

Baseado nessa diferença do contexto que o usuário se encontra, o modelo buscou analisar o nível de conhecimento do assunto ligado ao termo pesquisado para retornar uma informação mais adequada. Assim, alguns fatores foram levados em conta: o histórico de pesquisas anteriores, o perfil demográfico e o nível de conhecimento de usuários que se parecem com aquele que realiza a pesquisa.

7.2.2 Histórico

A variável *historico* influencia o nível de conhecimento da seguinte maneira: dado que o usuário acessou determinados itens que estão relacionados, há uma tendência de ele possuir conhecimento daquele assunto. Afinal, se o usuário pesquisa a 2 anos sobre um tema, já consta em seu histórico um imenso volume de termos pesquisados do assunto. No entanto há o problema de verificar quais itens são semelhantes. Por exemplo, dado que o histórico do usuário contém as palavras ”gene”, ”genótipo”, ”cromossomo” e ainda vários termos relacionados com esses, o sistema deve entender que o usuário provavelmente já tem um conhecimento do assunto de biologia, ou genética por exemplo. Portanto o sistema deve ser capaz de identificar tanto a semelhança como a proximidade entre termos pesquisados.

Durante a pesquisa, foi encontrado o problema de verificar qual a semelhança (ou proximidade) entre 2 termos. De acordo com vários autores, isto pode ser tratado com a ajuda de um tesauro. É um termo bastante utilizado na área de recuperação de informação. Os tesauros e os sistemas de classificação bibliográfica são as linguagens documentárias mais conhecidas(SOUZA, 2005). Os instrumentos para a representação da informação para indexação, armazenamento e recuperação de informações são considerados linguagens documentárias (SOUZA, 2005).

7.2.3 Perfil demográfico

A variável *perfil_demografico* também pode influenciar o nível de conhecimento do assunto pesquisado. O perfil demográfico corresponde as informações básicas do usuário como idade, sexo, cidade, estado, país, profissão, etc. Ao verificar que um usuário é de uma região o sistema pode inferir que ele tem um conhecimento maior ou menor do assunto. Por exemplo, compara-se um usuário da cidade de Manaus (Brasil), e um usuário de Napoli (Itália), ambos pesquisam sobre ”Floresta Amazônica”, po-

deria ser constatado que um tem mais ou menos conhecimento do que o outro. Da mesma forma, um usuário de 13 anos e outro de 45 devem obter informações diferentes ao pesquisar sobre o termo ”tecnologia”. Pode-se dizer que é utilizado o conceito de recomendação demográfica, pois são baseados em classes demográficas.

7.2.4 O Clustering

A variável denominada *clustering* nada mais é que a influência de usuários que tenham aproximadamente o mesmo nível de conhecimento sobre um tema, em relação àquele que realiza a pesquisa. O clustering divide os usuários do sistema em grupos, de tal maneira que usuários semelhantes (em relação a um assunto) fiquem em um mesmo grupo. Para isto pode-se utilizar o algoritmo k-means que está no apêndice B. Ao pesquisar um termo, o nível de conhecimento ser obtido, e um grupo de usuários serem estabelecidos de acordo com o clustering, será necessário então indicar conteúdos adequados ao usuário.

Para indicação do conteúdo que o usuário necessita, é utilizado a variável $informacoes$, dado o estado da variável clustering. Nela contém as informações do retorno da busca. Ela assume principalmente informações que usuários do mesmo grupo (formado no clustering) acessaram anteriormente e que tiveram boa avaliação. O conceito de sistemas de recomendação utilizado é o de colaboração.

Em relação a avaliação de um item, o usuário pode ou não avaliar explicitamente uma informação. No caso explícito o sistema pode exibir um formulário de notas que pode ser utilizado para avaliar a satisfação do conteúdo. O sistema pode capturar a importância para o usuário de forma implícita. Isto ocorre quando se aplica um coeficiente que mede a relação (através de uma divisão) entre a quantidade de elementos (palavras) com o tempo permanecido visualizando as informações.

