Um estudo do comportamento de Redes Bayesianas no prognóstico da sobrevivência no cancro da próstata

Ana Cristina Lopes Sarabando

2010

Mestrado de Informática Médica
Faculdade de Ciências | Faculdade de Medicina
Universidade do Porto

Orientador: Prof. Doutora Inês de Castro Dutra, DCC/FCUP & CRACS, Universidade do Porto
Co-orientador: Doutor Nuno Afonso Gomes Costa Maia, Hospital Infante D. Pedro, Serviço de Urologia

Agradecimentos

Agradeço à Prof. Doutora Inês de Castro Dutra por todos os conselhos, críticas construtivas e sugestões que, em última instância levaram este trabalho a bom porto e ao Doutor Nuno Maia pela sua infindável disponibilidade e preocupação.

Agradeço também ao meu marido, Paulo Miguel de Jesus Dias, pelo constante apoio e partilha de conhecimento da área da informática e programação.

Agradeço a Dr. Cláudia Camila Dias, do Serviço de Bioestatística e Informática Médica da Faculdade de Medicina da Universidade do Porto, pelo apoio e interesse demonstrado em colaborar no esclarecimento de dúvidas na área de Bioestatística e Redes Bayesianas.

Agradeço à minha Chefe de Serviço, Enfª Áurea Simões pelo apoio sempre prestado e pela flexibilidade para me permitir terminar esta dissertação

Por último, agradeço aos meus pais e a todos os meus amigos pelo imensurável apoio que me proporcionaram ao longo de toda a minha carreira académica.

Sumário

Esta dissertação faz um estudo da aplicação de redes bayesianas ao prognóstico da sobrevivência no cancro de próstata.

Com o auxílio de um médico especialista em cancro de próstata, construímos uma rede bayesiana a partir de dados clínicos de doentes. Aplicamos esta rede a um conjunto de dados de doentes, disponíveis no endereço http://lib.stat.cmu.edu/S/Harrell/data/descriptions/prostate.html (Andrews and Herzberg 1985), com o objectivo de estudar e avaliar a qualidade desta rede construída manualmente. A seguir, utilizamos algumas ferramentas de geração de redes bayesianas a fim de comparar a qualidade destas redes geradas automaticamente com a rede gerada manualmente. Os resultados mostram que as redes geradas automaticamente comportam-se tão bem ou melhor que a rede construída manualmente, além de apresentarem relações causais não existentes na rede gerada manualmente.

Neste trabalho utilizamos *software* tais como o SPSS, o WEKA, o GeNIe, o Mini-TUBA e o Aleph/SAYU na construção das redes manual e automáticas.

Fizemos um intenso estudo do estado da arte com o intuito de estabelecer uma base sólida para o trabalho. Foram procurados os métodos já existentes para prognóstico em doentes com cancro da próstata e observados os seus resultados, com o intuito de construir um instrumento útil e com resultados viáveis para a área médica em questão.

A mortalidade de doentes com cancro está a aumentar em todo o mundo e é necessário aproveitar os dados disponíveis para melhorar os cuidados de saúde e possivelmente reduzir a taxa de mortalidade por cancro.

Não tendo ainda sido desenvolvidos muitos estudos nesta área, pode considerar-se que este trabalho apresenta um impulso ao estudo e aplicação das redes bayesianas estáticas e dinâmicas na área médica e conduz, em última instância a um instrumento que consideramos útil para o prognóstico de doenças como o cancro da próstata.

Abstract

This thesis aims to study the use of bayesian networks in the prognostic of survival in prostate cancer.

With the help of a doctor, specialist in prostate cancer, we built a bayesian network based on clinical data from real patients. We applied the resulting network to a dataset, available at http://lib.stat.cmu.edu/S/Harrell/data/descriptions/prostate.html (Andrews and Herzberg 1985), to study and evaluate the quality of this manually built network. We also used several *software* tools to generate other bayesian networks in order to compare the results of these automatically generated networks with the manual one. Results shows that automatically generated networks behave as well or even better that the manually built one. They also present causal relations that are not available in the manually generated network.

In this work, we used several software applications, such as SPSS, WEKA, GeNIe, Mini-TUBA and Aleph/SAYU to build automatic and manual bayesian networks.

We also provide an extensive state of the art to provide a good starting point. We searched the already existing method used in prostate cancer and analysed their results. Our objective is to build a tool with viable results that may be used in this specific medical area in the future.

With the increase in the mortality of cancer patients, it seems important to use all the available data to improve healthcare and even reduce the cancer mortality rates.

Despite the lack of studies in this field, this work is a good step in the study and development of static and dynamic bayesian networks in the medical field and provides a useful tool for the prognostic of diseases such as prostate cancer.

Preâmbulo

Atul Gawande, cirurgião americano no Hospital de Boston e Professor na Faculdade de Medicina em Harvard, conta-nos: "Um dia, um cirurgião cardíaco, responsável pela área dos cuidados coronários, entrou no seu gabinete com 2240 electrocardiogramas (ECG). Ele demorou uma semana a examiná-los e separar aqueles em que achava que o doente estava a ter um ataque cardíaco na altura do registo." (Gawande 2002).

Um ECG é uma verdadeira aventura para um estudante de medicina, e a sua interpretação só melhora com a prática. Uma avaliação humana, mesmo de um especialista, falha com alguma facilidade.

Surge então a possibilidade de que um computador possa auxiliar a tarefa do ser humano, de forma a baixar a taxa de erros e optimizar a sua intervenção. Um sistema computacional pode ajudar um profissional da área da saúde de diversas formas. Neste trabalho, vamos concentrar-nos no apoio à decisão clínica utilizando técnicas de Inteligência Artificial, mais especificamente, em métodos de aprendizagem de máquina baseados em classificadores bayesianos ou probabilísticos.

Em 1990, William Baxt descreveu como uma rede neuronal artificial pode tomar decisões clínicas. O computador pode aprender com a experiência, tal como o ser humano. Na história de Gawande, o computador venceu o cardiologista ao classificar melhor 20% a mais de ECG's.

Na medicina os erros médicos são inaceitáveis, uma radiografia tem que ser bem interpretado, a dose de uma medicação não pode falhar, e nada sobre cada doente pode ser esquecido como alergias ou patologias associadas. Um cirurgião deve ser perfeito, não desperdiçando tempo, movimentos ou gota de sangue.

Esta perfeição médica é difícil de atingir, e só a rotina e a repetição podem ajudar. Com a repetição muitas funções mentais tornam-se automáticas e necessitam de menos esforço.

Muitos médicos podem considerar que cada doente é um caso diferente e que não há nada melhor para obter um diagnóstico do que a avaliação humana. Mas será mesmo assim? Quando um doente entra na urgência com uma dor abdominal e um cirurgião

tenta descartar a hipótese de apendicite aguda, ele não segue uma lista de procedimentos semelhante a todos os doentes?

Um algoritmo pode facilmente aproximar-se de um julgamento humano a fazer prognósticos ou diagnósticos. Pode-se então concluir que o ser humano em conjunto com o computador obterão claramente melhores resultados. Muitos estudiosos consideram que não, porém Paul Meehl, David Faust e Robin Dawes que estudaram diversas comparações de diagnósticos matemáticos e julgamento humano, constataram que a estatística venceu ou equiparou-se ao julgamento humano em todos os seus estudos (Dawes, Faust et al. 1989; Meehl 1996).

Índice

Índice de figuras

Índice de tabelas

Organização da tese

Este documento encontra-se estruturado da seguinte forma:

No capítulo 1 contextualizamos o nosso trabalho, definimos os seus objectivos e introduzimos o problema a ser abordado.

No Capítulo 2 é feito o estudo do estado da arte no que concerne a redes bayesianas e a sua aplicação no prognóstico de doenças. É feita uma descrição das redes bayesianas estáticas e dinâmicas.

O capítulo 3 apresenta uma descrição das ferramentas utilizadas neste trabalho.

O capítulo 4 descreve o processo de modelagem qualitativa e o trabalho efectuado com a base de dados para poder aplicar o algoritmo de redes Bayesianas. No mesmo capítulo descrevemos o processo de modelagem quantitativa com apoio do *software* GeNIe, com apoio do *software* WEKA, com apoio do *software* Aleph e com apoio do *software* mini-TUBA.

No Capitulo 5 apresentamos o problema médico e a rede inicial construída manualmente.

No Capitulo 6 apresentamos uma avaliação das redes sob os pontos de vista quantitativo e qualitativo.

O Capítulo 7 é dedicado às conclusões onde são debatidos os pontos-chave do trabalho bem como analisadas as reais potencialidades da aplicação de redes bayesianas ao prognóstico de sobrevivência no cancro da próstata. São ainda debatidas perspectivas futuras para sistemas deste género no último capítulo.

1 Introdução

Em 2001, (Sim, Gorman et al. 2001) referem que os Sistemas de Apoio à Decisão Clínica (SADC) melhoram a qualidade dos cuidados de saúde substancialmente. Os mesmos concluem que os SADC são pouco utilizados na prática clínica diária dos profissionais de saúde ou que os resultados que advêm dos SADC não são aplicados. Esta tendência tem vindo a alterar-se devido à grande variedade de áreas clínicas que são abrangidas pelos SADC e à nova posição adoptada pelos profissionais de saúde na utilização de sistemas informáticos na prática clínica.

A quantidade de dados médicos guardados digitalmente tem vindo a crescer de forma exponencial. A organização e selecção destes dados é uma necessidade, o que incentiva a procura de formas rápidas e lógicas de o fazer. Actualmente recorre-se a sistemas que agem racionalmente (de raciocínio probabilístico, estatístico ou técnicas de optimização) na gestão de dados médicos, para resolver problemas de diversas áreas como diagnóstico ou prognóstico. Alguns destes sistemas empregam métodos de aprendizagem de máquina (do inglês *machine learning*) para "aprender" algum modelo para os dados.

A aplicação de *machine learning* ao diagnóstico e detecção de tumores malignos é frequente, principalmente no tratamento de imagem médica e na área celular.

Existem três focos de preocupação nesta área: prever a susceptibilidade de desenvolver cancro, prever a possibilidade de um cancro tratado recidivar, e prever a capacidade de sobrevivência. O último foco, o que vamos explorar neste trabalho, procura prever um resultado (*outcome*) como esperança de vida, sobrevivência, progressão da doença e sensibilidade à terapêutica, após o diagnóstico da doença.

Para fazer esta previsão, utilizamos redes bayesianas. Estas são muito utilizadas e populares por serem classificadores probabilísticos. Dado um conjunto de observações, que no contexto deste trabalho, são os dados clínicos de cada doente, o que uma rede bayesiana permite fazer é atribuir uma probabilidade a um determinado doente de ter ou não esperança de vida, dada a sua informação clínica.