7.2.5 Modelo Proposto e os Problemas de Recomendação

No capítulo foi abordado os principais problemas de sistemas de recomendação. A elaboração do modelo buscou minimizar estes problemas. No entanto alguns, foram inevitáveis. Abaixo segue análise sobre quando os problemas conseguem ser minimizados e quando isto não acontece.

Problema do*novo usuário:* Ao ser implementado, este modelo tenta minimizar o problema do novo usuário, pois ele é um modelo híbrido e que aborda o aspecto demográfico. Assim, a idéia é que

mesmo sem ter um histórico formado anteriormente, o sistema possa apoiar no aspecto demográfica para avaliar o conhecimento do usuário. Em alguns casos pode não ser uma solução eficiente, e depender dos primeiros acessos do usuário para realmente considerar suas preferências através de seu acesso. No entanto se informações demográficas forem expressivas em relação a assunto, o modelo irá melhor se comportar.

Superespecialização: Este problema é um dos mais difíceis de ser minimizado se for levado em conta que sugerir algo novo (e ao mesmo tempo interessante) para um usuário é algo complexo. No modelo, isso é tratado entre a variável clustering e a nível de conhecimento. A idéia é sugerir informações úteis de um usuário para outro que tenha nível de conhecimento próximo. Se a informação for considerada de uma nova área para o segundo usuário, o problema de *superespecialização* poderá ser minimizado.

Avaliações esparsas: Este problema ocorre devido ao fato de alguns itens serem pouco avaliados. E quando são, mesmo que tenham boa avaliação geralmente não aparecem na recomendação. No modelo sugerido, isto pode ocorrer com freqüencia. Pois, enquanto uma informação não for descoberta, e acessada (consequentemente avaliada) por um usuário, ela não será sugerida. No entanto, o peso do usuário que irá avaliar, pode ser considerado de tal forma que mesmo com uma única avaliação ela possa ser sugerida para um usuário. Isto pode ser feito com a utilização de uma função que leve em conta não apenas as avaliações mas a quantidade de avaliações realizadas.

Capítulo 8

CONSIDERAÇÕES FINAIS

A abordagem da pesquisa foi feita em dois problemas considerados atuais, pois tem sido tratados em trabalhos há alguns anos. Foi utilizado redes bayesianas para propor modelos e soluções em software. A metodologia bayesiana foi utilizada por se tratar de problemas que apresentam muitas vezes ausência de informações.

8.1 Teste de Paternidade

O modelo proposto utiliza conceitos de vários outros, e de hereditariedade. Para os casos analisados, viu-se que ele se apresenta resultados de maneira satisfatória.

O teste de paternidade tem sido abordado por profissionais das áreas de biologia e de direito. Esta última tem tido trabalhos questionadores sobre o assunto, pois o teste é necessário em alguns casos judiciais.

No caso da exclusão de paternidade, quando determina-se que o suposto pai não é o pai biológico, há uma aceitação maior no cenário científico. No entanto para provar o contrário, metodologias tem sido propostas, e foi no intuito de abordar este problema, de afirmar que um indivíduo é o pai biológico, que este trabalho foi elaborado.

Em relação a união dos resultados de vários *loci* foi utilizado uma metodologia baseada em razão de verossimilhança por ser razoavelmente aceita. Outras abordagens de combinação de experimentos podem ser vistas em trabalhos estatísticos.

Além de ser modelado, o problema foi implementado em software, e constatou-se bons resultados. No entanto, a implementação pode ser melhorada através de integração com uma interface gráfica mais amigável para facilitar a utilização do usuário.

8.2 Sistemas de Recomendação

O problema de construir sistemas de recomendação eficientes foi abordado com o foco sobre os problemas dos existentes, e principalmente nas análises de outros autores sobre essas dificuldades.

Baseado nessa perspectiva, o núcleo do modelo sugerido está na variável nível de conhecimento, e por isso o modelo pode ser considerado ”baseado em conhecimento”. A principal vantagem é que essa abordagem busca uma maior precisão em relação a outros sistemas de recomendação.

O principal problema de sistemas baseado em conhecimento, consiste em adquirir o conhecimento. Logo pensou-se dois fatores que influenciasse o nível de conhecimento (histórico, e o fator demográfico) justamente para que quando um não for relevante o outro ter maior peso.

A elaboração através das redes bayesianas pode ser justificada devido a inexistência de informações e por as redes serem especializadas nessas abordagens nas quais muitos dados estão incompletos, ou ”ocultos”. Sugere-se para um próximo trabalho, a aplicação do modelo para que ele possa ser testado.

Referências Bibliográficas

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, v. 17, n. 6, p. 734–749, jun. 2005.

BALABANOVIC, M.; SHOHAM, Y. Fab: content-based, collaborative recommendation. *Commun. ACM*, ACM, New York, NY, USA, v. 40, n. 3, p. 66–72, 1997. ISSN 0001-0782.

BILLSUS, D.; PAZZANI, M. J. User modeling for adaptive news access. *Communications of the ACM*, Springer Netherlands, v. 10, p. 2–3, jun 2000.

BURKE, R. Hybrid recommender systems: Survey and experiments. *Communications of the ACM*, v. 12, p. 331–370, nov. 2002.

CASTILLO, E.; GUTIERREZ, J. M.; HADI, A. S. *Expert Systems and Probabilistic Network Models*. 1. ed. [S.l.]: Springer A., 1996.

CLAYPOOL, M.; GOKHALE, A.; MIR, T.; MURNIKOV, P.; NETES, D.; SARTIN, M. Combining content-based and collaborative filters in an online newspaper. In: *In Proceedings of ACM SIGIR Workshop on Recommender Systems*. [S.l.: s.n.], 1999.

GOLDBERG, D.; NICHOLS, D.; OKI, B. M.; TERRY, D. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, ACM, New York, NY, USA, v. 35, n. 12, p. 61–70, 1992. ISSN 0001-0782.

GOOD, N. Combining collaborative filtering with personal agents for better recommendations. *American Association of Artificial Intelligence*, p. 439–446, 1999.

HAMMOND, H. A.; JIN, L.; ZHONG, Y.; CASKEY, C. T.; CHAKARBORTY, R. Evaluation of 13 short tandem repeat loci for use in personal identification applications. *American Journal of Human Genetics*, v. 55, p. 175–189, July 1994.

HECKERMAN, D. *A Tutorial on Learning With Bayesian Networks*. [S.l.], 1995. Disponível em: <ftp://ftp.research.microsoft.com/pub/tr/tr-95-06.pdf>.

JENSEN, F. V. *Bayesian Networks and Decision Graphs*. Springer, 2001. Hardcover. (Information Science and Statistics). ISBN 0387952594. Disponível em: <http://www.amazon.ca/exec/obidos-/redirect?tag=citeulike09-20\&path=ASIN/0387952594>.

KLIR, G. J.; YUAN, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall PTR, 1995. Paperback. ISBN 0131011715. Disponível em: <http://www.amazon.ca/exec/obidos-/redirect?tag=citeulike09-20\&path=ASIN/0131011715>.

KOLMOGOROV, A. N. *Grundbegriffe der Wahrsheinlichkeitsrechnung (Fundamentos da Teoria de Probabilidade)*. Berlin: Springer, 1933.

KONSTAN, J. Recommender systems: A grouplens perspective. *Recommender Systems: Papers from the 1998 Workshop*, p. 60–64, 1998.

LACERDA, W.; BRAGA, A. Experimento de um classificador de padrões baseado na regra naive de bayes. *INFOCOMP Journal of Computer Science*, p. 60–64, 1998.

LAURITZEN, S.; SHEEHAN, N. A. Graphical models for genetic analyses. *Statistica Science*, v. 47, n. 4, p. 1235–1251, 2003.