Considerando os valores dos dados clínicos de doentes, construímos uma rede causal, com o auxílio de um especialista em cancro de próstata, utilizando a ferramenta GeNie (GeNIe Visited 2010), especifica para construção e teste de redes bayesianas e que possui uma interface gráfica para o desenho da rede. Além de avaliar esta rede, também utilizamos ferramentas para geração automática de redes bayesianas: WEKA (Mark, Eibe et al. 2009), MiniTUBA (Xiang, Minter et al. 2007), e SAYU (Srinivasan 2001; Davis, Burnside et al. 2005).

Utilizámos o conjunto de dados disponível em http://lib.stat.cmu.edu/S/Harrell/data/descriptions/prostate.html (Andrews and Herzberg 1985) como fonte para treino dos modelos de rede e para teste da nossa rede.

Houve vários aspectos a contemplar na aplicação das ferramentas ao conjunto de dados, pois este necessitou de vários processos de tratamento dos dados para cada ferramenta.

Os resultados mostram que para o problema médico em questão, as redes geradas automaticamente comportam-se tão bem ou melhor que a rede construída manualmente, além de apurarem relações causais não existentes na rede gerada automaticamente.

Resumindo, podemos dizer que redes bayesianas são um método simples, mas poderoso de auxiliar o prognóstico de doentes com cancro da próstata, permitindo melhorar as escolhas do tratamento para cada caso específico. Detalhou-se a evolução do cancro da próstata em vários doentes, permitindo melhorar o tratamento de futuros casos.

2 Estado de arte

2.1 Redes Bayesianas

Na Medicina encontramos diferentes situações que englobam lidar com um grande leque de informação, informação essa que necessita de uma avaliação rigorosa para se poder atingir os objectivos pretendidos, sejam eles a gestão de um serviço ou uma decisão terapêutica. Os sistemas de apoio à decisão clínica permitem agrupar toda a informação existente e retirar dela informação pertinente. As redes bayesianas são modelos que integram sistemas de apoio à decisão clínica e permitem capturar informação e lidar com situações de incerteza. Estas são úteis em situações de incerteza, pois determinam uma probabilidade de um certo acontecimento dependente das variáveis e das suas relações. A representação gráfica das relações de causalidade das variáveis do sistema permite uma visualização simplificada do sistema.

Uma rede Bayesiana pode ser definida como uma rede probabilística. É constituída por uma estrutura gráfica e uma distribuição de probabilidades a ela associada. Matematicamente podemos definir como B = (Pc,G) onde G é um grafo acíclico direccionado e Pc as probabilidades condicionais associadas a cada variável probabilística representada em um nó do grafo. O grafo é um grafo acíclico com nós e arcos.

Ao referirmo-nos ao "Markov Blanket" de um nó X, estamos a designar o nó X, os seus pais, os seus filhos e os outros pais dos seus nós filhos.

Cada nó representa uma variável que toma um valor dentro de um intervalo finito e cada arco representa uma relação probabilística entre as variáveis. Para cada variável do grafo estão definidas as probabilidades condicionais. Cada uma destas probabilidades descreve o efeito de uma combinação específica de valores.

Uma probabilidade condicional (Pc) é a probabilidade de determinado evento X ocorrer sabendo que ocorreu o evento Y, representado matematicamente por Pc(X|Y).

O Teorema de Bayes relaciona as probabilidades dos dois acontecimentos X e Y com as suas probabilidades condicionais mútuas, isto é:

$$P(X|Y) = P(Y|X)\frac{P(X)}{P(Y)} \qquad \text{(Teorema de Bayes)}$$

Os grafos não só permitem ao utilizador perceber quais as variáveis que influenciam outras variáveis, como também são o suporte para o cálculo computacional das probabilidades condicionais que são necessárias para os resultados ou aprendizagem. A leitura e interpretação dos grafos são bastante intuitivas e torna fácil a interpretação dos resultados, o que significa que, ao contrário de outros métodos tais como as redes neuronais que são consideradas "caixas negras", as redes bayesianas permitem uma clara interpretação semântica dos modelos.

As redes bayesianas permitem-nos obter resultados com um bom intervalo de confiança mesmo quando trabalhamos com grande quantidade de informação. São também modelos muito robustos que permitem pequenas alterações sem afectar a sua performance.

Outra vantagem das redes bayesianas, que é explorada neste trabalho, é o facto de permitir conjugar o conhecimento científico com os dados estatísticos de uma forma simples.

Normalmente, as redes bayesianas não representam eventos temporais, somente factos ocorridos em uma determinada data. Porém para alguns problemas pode ser importante representar mudanças de valores de uma variável ao longo do tempo. Isto é particularmente importante em ambientes médicos, onde o doente é acompanhado e pode ter seu quadro clínico alterado ao longo do tempo.

Já existem alguns sistemas desenvolvidos que permitem a introdução de informação colhida em diferentes momentos temporais. Estes modelos denominam-se de Redes Bayesianas Dinâmicas (DBN's). Em 2007, Marion Verduijn et al publicaram uma aplicação para prognóstico de situações em doentes sujeitos à cirurgia cardíaca – ProCarSur, utilizando DBN's (Verduijn, Rosseel et al. 2007). Esta aplicação lida com um vasto leque de variáveis e com vários valores para a mesma variável, valores estes que são registados em diferentes momentos temporais. Neste exemplo consideraram-se os seguintes momentos: colheita de dados no pré-operatório, dados durante a

cirurgia, dados das primeiras 24h na Unidade de Cuidados Intensivos pós-cirúrgicos e dados de todo o internamento pós-operatório.

As Redes Bayesianas Dinâmicas representam um poderoso método de modelo probabilístico para identificar padrões causais ou aparentes em bases de dados temporais.

A utilização de DBN's permite ao investigador identificar parâmetros significativos e relações entre eles. Elas são particularmente úteis quando há um grande número de variáveis e as suas relações são complexas, pois como já foi referido, uma das vantagens das redes bayesianas é o facto de obterem bons resultados mesmo lidando com grande número de dados.

Nas DBN podemos obter uma corrente causal de variáveis, isto é, uma sequência de eventos necessários para produzir um determinado resultado. Com outras técnicas de "machine learning" dificilmente se obtêm estas correntes causais.

As DBN também podem incluir conhecimento médico do especialista pois é possível forçar ou ignorar relações entre as variáveis.

Redes Bayesianas Dinâmicas são, assim como as redes bayesianas estáticas, um tipo de modelo de prognóstico. Os modelos de prognóstico, em saúde, *"são ferramentas que prevêem o resultado de uma doença ou do tratamento dessa mesma doença*" (Verduijn, Peek et al. 2007) num período de tempo futuro. As DBN's permitem-nos, além de prever o resultado, visualizar a influência das variáveis nesse mesmo resultado ao longo do tempo, prevendo o resultado individual de uma acção. Fornecendo-lhe dados específicos de alguns doentes podem obter um possível resultado de um acto médico ou de uma doença no futuro, mas para tal necessitamos de informação acerca do ciclo da doença do doente e das opções tomadas para tratamentos ao longo do processo.

Segundo vários autores (Kononenko 1993; Chickering, Heckerman et al. 1997; Ghahramani 1998; Lahdesmaki and Shmulevich 2008; van Gerven, Taal et al. 2008) as DBN's trazem vantagens quando comparadas com outros métodos como os processos de decisão de Markov, redes neuronais, ou árvores de decisão, que permitem também a construção de modelos de prognóstico em medicina. Alguns destes modelos (em especial as redes neuronais) funcionam como "caixas pretas", isto

é, não permitem ao utilizador saber como atingiram uma determinada conclusão de prognóstico.

Em relação aos outros modelos, as DBN's podem ser usadas para representar conhecimento médico descrito em causas e efeitos que obtêm dos dados, conhecimentos de especialistas, e/ou literatura. Nestas redes não são feitas suposições restritivas acerca da interacção entre variáveis. (Ghahramani 1998; van Gerven, Taal et al. 2008).

As DBN's representam um método poderoso para identificar possíveis padrões causais em conjuntos de dados heterogéneos e complexos, com tolerância para grande variabilidade de dados e não linearidade. Segundo Van Gerven (2008), pelas suas vantagens, por ser uma generalização do processo de decisão de Markov e por não restringir as relações entre as variáveis, as DBN's permitem resolver algumas das limitações actuais dos modelos de prognóstico (van Gerven, Taal et al. 2008).

Os mesmos autores elucidam-nos sobre os passos necessários de forma a obter um modelo de prognóstico satisfatório de Redes Bayesianas Dinâmicas. Este modelo permite obter uma rede de relações entre variáveis que emergem ao longo do tempo escolhido para o caso em estudo. (Verduijn, Peek et al. 2007) apresentam um modelo de aprendizagem da rede.

A aprendizagem é uma das características que define os sistemas baseados em inteligência artificial. É difícil definir "aprendizagem" pelo que a maioria dos artigos concorda que é uma característica dos sistemas adaptativos que são capazes de melhorar o seu comportamento em função da sua experiência passada, por exemplo, resolver problemas semelhantes. A aprendizagem da rede é um processo que permite a optimização do modelo através da descoberta de novos padrões a partir de novos dados, mais extensos e complexos.

O desenvolvimento de redes bayesianas Dinâmicas e a aprendizagem destas mesmas redes remonta a 1997, com Chickering et al, mas sem aplicação a dados em concreto, apenas a visão matemática dos métodos. (Chickering, Heckerman et al. 1997; Ghahramani 1998).

A aprendizagem em redes bayesianas inicia-se sobre um modelo pré-existente elaborado por conhecimento prévio de especialistas, pelos dados ou ambos (*prior knowledge*). Este conhecimento prévio é representado na forma de uma distribuição

de probabilidade numa estrutura do modelo e parâmetros. Esta representação é melhorada utilizando novos dados de forma a obter uma nova distribuição de probabilidade sobre a estrutura do modelo e os parâmetros.

Existem também métodos de aprendizagem dos parâmetros e da estrutura da rede bayesiana e métodos de aprendizagem através de dados incompletos.

A aprendizagem é uma combinação de "prior knowledge" com novos dados e obtém melhorias no conhecimento (melhora as probabilidades da estrutura de uma rede bayesiana pré-existente). Segundo Cruz obter sucesso com a aprendizagem do modelo nem sempre é garantido. Em qualquer método utilizado é importante existir um bom conhecimento do problema e das limitações dos dados disponíveis assim como das hipóteses e limitações do algoritmo aplicado. (Cruz and Wishart 2007)

2.2 Escolha do problema médico

O cancro é a causa líder de morte no mundo e o número total de casos está, de forma global, a aumentar. O número global de mortes por cancro é previsto aumentar 45% de 2007 até 2030 (de 7.9 milhões para 11.5 milhões de mortes), influenciado em parte por um aumento e envelhecimento da população global (WHO visited 2010).

Quem vive diariamente num serviço de saúde apercebe-se deste aumento exponencial de casos que surgem para tratamento de cancro, seja cirúrgico, farmacológico, curativo ou paliativo. Os casos aumentam, a variabilidade é grande entre cada caso, mas a doença começa a ser uma ameaça cada vez mais estudada e previsível.

Vários cancros poderiam ser problemas propostos e aplicados, mas a escolha do problema levou em consideração aspectos práticos e de utilidade. Esta foi feita a partir de considerações sobre a praticidade de se construir um sistema especialista modelando o problema e também sobre os possíveis benefícios que um sistema traria à área médica envolvida. Os critérios adoptados são descritos a seguir:

a) Nível de complexidade do problema médico

A obtenção de informação médica é um tema muito complexo que coloca em causa diversos problemas como a segurança da informação, a confidencialidade, etc. Quanto mais complexo for o problema médico, mais difícil seria disponibilizarem uma base de dados com informação de doentes. Como já foi referido neste trabalho, o cancro é

a causa líder de morte no mundo e o número total de casos está, de forma global, a aumentar, influenciado em parte por um aumento e envelhecimento da população global.

Os casos aumentam, a variabilidade é grande entre cada caso e a doença começa a ser uma ameaça cada vez mais estudada e previsível, pelo que este tema seria pertinente e de maior facilidade em encontrar informação.

b) Possível benefício que o sistema de decisão traria ao problema

Este item considerou os possíveis benefícios em termos financeiros e de ganhos em saúde com um sistema de apoio à decisão clínica. Os ganhos em saúde podem-se referir a anos de vida, melhora de qualidade de vida, ou ganhos monetários obtidos com a escolha de melhores procedimentos médicos.

c) Disponibilidade do médico especialista que auxiliou na construção do sistema

Foi necessário contactar com um médico especialista com interesse em conhecer sistemas de apoio à decisão clínica e que mostrasse disponibilidade e abertura para auxiliar durante a construção do sistema.

2.3 Utilização das redes bayesianas para o prognóstico de sobrevivência no cancro da próstata

A aplicação de métodos de tratamentos de dados e de *machine learning* no cancro não é novidade. As primeiras aplicações foram no apoio à detecção e diagnóstico e iniciaram-se há aproximadamente 25 anos (finais da década de 80) no diagnóstico de cancro da mama e do sistema digestivo (Graham, Paplanus et al. 1990; Wolberg, Street et al. 1994).

Kononenko em 1993 refere a aplicação de sistemas de aprendizagem de árvores de decisão e do classificador Naive Bayes em problemas de diagnóstico (localização de tumores primários, recorrência do cancro da mama, tumores da tiróide e reumatologia) (Kononenko 1993).

O objectivo de obter prognósticos no cancro é diferente do objectivo da detecção e diagnóstico. No prognóstico é medicamente útil saber informações dos riscos a que o

doente pode estar sujeito, conhecer a probabilidade de recorrência do cancro e qual a probabilidade de sobrevivência de um doente ao longo do tempo.

Os modelos de prognóstico podem prever doenças e seus tratamentos. São importantes para ajudar no tratamento da grande quantidade de dados que surgem actualmente de doentes que sofreram de cancro e que ao longo da doença viveram diferentes tratamentos e diferentes respostas obtendo de diferentes formas e tempos o mesmo resultado, a morte. A morte é um importante resultado pois é o fim clínico do processo de cuidar.

Como já foi referido, existem alguns sistemas desenvolvidos para aplicação de DBN's em dados médicos, como por exemplo o ProCarSur (doentes sujeitos a cirurgia cardíaca) (Verduijn, Rosseel et al. 2007). Em 2007 Zuoshuang Xiang *et al* desenvolvem também o mini-TUBA, um sistema de modelação na internet que permite a investigadores clínicos e biomédicos obterem inferências clínicas e antecipações utilizando DBN's com dados temporais (Xiang, Minter et al. 2007).

Utilizando o ProCarSur, Verduijn *et al* obtêm uma DBN de dados clínicos de doentes sujeitos a cirurgia cardíaca usando um método de aprendizagem da rede a partir de modelos locais. Ele considera que se deve determinar um *Markov Blanket* apropriado para cada variável, permitindo assim seleccionar o melhor subgrupo em características preditivas para cada variável. Através dos modelos locais é possível representar a probabilidade de distribuição de cada variável determinando quais os seus pais na rede. A rede final é obtida através dos dados por uma estratégia de busca local (Verduijn, Peek et al. 2007).

A rede final obtida por Verduijn et al para cirurgia cardíaca desenvolvida através de modelos locais é uma rede bayesiana que não permite aprendizagem (Figura 1). A aprendizagem desta rede é aplicada na sua elaboração, isto é, a rede não surge somente de *prior knowledge* nem somente de dados, mas de uma combinação dos dois.

A Figura 1 mostra um exemplo de DBN para cirurgia cardíaca, onde cada nível corresponde aos diferentes tempos de recolha de dados dos doentes em causa e mostra quais as variáveis medidas em cada momento.

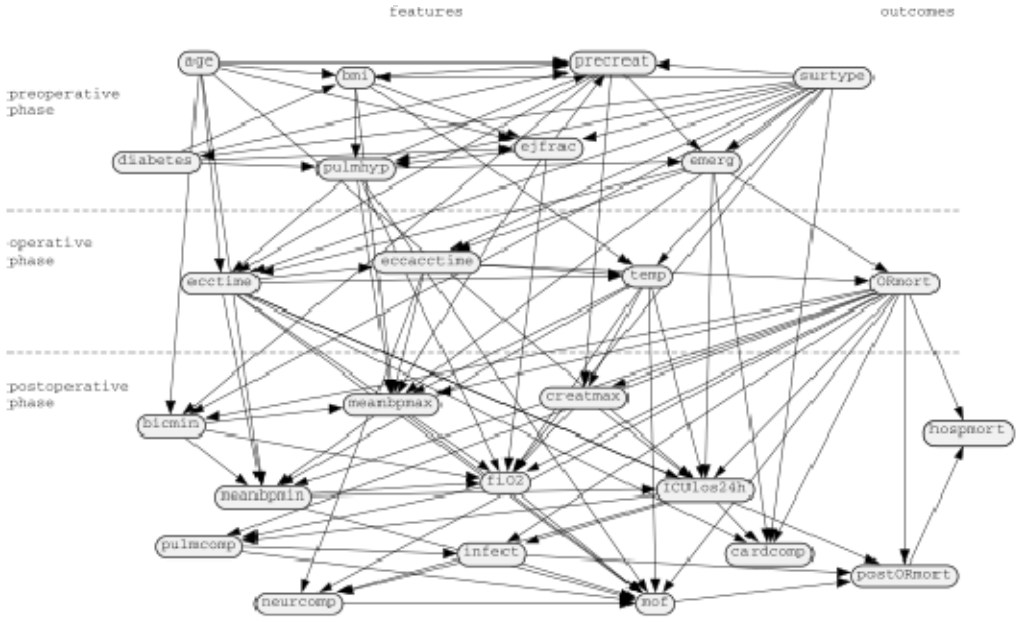

Figura 1: DBN para cirurgia cardíaca desenvolvida através de modelos em árvore locais (Verduijn, Peek et al. 2007)

Perante a pesquisa efectuada, tornou-se evidente a utilidade de "machine learning" de forma a permitir aos computadores "aprender" através de exemplos passados e a detectar padrões em grandes quantidades de dados complexos. Esta aplicação no prognóstico do cancro não é recente na área de diagnóstico, mas é recente e ainda pouco explorada na previsão da sobrevivência.

É importante contribuir para auxiliar o médico na tomada de decisão no que respeita ao tratamento do cancro tendo em conta os resultados que advêm de tais decisões.

Se existirem opções com probabilidades para os diferentes resultados que podem surgir consoante o tratamento escolhido e as características do caso clínico, o médico mais facilmente utiliza o seu conhecimento científico em decisões ponderadas matematicamente.

As redes bayesianas são alvo de vários estudos e artigos publicados. São as que apresentam mais vantagens como modelos de prognóstico apesar de a maioria dos estudos na previsão da sobrevivência do cancro aplicarem como método de "machine learning" as redes neuronais artificiais (ANN's) (Cruz and Wishart 2007).

3 Material e Métodos

A pesquisa bibliográfica efectuada para o desenvolvimento deste trabalho incidiu sobre os seguintes termos: dynamic bayesian networks, prognostic, cancer, learning Bayesian networks.

A pesquisa foi efectuada nas seguintes bases de dados: Medline, PubMed, Google Scholar, Scopus e ScienceDirect (editora Elsevier).

Os artigos encontrados foram filtrados por tema (incluindo só os da área da saúde, bioinformática e/ou medicina). Os termos em questão incluíam: "dynamic bayesian networks", "prognostic and cancer", "learning Bayesian networks", "machine learning and cancer prediction".

A relevância de cada artigo foi discutida tendo em conta os títulos e os resumos. Foram identificados alguns artigos que tenham utilizado métodos de redes bayesianas dinâmicas e dados na área da saúde (moleculares, clínicos, histológicos, epidemiológicos). Foram seleccionados também os que determinam prognóstico no cancro.

No Scopus em particular foram encontrados 14 artigos após a filtragem. Desses 14, 8 são relacionados com genética, 4 são tentativas matemáticas de melhorar uma rede bayesiana dinâmica e 2 são aplicações de redes bayesianas dinâmicas com dados médicos em situações de diagnóstico/prognóstico.

Esta pesquisa demonstrou que ainda há muita investigação a desenvolver na aplicação de redes bayesianas Dinâmicas na área da saúde. Em artigos de 1995 inicia-se a abordagem de assuntos sobre modelos de prognóstico na área da saúde, mas sem aplicações clínicas ou aplicações de redes bayesianas. Os artigos encontrados nesta filtragem iniciam-se em 2005 até aos dias de hoje (2010).

Alguns artigos utilizados para o desenvolvimento deste trabalho e aprofundamento de alguns temas, foram seleccionados tendo em conta as referências bibliográficas de outros artigos.

3.1 O software – open source

Neste trabalho, optamos por utilizar apenas ferramentas livres e *open source*. Desta forma, o *software* apresentado nesta secção pertence a esta classe.

3.1.1 GeNIe

GeNIe é a interface gráfica para o SMILE.

SMILE (Structural Modeling, Interface, and Learning Engine) é uma biblioteca de classes de C++ que implementa métodos gráficos de decisões. A sua interface gráfica é o GeNIe, um ambiente de desenvolvimento amigável para gráficos de modelos de decisão.

Ambos foram desenvolvidos no Laboratório de Sistemas de Decisão na Universidade de Pittsburgh, disponíveis para a comunidade desde Julho de 1998.