LINS, A. M.; MICKA, K. A.; SPRECHER, C. J.; TAYLOR, J. A.; BACHER, J. W.; RABBACH, D.; BEVER, R. A.; CREACY, S.; SCHUMM, J. W. Development and population study of an eight-locus short tandem repeat (str) multiplex system. *Journal of Forensic Sciences*, v. 43, p. 1168–1180, 1998.

LUNA, J. E. U. *Algoritmos EM para Aprendizagem de Redes Bayesianas a partir de Dados Incompletos*. Dissertação (Mestrado) — Departamento de Computação e Estatística da Universidade Federal de Mato Grosso do Sul, 2004.

MANBER, U.; PATEL, A.; ROBISON, J. Experience with personalization of yahoo! *Commun. ACM*, ACM, New York, NY, USA, v. 43, n. 8, p. 35–39, 2000. ISSN 0001-0782.

MARSHALL, T. C.; SLATE, J.; KRUUK, L. E. B.; PEMBERTON, J. M. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, v. 7, p. 639–655, 1998.

MENDEL, G. Versuche über pflanzen-hybriden - vorgelegt in den sitzungen vom 8. Tradução para o inglês. 1865. Disponível em: <http://www.mendelweb.org>.

MONTANER, M.; LOPEZ, B.; LA, J. L. D. *Kluwer Academic Publishers. Printed in the Netherlands. A Taxonomy of Recommender Agents on the Internet*. 2003.

MORRIS, R. D. Bayesian research at the nasa ames research center. Computational Sciences Division. Acesso em 27 de agosto de 2008. 2003. Disponível em: <http://ti.arc.nasa.gov-/publications/pdf% -/0482.pdf>.

NAKANO, F. *Um Novo Modelo para Cálculo de Probabilidade de Paternidade - Concepção e Implementação*. Tese (Doutorado) — Universidade de São Paulo, 2007.

O'DONOVAN, J.; SMYTH, B. Trust in recommender systems. In: *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2005. p. 167–174. ISBN 1-58113-894-6.

PENA, S. D. Thomas bayes: O 'cara'! *Revista Ciência Hoje*, v. 38, n. 228, p. 22–29, 2006.

REAL, R. Redes bayesianas aplicadas a reconhecimento de spam. Acesso em 20 de junho de 2008. 2003. Disponível em: <http://www.inf.ufrgs.br/procpar/disc/cmp135/trabs/rodrigo/T2/html_spam-/spam.html>.

REATEGUI, E. B.; CAZELLA, S. C. Sistemas de recomendação. In: *Anais do Encontro Nacional de Inteligência Artificial. XXV CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO*. [s.n.], 2005. p. 306–348. Disponível em: <http://www.sbc.org.br/bibliotecadigital/download-.php?paper=415>.

RESNICK, P.; VARIAN, H. R. Recommender systems. *Commun. ACM*, ACM, New York, NY, USA, v. 40, n. 3, p. 56–58, 1997. ISSN 0001-0782.

RICCI, F.; ARSLAN, B.; MIRZADEH, N.; VENTURINI, A. A case-based travel advisory system. *Seventh European Conference on Case Based Reasoning*, Springer, Berlin, ALLEMAGNE, v. 2416, n. 6, p. 613–627, 1997.

RUSSEL, S. J.; NORVIG, P. *Inteligência Artificial*. 2. ed. [S.l.]: Campus, 2004.

SAHEKI, A. H. *Construção de uma Rede Bayesiana Aplicada ao Diagnóstico de Doenças Cardíacas*. Dissertação (Mestrado) — Escola Politécnica da Universidade de São Paulo USP, 2005.

SILVA, J. M. V. Teste de paternidade por análise de dna. 2001. Disponível em: <http://www.ufv.br-/dbg/BIO240/TP120.htm>.

SOUZA, R. R. *Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais*. Tese (Doutorado) — Escola de Ciência da Informação, Universidade Federal de Minas Gerais, 2005.