GeNIe é um *software* que pode ser utilizado para criar modelos de decisão intuitivos utilizando uma interface gráfica de click-and-drop.

3.1.2 WEKA

WEKA é um *software* disponível online (também *open source*), desenvolvido na Universidade de Waikato na Nova Zelândia. Este *software* permite implementar diversos algoritmos sobre bases de dados, assim como possui diversas ferramentas que permitem transformar a base de dados (por exemplo, discretizar as variáveis contínuas, função que foi utilizada e descrita no Capítulo 5.2).

Neste estudo pretendeu-se explorar a aplicação dos classificadores de redes bayesianas, e vários estão implementados no WEKA. Aplicaram-se estes classificadores sobre a base de dados utilizada em todo este trabalho.

Todos os algoritmos de redes bayesianas implementados no WEKA assumem que a base de dados cumpre os seguintes requisitos: todas as variáveis são discretas e finitas e não existem valores em falta. No final obtiveram-se duas redes pois utilizaram-se os dados discretizados pelos dois processos descritos no Capítulo 5.2.

A aprendizagem de uma rede bayesiana implica naturalmente dois passos: primeiro aprender a estrutura da rede e depois as tabelas de probabilidades.

3.1.3 Mini-TUBA

MiniTUBA (*medical inference by network integration of temporal data using Bayesian analysis*) é uma ferramenta disponível na Web que permite a investigadores clínicos e/ou biomédicos obterem uma análise bayesiana dinâmica utilizando bases de dados temporais. Com o MiniTUBA é possível explorar a forma como dados médicos recolhidos ao longo do tempo podem ser utilizados em inferência estatística.

Foi desenvolvido por Z.Xiang et al, *Center for Computational Medicine and Biology, University of Michigan,* Ann Arbor, MI, EUA.

Os utilizadores são capazes de actualizarem a base de dados continuamente e melhorar os resultados. Mini-TUBA oferece prognósticos e sugestões de intervenções baseado num processo automático de aprendizagem utilizando todos os dados disponíveis.

A utilização do mini-TUBA permite obter uma DBN a partir de dados inseridos pelo utilizador, configurada consoante os resultados que se pretende obter (definindo os parâmetros da rede bayesiana). Esta DBN permite ao utilizador obter inferências sobre a contribuição de uma intervenção sobre um caso específico.

A Figura 2 ilustra o processo de aprendizagem das redes bayesianas dinâmicas utilizadas no mini-TUBA. O *software* permite ao utilizador fazer login (com um login geral ou pedindo autorização para ter uma conta privada) de forma a disponibilizar a introdução de novos dados para um novo estudo, ou a alterar os parâmetros de um estudo já existente de forma a calcular nova rede. Toda a informação introduzida pelo utilizador fica guardada numa base de dados MySQL que é lida pelo *software* de forma a realizar uma análise bayesiana. Os resultados são apresentados na página do estudo do utilizador, permitindo a qualquer momento alterar os dados ou os parâmetros de cálculo.

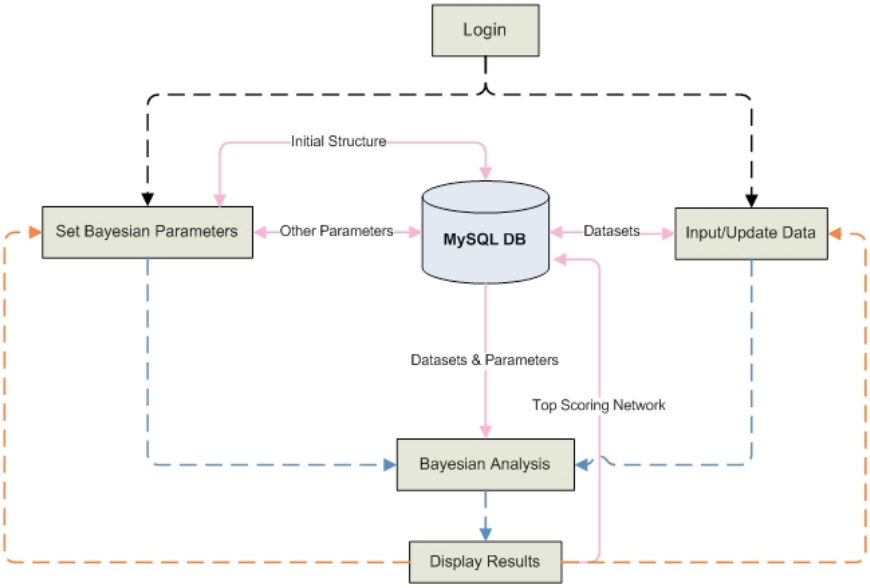

Figura 2: mini-TUBA workflow (Xiang, Minter et al. 2007)

A vantagem desta aplicação consiste na facilidade em conseguir uma aprendizagem da rede, bastando para tal introduzir mais dados na base de dados. A dificuldade surge com grande quantidade de dados, pois o cálculo da rede de um projecto mini-TUBA, no máximo de dados que ele permite, pode demorar até 144 h (o que representa um investimento computacional significante). Outra desvantagem é que a aprendizagem da rede implica o cálculo de uma nova rede. Por outro lado este sistema permite facilidade de acesso à mesma (disponível na Web, ver exemplo de interface na Figura 3) e a possibilidade de aplicação do *software* em qualquer base de dados introduzida pelo utilizador.

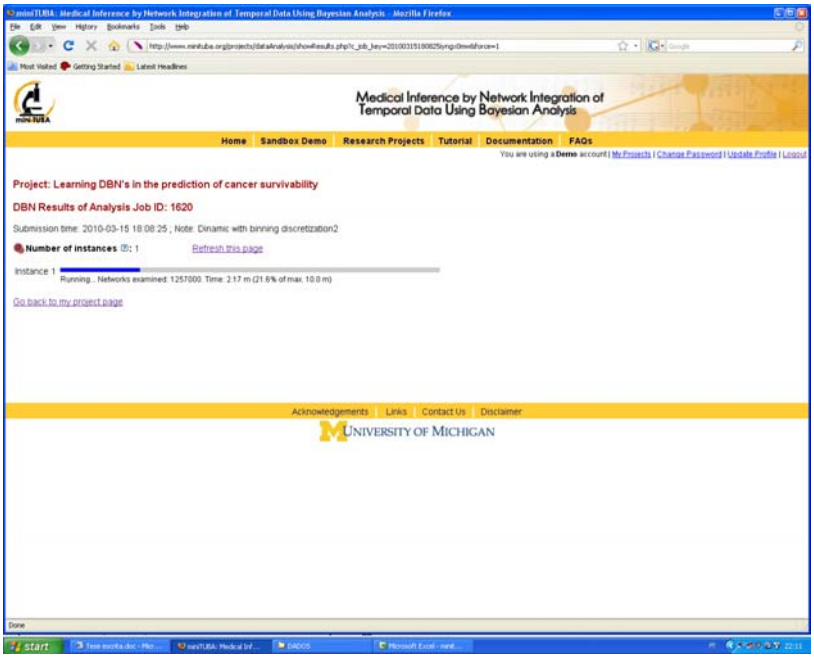

Figura 3: mini-TUBA na Web (Xiang, Minter et al. 2007)

3.1.4 Aleph e SAYU

As ferramentas apresentadas nas secções 3.1.1, 3.1.2 e 3.1.3 são consideradas de aprendizagem proposicional. Nesta secção apresentamos uma ferramenta de aprendizagem estatística relacional que combina regras de primeira ordem com redes bayesianas. Para a aprendizagem das regras utilizamos o sistema Aleph (Srinivasan 2001) e para a integração das regras numa rede bayesiana utilizamos o sistema SAYU (Davis, Burnside et al. 2005). Aleph foi desenvolvido em Prolog por Ashwin Srinivasan, na Universidade de Oxford, Inglaterra. SAYU foi desenvolvido por Davis et *al*, na Universidade de Wisconsin, Madison, EUA.

Dados um conjunto de exemplos positivos e negativos e suas descrições em uma linguagem de primeira ordem, e uma linguagem de restrições, Aleph busca encontrar o melhor modelo para os exemplos através da construção de regras de primeira ordem. SAYU (*Score As You Use*) utiliza estas regras para construir uma rede bayesiana. Em outras palavras, SAYU avalia cada regra de acordo com sua contribuição probabilística na rede construída utilizando esta regra como variável.

4 Modelo Qualitativo da rede Bayesiana

4.1 Processo para criação do modelo da rede

Para a criação do modelo estrutural da rede, em conjunto com o especialista da área, foram necessárias diversas reuniões que permitissem a partilha de conhecimento médico.

Nestas reuniões, procurou-se identificar quais as informações relevantes, relativas ao problema médico, que poderiam ser representadas como variáveis na rede.

Esta informação cedida pelo especialista foi cruzada com fontes bibliográficas e foram encontradas as variáveis relevantes assim como as relações entre elas.

A estrutura obtida inicialmente necessitou de diversos ajustes que foram realizados ao longo das reuniões com o especialista.

4.2 A escolha das variáveis

A escolha das variáveis foi realizada tendo em conta três diferentes momentos temporais do desenvolver da doença, que se iniciam com a colheita de dados do doente e terminam 5 anos depois da cirurgia de Prostatectomia Radical (PR). Esta divisão temporal veio facilitar a selecção das variáveis de uma forma mais organizada, permitindo também a sua posterior utilização em modelos de redes bayesianas dinâmicas.

Os três momentos temporais são:

(1) o momento da colheita de dados do doente

(2) quando se obtêm os resultados dos exames auxiliares de diagnóstico realizados

(3) 5 anos após cirurgia de prostatectomia radical

As variáveis de cada momento foram obtidas de uma forma criteriosa. Inicialmente o especialista propôs as que achava relevante derivado da sua experiência médica, e posteriormente foi confrontado com a opinião de autores de forma a completar o seu raciocínio clínico.

Joseph A. Cruz (Cruz and Wishart 2007) refere algumas variáveis que diversos autores consideram pertinentes no prognóstico do cancro (em geral) como a história familiar, a idade, a dieta, o peso, hábitos de vida de risco (fumador ou alcoólico), exposição a ambientes carcinogénicos, etc. Colozza et *al* refere também a importância de conjugar estes dados familiares com dados clínicos como resultados de exames ou dados moleculares, pois o avanço da tecnologia tem-nos permitido obter cada vez mais pormenores de cada doença e de cada doente particularmente (Colozza, Cardoso et al. 2005).

Na área específica do cancro da próstata, (Revett, Magalhães et al. 2006) consideram que factores como a dieta, a hereditariedade, factores ambientais que afectam as hormonas masculinas observados em estudos epidemiológicos, podem ser a causa da doença. Os mesmos autores consideram dois factores importantes para a detecção precoce do cancro da próstata: o toque rectal e os valores do *Prostate Specific-Antigen (PSA),* por outro lado colocam a idade como um importante factor de incidência. No seu artigo, (Revett, Magalhães et al. 2006), obtêm um conjunto de variáveis que consideram relevantes como causadoras do cancro da próstata, através da investigação de um conjunto de dados médicos com informação clínica de doentes com cancro da próstata, utilizando métodos de *machine learning*.