VALENTIM, F. L. Estudo e implementação de algoritmos de inferência e aprendizado em redes bayesianas. (Monografia) - Departamento de Ciência da Computação, Universidade Federal de Lavras. 2007. Disponível em: <http://www.dcc.ufla.br>.

WATSON, J.; CRICK, F. Molecular structure of nucleic acids. a structure for deoxyribose nucleic acid. *Nature*, p. 737–738, 1953.

YUDKOWSKY, E. An intuitive explanation of bayesian reasoning. Bayes' Theorem for the curious and bewildered; an excruciatingly gentle introduction. 2008. Disponível em: <http://yudkowsky.net/bayes/bayes.html>.

ZWEIG, G. G. *Speech Recognition with Dynamic Bayesian Networks*. Tese (Doutorado) — UNIVERSITY of CALIFORNIA, BERKELEY, 1998.

Apêndice A

Algoritmos de Inferência Bayesiana

A.1 Algoritmo Forward Sampling

O algoritmo segue alguns passos: Obtém-se estados para as variáveis sem pais, utilizando um sorteio de acordo com as *prioris*. A seguir o restante das variáveis vão assumindo um estado (sorteado) de acordo com estados dos seus pais, e suas probabilidades condicionais. Várias configurações são geradas e algumas são descartadas (aquelas que não correpondem aos estados das evidências). É feito a contagem de frequencia de uma variável X assumir um valor x ($X = x|e$). Assim obtém-se a fração de configurações do estado de consulta. Na figura A.1 segue o pseudo-código que está em (RUSSEL; NORVIG, 2004).

A.2 Algoritmo de Enumeração

O algoritmo de inferência por enumeração é baseado no cálculo do somatório em termos da distribuição conjunta total. Uma consulta de uma variável X dado as evidências pode ser obtida, e o funcionamento do algoritmo segue a equação A.1. A figura A.2 explicita o pseudo-código que está em (RUSSEL; NORVIG, 2004).

$$P(X|e) = \alpha P(X,e) = \alpha \sum_{y} P(X,e,y) \tag{A.1}$$

```
função FORWARD-SAMPLING(X, e, rb, N) retorna uma estimativa de P(X|e)

entradas: X, a variável de consulta
          e, evidência especificada como um evento
          rb, uma rede bayesiana
          N, o número total de amostras a serem geradas
variáveis locais: N, um vetor de contangens sobre X, inicialmente zero

para j = 1 até N faça
    x <- AMOSTRA-A-PRIORI(rb)
    se x é consistente com e então
        N[x] <- N[x] + 1 onde x é o valor de X em x

retornar NORMALIZAR (N[X])
```

Figura A.1: Algoritmo Fowarding Sampling

```
Função ASK-ENUMERAÇÃO (X, e, Rb) retorna uma distribuição sobre X
    entradas: X, a variável de consulta
    e, valores observados para variáveis E (evidência)
    Rb, uma rede bayesiana com variáveis {X} U E U Y /* Y = variáveis ocultas */

Q(X) <- uma distribuição X, inicialmente vazia
Para cada valor xi de X faça
    estender e com valor xi para X
    Q(xi) <- ENUMERAR-TODOS (VARS[rb], e)
Retornar NORMALIZAR(Q(x))

Função ENUMERAR-TODOS(vars, e) retorna um número real
    Se VAZIO(vars) então retornar 1,0
    Y <- PRIMEIRO(vars)
    se Y tem valor y em e
        então retornar P(y | pais(Y)) x ENUMERAR-TODOS (RESTO(vars), e)
        senão retornar Somatorio(P(y | pais(Y)) x ENUMERAR-TODOS(RESTO(vars), ey )
            onde ey é e estendido com Y = y
```

Figura A.2: Algoritmo de inferência por enumeração

Apêndice B

Algoritmos de Clustering

B.1 K-means

O algoritmo k-means segue basicamente os passos abaixo para k grupos e n elementos:

1. Seleciona-se os k primeiros elementos. Eles serão os centros de cada grupo. Portanto k deve ser o número de grupos que se deseja obter.