Alan W. Partin et al (Partin, Mangold et al. 2001) surgem com as *Partin's Tables*, tabelas que permitem predizer o estadio patológico da doença através da conjugação de variáveis clínicas pré-operatórias tais como o valor do PSA, *Gleason Stage* e o estadio clínico. Os resultados obtidos com a utilização destas tabelas podem facilitar a decisão terapêutica para um determinado doente.

Os dados obtidos a partir da realização de biopsia prostática transrectal ecoguiada também são factores determinantes de decisão terapêutica. Um relatório completo deste exame pode conter os seguintes dados: valor de PSA, volume da próstata, aspecto das vesículas seminais, presença de nódulos hipoecogénicos e localização e em termos histológicos o *Gleason Score* e a presença ou ausência de penetração da cápsula.

A conjugação do conhecimento extraído nestas pesquisas com a experiência do especialista permitiu-nos obter as variáveis finais que se podem ver na Tabela 1.

Tabela 1: Variáveis

Nome da variável	Abreviatura
Idade	idade
Peso	wt
História familiar de cancro	hx_familiar_ca
International Prostate Symptom Score	IPSS
Pressão arterial sistólica	sbp
Pressão arterial diastólica	dbp
Hemoglobina	hemoglobina
Nódulos hipoecogénicos	nodulos_hipoecogenicos
Prostate specific-antigen	PSA
Estadio Clínico	estadio_clinico
Gleason Stage (ouu Gleason Score)	gleason_stage
Penetração da cápsula prostática	penetração_capsula
Envolvimento das vesiculas seminais	env_vesiculas_sem
Envolvimento linfático	env_linfatico
Doubling time PSA	doubling_time_psa
Tamanho da próstata	volume_prostata
Metástases ósseas	metastases_osseas
Status após 5 anos da cirurgia	sobrevivência_5anos

4.3 As variáveis

4.3.1 Idade

Refere-se à idade do doente no momento em que é realizada a colheita de dados iniciais, na primeira consulta médica de Urologia. A idade é medida em anos e a escala utilizada neste estudo é a seguinte:

<= 40

41-50

51-60

61-70

71-80

81-90

>90

A utilização desta escala com intervalos de 10 anos foi propositada. Existem diversos estudos que comprovam que a idade avançada influência o prognóstico do cancro da próstata, assim, esta divisão permite tirar conclusões mais pormenorizadas.

4.3.2 Peso

Refere-se ao peso do doente no momento em que é realizada a colheita de dados iniciais, na primeira consulta de Urologia. O peso é medido em Kilogramas (Kg) e a escala utilizada neste estudo é a seguinte:

<=80

80-88

>88

A utilização desta escala foi baseada no Índice de Massa Corporal (IMC). Uma pessoa com peso superior a 88 kg e uma estatura média (considerando uma estatura média de 1,70 m), é considerada pelo IMC no limite para obesidade de grau I.

4.3.3 História familiar de cancro

Diversos estudos apoiam a teoria de que a presença de cancro em outros membros da família é um elevado factor de risco (genético) para a pessoa vir a desenvolver cancro.

A escala utilizada neste estudo é a ausência ou presença de casos de cancro na família.

4.3.4 International Prostate Symptom Score

International Prostate Symptom Score é um questionário que foi desenvolvido e validado pelo comité multidisciplinar da Associação Americana de Urologia para poder classificar o risco de aumento do volume prostático através dos sintomas do doente (Barry, Fowler et al. 1992).

O questionário tem 7 perguntas que envolve a frequência urinária, nocturia, fraco jacto urinário, intermitência, hesitação, incompleto esvaziamento da bexiga e urgência urinária. A escala é a seguinte:

0-7 Ligeiramente sintomático

8-19 Moderadamente sintomático

20-35 Severamente sintomático

4.3.5 Pressão arterial

A pressão arterial foi dividida em duas variáveis, pressão arterial sistólica e pressão arterial diastólica. Esta divisão permite detectar mais facilmente uma pressão arterial alta, pois basta uma delas estar acima do valor considerado normal. Outro motivo para esta divisão prende-se com o facto de esta variável estar assim apresentada no conjunto de dados utilizado neste estudo. A pressão arterial é medida em mmHg e a escala utilizada é a seguinte:

Pressão arterial sistólica:

<14

>=14 – Pressão arterial alta

Pressão arterial diastólica:

<9

>=9 – Pressão arterial alta

4.3.6 Hemoglobina

A hemoglobina é medida através de análises clínicas efectuadas ao doente. Esta variável surge em diferentes tempos desde que o doente tem a primeira consulta até 5 anos após a cirurgia de prostatectomia radical, pois em diversos momentos são

realizadas análises clínicas ao doente. Os momentos pertinentes para este estudo são: as primeiras análises efectuadas, as pós-cirurgicas e as finais (5 anos após cirurgia).

Os valores de hemoglobina são medidos em gr/dl e a escala utilizada neste estudo é a seguinte:

<12,5

12,5-18

>=18

A utilização desta escala justifica-se porque, segundo estudos laboratoriais, os valores normais de hemoglobina numa pessoa saudável se situam entre os 12,5 gr/dl e os 18 gr/dl.

4.3.7 Nódulos hipoecogénicos

A presença de nódulos hipoecogénicos na próstata está relacionado com a presença de cancro da próstata, mas não é específico. A ecografia prostática é muito sensível na detecção destes nódulos, e como tal obtêm bastantes resultados falsos positivos. A probabilidade de encontrar cancro da próstata em doentes com nódulos hipoecogénicos detectados na ecografia varia de 15% a 37% (Pontes, Ohe et al. 1984).

4.3.8 *Prostate Specific-Antigen* (PSA)

O PSA é uma proteína produzida pela glândula prostática. O seu valor é medido através de análises ao sangue. Níveis altos de PSA no sangue estão directamente relacionados com alterações prostáticas, que podem ser benignas ou malignas (cancro da próstata).

Os valores de PSA são medidos em ng/ml e a escala utilizada neste estudo é a seguinte:

< = 4

> 4

A utilização desta escala é baseada no estudo efectuado por Thompson et *al.*, que verifica a maior probabilidade de incidência de cancro da próstata em homens com valores de PSA >4 ng/ml (Thompson, Pauler et al. 2004).

4.3.9 Estadio Clínico

A determinação do estadio clínico envolve observação clínica, toque rectal e estudos auxiliares de diagnostico (como por exemplo a Eco, a TAC, a RM, a cintigrafia óssea, …). Através destes exames determina-se os estadios clínicos T1, T2, T3 e T4, sendo o T1 o menos agressivo e o T4 o mais agressivo (já com doença disseminada).

4.3.10 Gleason Stage

O *Gleason Stage* é determinado pelo patologista, na observação microscópica dos fragmentos de tecido retirados durante a realização de uma biopsia prostática. Estes fragmentos de glândula prostática são examinados e o seu padrão histológico classificado consoante os dois tipos de células presentes em maior quantidade. A escala de Gleason varia de 1 a 5 consoante a arquitectura estrutural das células glandulares que estaria dependente de maior ou menor grau de diferenciação celular e a sua estrutura tecidular. O Gleason Stage ou score resulta da soma dos dois principais padrões encontrados. Esta escala para o cancro da próstata foi desenvolvida pelos Dr. Gleason e Mellinger em 1974 como um método para prever o comportamento do cancro da próstata (Gleason and Mellinger 1974).

4.3.11 Penetração da cápsula prostática

Esta variável poderá ser inferida no momento da realização de exames auxiliares de diagnóstico como a ecografia prostática ou ressonância magnética pélvica mas é apenas confirmada pelo estudo histológico de peça de prostatectomia radical.

A escala utilizada para este estudo é a presença ou ausência de penetração de células tumorais na cápsula prostática.

4.3.12 Envolvimento das vesículas seminais

As vesículas seminais são duas glândulas adjacentes à próstata e que produzem o liquido seminal. Quando as células tumorais atingem as vesículas seminais significa cancro da próstata localmente avançado e doença mais agressiva com maior probabilidade de metastização à distância.

A escala utilizada para este estudo é a presença ou ausência de envolvimento das vesículas seminais no cancro da próstata.

4.3.13 Envolvimento linfático

À semelhança do envolvimento das vesículas seminais, a presença de células tumorais nos gânglios linfáticos ilíacos, é sinal de mau prognóstico por indicar doença já metastizada.

A escala utilizada para este estudo é a presença ou ausência de envolvimento dos gânglios ilíacos no cancro da próstata.

4.3.14 *Doubling time* PSA

PSA *doubling time (PSADT)* é o tempo que os valores de PSA demoram até atingir o dobro. É calculado com a fórmula:

$$PSADT = \log(2)\frac{dT}{(\log(B) - \log(A))}$$

onde A e B são as medidas iniciais e finais de PSA e dT é o tempo entre as duas medidas.

O PSADT após prostatectomia radical está directamente relacionado com a mortalidade específica por cancro da próstata. A escala de discretização foi seleccionada tendo em conta a divisão efectuada em outros estudos realizados (Teeter, Presti et al. 2009). O *doubling time* PSA é medido em meses e a escala é a seguinte:

<3

3-8,9

9-14,9

> 15

4.3.15 Tamanho da próstata

Uma próstata com um volume superior a 20 gr, significa uma próstata de volume, contudo anormal. Será um aumento benigno (hiperplasia benigna da próstata). Apenas nos casos de doença avançada localmente o volume da próstata poderá ser maioritariamente resultante do crescimento maligno. A determinação do volume da próstata utiliza normalmente a fórmula da elipse da próstata:

$$Vp = \frac{\pi}{2} \times (diâmetro\_transverso \times diâmetro\_anterior\_posterior \times diâmetro\_cefalocaudal)$$

Esta é a fórmula mais usada em urologia para determinar o volume aproximado da próstata. O tamanho da próstata é medido em gramas (gr) e a escala utilizada é a seguinte:

<20

>=20

4.3.16 Metástases ósseas

O local mais comum de disseminação do cancro da próstata é o osso. A presença ou não de metástases ósseas no momento do diagnóstico é um elemento fundamental na decisão do tratamento.

As metástases ósseas são um factor de mau prognóstico. Dos doentes com cancro da próstata que desenvolvem metástase, 50% morrem num prazo de 30 meses. Num estudo verificou-se que na autópsia de doentes que morreram de cancro da próstata avançado, as metástases ósseas estavam presentes em 80% a 90% dos casos.

A escala utilizada para este estudo é a presença ou ausência de metástases ósseas.