2. Para cada elemento $k+1$ até n, verifica em qual o centro mais próximo dele. Se for o segundo centro por exemplo, o elemento entra no grupo daquele centro.

3. Em cada grupo, obtém um novo centro, através da média aritmética daquele grupo.

4. Para todos os elementos, verifica em qual centro ele deve ficar, se ele vai deslocar para outro ou permanecer no mesmo.

5. Realiza o passo 3, e o 4 até não haver mais mudança de centro (até que os centros fiquem estáveis).

Os centros, nada mais são, que bons representantes de um grupo. A escolha no início dos k primeiros elementos poderia ser diferente escolhendo centros aleatórios (através de um sorteio).

Apêndice C

Sintaxe dos Arquivos de Entrada

Abaixo segue a sintaxe geral dos arquivos de entrada do aplicativo que foi construído para realizar o teste de paternidade.

C.1 Arquivo de Freqüencia Populacional, corresponde a Probabilidade a *Priori* (figura C.1)

Este arquivo é utilizado para entrar com os dados referente a freqüencia populacional de cada genótipo para cada *loci*. Por isso segue a sintaxe abaixo.

```
loci genotipo frequencia
```

C.2 Arquivo de Tabelas de Probabilidade Condicional (figura C.2)

O arquivo de tabelas de probabilidade condicional é utilizado para obter os dados que serão executados no algoritmo de inferência. Sua sintaxe é a seguinte:

```
loci gen_pai(1) gen_pai(2) G1=Pr(G1) G2=Pr(G2) ... Gn=Pr(Gn)
```

C.3 Arquivo de Entrada referente a família do exemplo 1 (figura C.3) e referente a família do exemplo 2 (figura C.4)

São os arquivos textos que representam as famílias dos indivíduos. Nas primeiras linhas do arquivo, iniciadas pelo símbolo '>' escreve-se os possíveis genótipos para cada *loci*. Logo após, cada membro da família é representado uma vez a cada *loci*. Os indivíduos que são pai biológico e suposto pai são marcados com PB e SP, respectivamente. Segue exemplos:

```
# exemplo de possíveis genótipos para o loci 1
> 1 AA Aa aa

# indivíduo 4 que possue os pais 1 e 2, e em relação ao loci 1 tem genótipo Aa
1 4 1 2 Aa

# indivíduo 3 que não possue pais, seu genótipo é desconhecido em relação ao loci 2,
# ele representa também o pai biológico
3 2 _ _ _ PB
```

C.4 Trecho de um arquivo XML gerado pelo programa é exemplificado na figura C.5

```
1 AA 0.442
1 Aa 0.247
1 aa 0.311

2 BB 0.485
2 Bb 0.265
2 bb 0.250

3 XX 0.185
3 Xx 0.465
3 xx 0.350

4 MM 0.34
4 MN 0.35
4 NN 0.31
```

Figura C.1: Arquivo texto que corresponde a probabilidade a *priori*

```
# ------------------------------------------------------------------------
# loci gen_pai(1) gen_pai(2) Pr(G1) Pr(G2) ... Pr(Gn)
# ------------------------------------------------------------------------

1 AA AA AA=1 Aa=0 aa=0
1 AA Aa AA=0.5 Aa=0.5 aa=0
1 AA aa AA=0 Aa=1 aa=0
1 Aa Aa AA=0.25 Aa=0.5 aa=0.25
1 Aa aa AA=0 Aa=0.5 aa=0.5
1 aa aa AA=0 Aa=0 aa=1

2 BB BB BB=1 Bb=0 bb=0
2 BB Bb BB=0.5 Bb=0.5 bb=0
2 BB bb BB=0 Bb=1 bb=0
2 Bb Bb BB=0.25 Bb=0.5 bb=0.25
2 Bb bb BB=0 Bb=0.5 bb=0.5
2 bb bb BB=0 Bb=0 bb=1

3 XX XX XX=1 Xx=0 xx=0
3 XX Xx XX=0.5 Xx=0.5 xx=0
3 XX xx XX=0 Xx=1 xx=0
3 Xx Xx XX=0.25 Xx=0.5 xx=0.25
3 Xx xx XX=0 Xx=0.5 xx=0.5
3 xx xx XX=0 Xx=0 xx=1

4 MM MM MM=1 MN=0 NN=0
4 MM MN MM=0.5 MN=0.5 NN=0
4 MM NN MM=0 MN=1 NN=0
4 MN MN MM=0.25 MN=0.5 NN=0.25
4 MN NN MM=0 MN=0.5 NN=0.5
4 NN NN MM=0 MN=0 NN=1
```