4.3.17 Status após 5 anos da cirurgia

Esta variável é o *outcome* do nosso conjunto de variáveis. Um dos objectivos deste estudo é procurar relações entre as variáveis e de que forma estas mesmas influenciam a mortalidade por cancro da próstata cinco anos após prostatectomia radical.

A escala utilizada para este estudo consistiu no estado de doente morto ou vivo, cinco anos após a cirurgia.

4.4 Limitações na escolha das variáveis

Algumas variáveis que foram consideradas pertinentes não puderam ser incluídas. Estas variáveis estão relacionadas com alterações genéticas que poderão ser indicativas de doentes que sofrem de cancro da próstata e têm que ser sujeitos a uma prostatectomia radical. Como referia Colozza et *al* (Colozza, Cardoso et al. 2005), o avanço tecnológico já nos permite obter dados mais aprofundados das doenças. Era importante para este estudo englobar a investigação genética e poder acrescentar parâmetros como alterações cromossómicas, o que não foi possível pois não há ainda investigação suficiente neste campo, nem dados disponíveis para utilização em estudos posteriores.

4.5 Métodos para a estruturação da rede

A estruturação da rede, visando obter um modelo que melhor representasse o cancro da próstata e a sua projecção no doente, e que fosse de fácil compreensão, foi feita através dos seguintes métodos:

- Inserção ou remoção de arcos de forma a representar as relações causais existentes entre as variáveis no problema médico.
- Fusão de nós, isto é, nós mais específicos foram trocados por nós mais genéricos.
- Divisão de nós. O processo oposto à fusão de nós, isto é, nós genéricos foram trocados por nós mais específicos.

- Remoção de nós por serem desnecessários ou de difícil arranjo no restante da rede (nós livres na rede).
- Inserção de nós que não haviam sido considerados anteriormente.

Nenhuma variável foi considerada mais importante do que outra e a estrutura final da rede foi aquela que representasse de forma mais clara e de fácil compreensão a real relação entre as variáveis.

4.6 Evolução dos modelos construídos

A construção da estrutura da rede foi feita manualmente e de forma iterativa, exigindo diversas reuniões até se obter uma rede final. Para um melhor resultado foi necessário conjugar os dados do conhecimento médico com a bibliografia científica.

O principal objectivo deste modelo é pré-seleccionar as variáveis significativas numa rede que possa avaliar o prognóstico de doentes com cancro da próstata sujeitos a prostatectomia radical, e posteriormente relacioná-las entre si.

A partir de sessões de esclarecimento com o médico especialista, foram realizadas algumas alterações, assim como a conexão dos nós da rede.

Neste modelo, os nós da rede foram divididos em 3 momentos temporais diferentes. Esta divisão não altera em nada as relações entre as variáveis nem o resultado final da rede, mas vem facilitar a construção de uma rede dinâmica numa fase posterior deste estudo.

1º Momento: entrevista com o doente na primeira consulta (fase pré-diagnóstico)

2º Momento: após se obter os resultados dos exames auxiliares de diagnóstico (fase pós-diagnóstico)

3º Momento: após o doente ter sido sujeito a prostatectomia radical e se obter o resultado de alguns exames realizados no pós-operatório (fase pós-operatório).

A estrutura final da rede é visível na Figura 4.

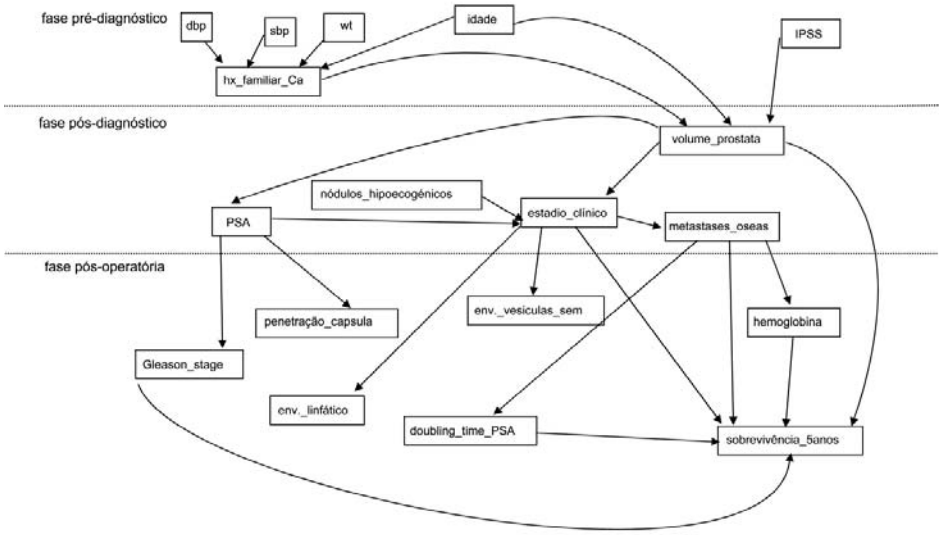

Figura 4: Modelo final da rede

4.7 Conclusão

Os atributos presentes nesta rede são os que melhor identificam um doente com cancro da próstata. Se pudéssemos ter disponível uma base de dados com valores para estes atributos, e de preferência recolhidos ao longo de 5 anos após o diagnóstico da doença, seria o ideal para poder obter uma rede bayesiana (estática ou dinâmica) que nos fornecesse informação relevante para o tratamento destes doentes.

Infelizmente neste trabalho isso não foi possível, e para podermos continuar com o seu desenvolvimento utilizámos uma base de dados disponível em http://lib.stat.cmu.edu/S/Harrell/data/descriptions/prostate.html (Andrews and Herzberg 1985), que descrevemos no capítulo seguinte.

As relações entre as variáveis foram escolhidas pelo especialista tendo em conta a sua experiência, e são posteriormente comparadas com as obtidas nas redes bayesianas construídas mais à frente neste trabalho.

5 Modelos Quantitativos

Para uma utilização deste modelo em processos de inferência, de forma a permitir obter probabilidades de ocorrência de certos fenómenos, foi necessário trabalhar a base de dados que contém informação de 502 doentes a quem foi diagnosticado cancro da próstata. Esta base de dados está disponível para estudos no repositório http://lib.stat.cmu.edu/S/Harrell/data/descriptions/prostate.html (Andrews and Herzberg 1985).

Esta base de dados inicialmente continha 18 atributos, incluindo o atributo de decisão (*outcome*), com 27 valores em falta (*missing values*). Esta foi adaptada tendo em conta a sua aplicação.

Foram retirados do modelo final da rede (Figura 4) os atributos que não estavam presentes na base de dados e foram retirados da base de dados os atributos que não estavam presentes na rede. Esta redução de atributos torna a base de dados mais fácil de trabalhar e dá ênfase às características de classificação dos dados.

O modelo necessita da aplicação das regras que os algoritmos de Redes Bayesianas impõem. Para tal, o conjunto de dados médicos utilizado na construção da rede deve obedecer a algumas particularidades que são: todas as variáveis devem ser discretas e finitas e não devem existir valores em falta (*missing values*).

5.1 Gestão dos atributos

Para a construção dos modelos quantitativos seleccionamos da base de dados 11 dos 18 atributos que ela contém. Como já referimos atrás, só utilizámos os atributos da base de dados que também estão presentes na selecção feita para o modelo final da rede (Figura 4).

As variáveis que foram utilizadas nas Redes Bayesianas obtidas neste estudo são as que estão apresentadas na Tabela 2.

Tabela 2: Atributos seleccionados da base de dados

Nome da variável	Abreviatura
Idade	age
Peso	wt
História familiar de cancro	hx
Pressão arterial sistólica	sbp
Pressão arterial diastólica	dbp
Hemoglobina	Hg
Estadio Clínico	stage
Doubling time PSA	dtime
Tamanho da próstata	size
Metástases ósseas	bm
Status após 5 anos da cirurgia	status

Na Figura 5 podemos observar uma distribuição matricial dos valores da base de dados dispostos por atributo e por doente. Esta figura permite uma visualização dos dados e a sua distribuição dos valores sem necessidade de recorrer à extensa base de dados de 502 doentes, que que aumentaria demasiado o número de páginas desta dissertação.

Nesta Figura, cada ponto significa um exemplo e cada cor corresponde ao valor do atributo classe, que no nosso caso, pode ser *dead* ou *alive.* Cada quadrado apresenta um cruzamento de um par de atributos e como os exemplos se distribuem através destes atributos pelos valores da classe. Além de mostrar a relação entre os atributos dos 502 pacientes, esta figura ilustra que não é trivial extrair um padrão baseado num único atributo ou conjunto de atributos que possa ser indicativo de sobrevivência ou não.

O *outcome* Status que pode ter o valor *dead* (= 0) ou *alive* (=1) possui 354 doentes mortos e 148 vivos.

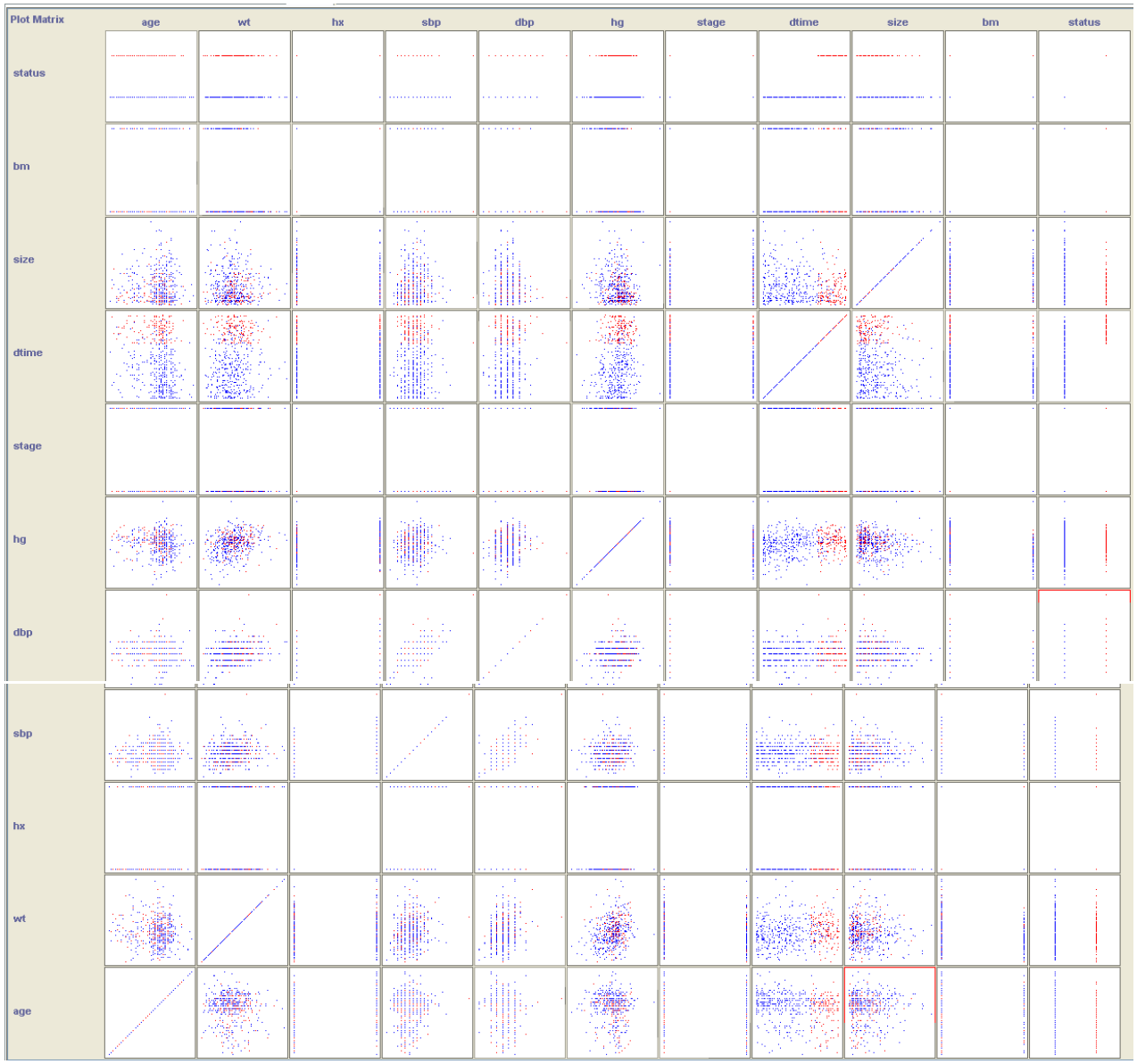

Figura 5: Matriz representativa dos valores de cada atributo para cada doente ("scatterplot" obtido com o *software* WEKA)

5.2 Necessidade de discretização das variáveis contínuas

Os algoritmos utilizados em sistemas de aprendizagem de máquina, normalmente fazem a sua própria discretização interna no caso dos atributos com valores numéricos. Como existe um padrão de referência clínico para vários destes atributos como descrito na sub-secção 4.4., fizemos um primeiro filtro para transformar todos os atributos numéricos nos intervalos clínicos apropriados.

Mantivemos contudo os dados originais com o intuito de observar se os classificadores gerados produziam outro tipo de intervalo. Este estudo pode ser importante para uma posterior revisão dos intervalos clínicos actualmente empregues como padrão de referência.

Foram utilizados dois processos de discretização das variáveis. No primeiro processo foram escolhidos os intervalos baseados em conhecimento científico, para que possa representar da melhor forma a realidade, permitindo assim resultados mais reais e uma melhor interpretação dos resultados na área médica (intervalos na Tabela 3).

O segundo processo para a discretização das variáveis foi utilizar um filtro para discretização no WEKA. Através da aplicação do filtro *weka.filters.unsupervised.attribute.Discretize*, discretizou-se um grupo de atributos numéricos da base de dados em atributos nominais. Esta discretização é realizada através de *binning* simples (neste trabalho optou-se por agrupar em 5 intervalos).

Tabela 3: Intervalos de discretização. A justificação para a escolha destes intervalos está descrita aquando a apresentação de cada variável (Subsecção 4.4.).

Variável	Intervalo de discretização
Idade (age)	<= 40 41-50 51-60 61-70 71-80 81-90 >90
Peso (wt)	<=80 80-88 >88
História familiar (Hx)	Sim Não
Pressão arterial sistólica (Sbp)	<14 >=14
Pressão arterial diastólica (Dbp)	<9 >=9
Hemoglobina (Hg)	<12,5 12,5-18 >=18
Estadio clinico (Stage)	1 2 3 4
Tamanho da prostata (Size)	<20 >=20
Metastases osseas (bm)	Sim Não
Doubling Time (dtime)	<3 3-8,9 9-14,9 > 15
Status	Dead Alive

5.3 Missing values

Os valores em falta na base de dados podiam comprometer o resultado que viríamos a obter nas diferentes redes (estática e dinâmica).

Para podermos construir uma rede bayesiana, foi necessário preencher os valores em falta das variáveis que utilizámos, pois um dos critérios para uma rede bayesiana é que não devem existir *missing values*.

Poderíamos ter utilizado várias alternativas para o preenchimento destes dados, mas segundo Howell (Howell 1999), *Expectation/Maximization* (EM) é o algoritmo referido como o mais importante na literatura recente. Para tal utilizámos este algoritmo através do *software* SPSS e foi possível reestimar os parâmetros em falta na base de dados.

O SPSS permite efectuar uma análise da base de dados em relação aos dados em falta, conseguindo um diagnóstico da situação em análise (Tabela 4), permitindo-nos obter alguns resultados estatísticos como o número de amostras (N) em cada variável, a média dos seus valores e o desvio padrão. Também nos indica o número de valores em falta em cada variável que foram introduzidos utilizando o algoritmo EM.

Tabela 4: Estatistica da aplicação do algoritmo EM

				Faltas	
	N	Média	Desvio padrão	contagem	Percentagem
age	501	71,46	7,081	1	20
wt	500	99,03	13,436	2	40
sz	499	14,69	12,338	3	60

5.4 Modelo I - Modelo construído manualmente

Para obter uma rede final neste modelo, utilizou-se a rede construída manualmente pelo especialista.

O passo seguinte à criação da estrutura de uma Rede Bayesiana é a especificação das suas probabilidades. Essas probabilidades podem ser obtidas de duas maneiras: a partir de estudos de especialidade ou inferindo de uma base de dados. É possível também a combinação das duas alternativas. O objectivo final é sempre obter as probabilidades que melhor correspondam à realidade do problema modelado pela Rede Bayesiana.

Para a quantificação da Rede Bayesiana do modelo final obtido, foi utilizada a inferência através da base de dados.

A estrutura inicial da rede obtida a partir de estudos de especialidade foi usada como ponto de partida para a construção das tabelas de probabilidades. Foi necessário construir uma tabela para cada nó, com campos para preenchimento das probabilidades correspondentes à combinação das categorias de todos os pais. O preenchimento das tabelas fez-se utilizando o programa Excel por inferência da base de dados utilizada neste estudo.

5.4.1 Probabilidades marginais

As probabilidades marginais são as mais simples de se obter e correspondem a nós sem pais. Estas probabilidades correspondem à prevalência da doença na população em questão, isto é, nos 502 doentes com cancro da próstata.

As tabelas 5 a 8, representam as probabilidades marginais da rede.

A primeira e segunda linha da tabela representam as categorias do nó, e a terceira a sua prevalência para cada categoria

Tabela 5: Probabilidades marginais para Pressão Arterial Sistólica (sbp)

Sim	não
P(sbp>=14)	P(sbp<14)

63.3%	36.7%

Tabela 6: Probabilidades marginais para Pressão Arterial Diastólica (dbp)

Sim	não
P(dbp>=9)	P(dbp<9)
35.5%	64.5%

Tabela 7: Probabilidades marginais para Idade

P(id<=40)	P(41>id>50)	P(51>id>60)	
0.0%	1.0%	9.6%	
P(61>id>70)	P(71>id>80)	P(81>id>90)	P(id>91)
19.7%	65.9%	3.8%	0.0%

Tabela 8: Probabilidades marginais para Peso

normal	acima	obeso
P(wt<=80)	P(80<wt<=88)	P(wt>88)
6.8%	14.1%	79.1%

5.4.2 Probabilidades condicionais

Uma distribuição de probabilidade condicional refere-se à probabilidade de ocorrer um evento A sabendo que ocorreu um outro evento B, e representa-se por P(A|B). Os nós da rede que têm pais, têm probabilidades condicionais em relação aos respectivos pais.

Temos sempre presente uma combinação de distribuições de probabilidades associada a uma estrutura gráfica de uma Rede Bayesiana. Para cada variável (Vi) no gráfico é especificado um conjunto de distribuições de probabilidades condicionais P(Vi|π(Vi)) sendo π(Vi) os pais de Vi; cada uma destas distribuições descreve o efeito conjunto de uma específica combinação de valores para os pais de cada variável Vi, na distribuição de probabilidades, sobre os valores de Vi.

Uma das grandes dificuldades na aplicação prática de redes bayesianas (quando calculadas manualmente) é o excessivo número de probabilidade necessário para a quantificação de uma rede. Quanto maior o número de pais de um nó, mais complexa e extensa é a tabela de probabilidades condicionais. Os cálculos foram efectuados tendo em conta:

$$P(V_i|\pi(V_i) = \frac{P(V_i \cap \pi(V_i))}{P(\pi(V_i))} \text{ se } P(\pi(V_i)) > 0$$

O teorema de Bayes diz que:

$$P(V_i|\pi(V_i) = \frac{P(\pi(V_i)|V_i) \times P(V_i)}{P(\pi(V_i))} \text{ se } P(\pi(V_i)) > 0$$

5.4.3 Modelo final da rede - GeNIe

O modelo final da rede foi construído no GeNIe, *software* disponível online gratuitamente. Foram desenhados os nós da rede (pais e filhos) com as mesmas relações da rede construída manualmente em conjunto com o especialista. Em cada nó é possível introduzir as tabelas de probabilidades marginais e condicionais. Podemos ver o aspecto final da rede na Figura 6.

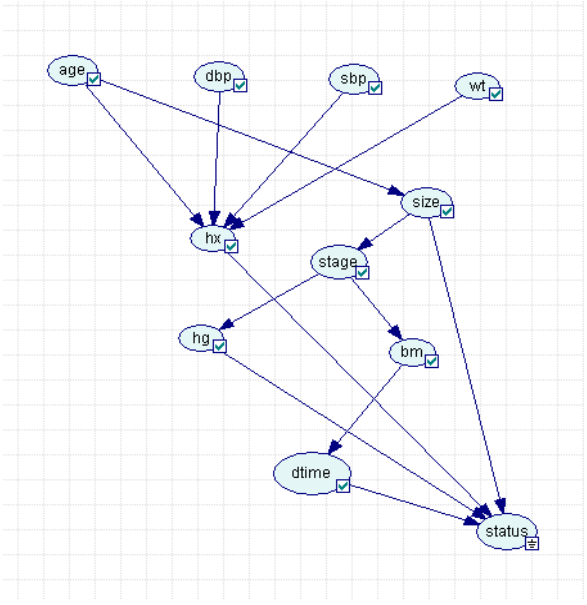

Figura 6: Aspecto final da rede construída no GeNIe

5.4.4 Determinação da exactidão da rede

Foram realizadas diversas tentativas de forma a podermos testar os 502 casos de doentes na rede da Figura 6, mas em nenhum *software* conseguimos introduzir a rede no formato que o GeNIe nos oferece.