Figura C.2: Arquivo texto que corresponde a tabela de probabilidade condicional

```
> 1 AA Aa aa
> 2 BB Bb bb
> 3 XX Xx xx
> 4 MM MN NN
1 1 _ _ aa
1 2 _ _ _ PB
1 3 _ _ _ SP
1 4 _ _ aa
1 5 1 2 Aa
1 6 3 4 Aa
1 7 3 4 Aa
1 8 3 4 Aa

2 1 _ _ BB
2 2 _ _ _ PB
2 3 _ _ _ SP
2 4 _ _ Bb
2 5 1 2 BB
2 6 3 4 BB
2 7 3 4 Bb
2 8 3 4 BB

3 1 _ _ xx
3 2 _ _ _ PB
3 3 _ _ _ SP
3 4 _ _ Xx
3 5 1 2 Xx
3 6 3 4 xx
3 7 3 4 XX
3 8 3 4 xx

4 1 _ _ MN
4 2 _ _ _ PB
4 3 _ _ _ SP
4 4 _ _ MN
4 5 1 2 MM
4 6 3 4 MM
4 7 3 4 MM
4 8 3 4 MN
```

Figura C.3: Arquivo texto que corresponde a família do primeiro exemplo

```
> 1 AA Aa aa
> 2 BB Bb bb
> 3 XX Xx xx
> 4 MM MN NN

1 1 _ _ Aa
1 2 _ _ Aa
1 3 _ _ aa
1 4 _ _ _ PB
1 5 1 2 _ SP
1 6 3 4 aa

2 1 _ _ Bb
2 2 _ _ Bb
2 3 _ _ BB
2 4 _ _ _ PB
2 5 1 2 _ SP
2 6 3 4 Bb

3 1 _ _ XX
3 2 _ _ Xx
3 3 _ _ Xx
3 4 _ _ _ PB
3 5 1 2 _ SP
3 6 3 4 xx

4 1 _ _ MN
4 2 _ _ NN
4 3 _ _ NN
4 4 _ _ _ PB
4 5 1 2 _ SP
4 6 3 4 MN
```

Figura C.4: Arquivo texto que corresponde a família do segundo exemplo

```
<probabilidade_priori>

   <priori>
      <loci>1</loci>
      <genotipo>AA</genotipo>
      <valor>0.442</valor>
   </priori>

   <priori>
      <loci>1</loci>
      <genotipo>Aa</genotipo>
      <valor>0.247</valor>
   </priori>

   <priori>
      <loci>1</loci>
      <genotipo>aa</genotipo>
      <valor>0.311</valor>
   </priori>

   <priori>
      <loci>2</loci>
      <genotipo>BB</genotipo>
      <valor>0.485</valor>
   </priori>

   <priori>
      <loci>2</loci>
      <genotipo>Bb</genotipo>
      <valor>0.265</valor>
   </priori>

   <priori>
      <loci>2</loci>
      <genotipo>bb</genotipo>
      <valor>0.250</valor>
   </priori>
```

Figura C.5: Exemplo de trecho de arquivo XML gerado pelo aplicativo a partir de um arquivo de entrada de freqüencias genotípicas