A única forma que encontrámos foi de testar manualmente no GeNIe caso a caso.

Foram escolhidos aleatoriamente 10 doentes da base de dados e foram testados no GeNIe. Em 10 doentes, 7 foram correctamente classificados. Logo podemos considerar que a rede manual apresenta uma exactidão de 70%, com pouca segurança pois uma amostra de 10 doentes num total de 502, é uma amostra pouco significativa.

5.5 Modelo II - WEKA – Machine Learning algorithms in Java

5.5.1 Aprendizagem da estrutura da rede

A aprendizagem de uma rede bayesiana implica naturalmente dois passos: primeiro aprender a estrutura da rede e depois as tabelas de probabilidades.

O primeiro passo apresenta diversas formas de resolução no Weka. Neste trabalho utilizámos a técnica de *cross-validation* (com 10 *folds*), pois permite prever o comportamento da rede no futuro (com outros dados) através da avaliação da exactidão da classificação obtida.

Para obtermos a estrutura da rede utilizamos o classificador *BayesNet*, indicado no Weka como: *weka.classifiers.bayes.BayesNet*. Este algoritmo tem algumas opções que podemos configurar (Figura 7).

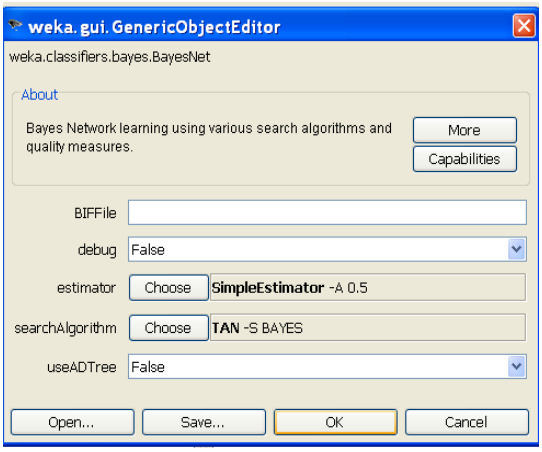

Figura 7: Opções de configuração do algoritmo BayesNet

BIFFile – Podemos colocar o nome de um ficheiro no formato BIF XML. A rede bayesiana gerada pelos dados pode ser comparada com a rede bayesiana representada no formato BIF XML. As estatísticas calculadas são o número de ligações extra e em falta.

debug – Se colocado TRUE, o classificador pode fornecer informação adicional

estimator – O algoritmo escolhido encontra as tabelas de probabilidades condicionais da rede bayesiana

searchAlgorithm – Selecciona o método usado para encontrar a estrutura da rede

useADTree – Quando TRUE, o tempo de aprendizagem da rede diminui. Quando FALSE, o algoritmo para a aprendizagem estrutura da rede fica mais lento e funciona com menos memória. Por defeito ADTree é TRUE.

Existem vários métodos para encontrar a estrutura de uma rede bayesiana, e no weka estes algoritmos estão divididos por áreas: *local score metric, conditional independence tests, global score metrics e fixed structure*. Para cada uma destas áreas podem ser implementados diferentes algoritmos como por exemplo: *hill climbing, simulated annealing e tabu search.*

O algoritmo que pretendemos usar neste trabalho para obter a estrutura da rede é o TAN (*Tree Augmented Naive Bayes*), pois é o que nos permite obter uma rede com uma estrutura de melhor visualização das relações entre variáveis. A rede (ou árvore como é denominada neste algoritmo) é formada através do cálculo da árvore de extensão de peso mínimo utilizando o algoritmo Chow-Liu (Chow and Liu 1968). Obtemos assim um grafo dirigido acíclico; isto é, para qualquer vértice x, não há nenhuma ligação dirigida começando e acabando em x.

Na configuração das outras opções, manteve-se as que vêm por defeito.

5.5.2 As redes finais

A primeira rede (rede A) resulta dos dados discretizados pelo filtro do WEKA *weka.filters.unsupervised.attribute.Discretize*, a segunda (rede B) dos dados dicretizados nos intervalos com significado médico.

O texto que segue as imagens é o resultado obtido em cada uma das redes.

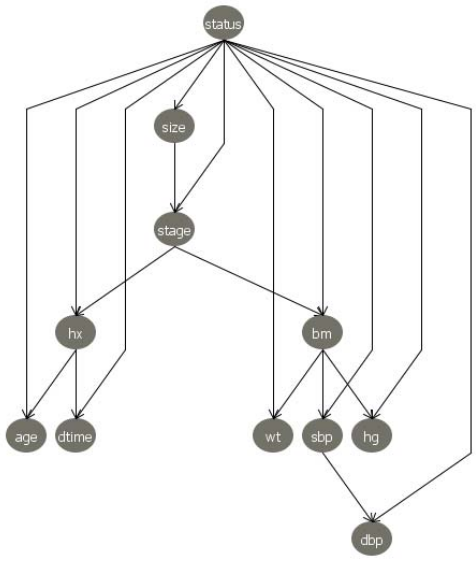

Figura 8: Rede A obtida no WEKA (dados discretizados com o filtro weka.filters.unsupervised.attribute.Discretize)

Tabela 9: Sumário da cross-validation rede A

1	=== Stratified cross-validation ===
2	=== Summary ===
3	
4	Correctly Classified Instances 441 87.8486 %
5	Incorrectly Classified Instances 61 12.1514 %
6	Kappa statistic 0.7161
7	Mean absolute error 0.1536
8	Root mean squared error 0.2856
9	Relative absolute error 36.9069 %
10	Root relative squared error 62.6427 %
11	Coverage of cases (0.95 level) 99.8008 %
12	Mean rel. region size (0.95 level) 69.9203 %
13	Total Number of Instances 502
14	
15	=== Detailed Accuracy By Class ===
16	
17	TP Rate FP Rate Precision Recall F-Measure ROC Area Class
18	0.893 0.155 0.932 0.893 0.912 0.951 dead
19	0.845 0.107 0.767 0.845 0.804 0.951 alive

20	Weighted Avg. 0.878 0.141 0.883 0.878 0.88 0.951
21	
22	=== Confusion Matrix ===
23	
24	a b <-- classified as
25	316 38 | a = dead
26	23 125 | b = alive

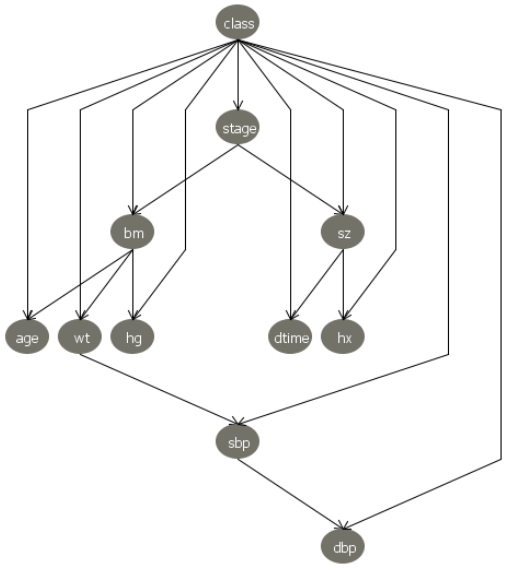

Figura 9: Rede B obtida no WEKA (dados discretizados em intervalos com significado médico)

Tabela 10: Resultado da cross-validation rede B

1	=== Stratified cross-validation ===
2	=== Summary ===
3	
4	Correctly Classified Instances 358 71.3147 %
5	Incorrectly Classified Instances 144 28.6853 %
6	Kappa statistic 0.273
7	Mean absolute error 0.3349
8	Root mean squared error 0.4204
9	Relative absolute error 80.4763 %
10	Root relative squared error 92.2058 %

```
11  Coverage of cases (0.95 level)         99.8008 %
12  Mean rel. region size (0.95 level)     87.4502 %
13  Total Number of Instances              502
14
15  === Detailed Accuracy By Class ===
16
17                TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
18                  0.426    0.167     0.516     0.426     0.467     0.755     alive
19                  0.833    0.574     0.776     0.833     0.804     0.755     dead
20  Weighted Avg.   0.713    0.454     0.7       0.713     0.704     0.755
21
22  === Confusion Matrix ===
23
24    a   b   <-- classified as
25   63  85 |   a =  alive
26   59 295 |   b =  dead
```

5.5.3 Interpretação do resultado

Rede A com resultados na Tabela 9

As linhas 4 e 5 indicam a performance do classificador com os dados de treino (treinamos o algoritmo com os dados e depois verificamo-lo com os próprios dados treinados– *cross-validation*).

A linha 4 significa que dos 502 doentes, 441 foram correctamente classificados, o que dá uma *accuracy* de 87,8%.

A linha 5 significa que dos 502 doentes, 61 foram mal classificados.

As linhas 6 a 12 indicam probabilidades de erro e valores médios.

As linhas 17 a 20 significam a probabilidade de doentes bem classificados em cada classe do status (*dead* ou *alive*).

As linhas 24 a 26 mostram uma matriz de confusão que indica quantas instâncias de cada classe foram classificadas de forma correcta e incorrecta. Se houve 100% de classificação correcta podemos esperar uma matriz de confusão onde todo elemento da diagonal secundária é igual a zero.

O resultado de validação cruzada (os dados de treino são misturados e reamostrados para classificação com a rede criada, a experiência é repetida 10 vezes) é mostrado entre as linhas 4 e 26 da Tabela 9.

Rede B com resultados na Tabela 10

As linhas 4 e 5 indicam a performance do classificador com os dados de treino (treinamos o algoritmo com os dados e depois verificamo-lo com os próprios dados de treino – *cross-validation*).

A linha 4 significa que dos 502 doentes, 358 foram correctamente classificados, o que dá uma *accuracy* de 71.3%.

A linha 5 significa que dos 502 doentes, 144 foram mal classificados.

As linhas 6 a 12 indicam probabilidades de erro e valores médios.

As linhas 17 a 20 significam a probabilidade de doentes bem classificados em cada classe do status (*dead* ou *alive*).

As linhas 24 a 26 mostram uma matriz de confusão que indica quantas instâncias de cada classe foram classificadas de forma correcta e incorrecta. Se houve 100% de classificação correcta podemos esperar uma matriz de confusão onde todo elemento da diagonal secundária é igual a zero.

O resultado de validação cruzada (dados de treino são misturados e reamostrados para classificação com a rede criada, a experiência é repetida 10 vezes) é mostrado entre as linhas 4 e 26 da Tabela 10.

5.5.4 Conclusões

A primeira rede (Figura 8) que resulta dos dados discretizados pelo filtro do Weka apresenta uma melhor exactidão (87,8%), o que significa que classificou correctamente um maior número de casos. Este resultados indica-nos que utilizando o filtro de discretização do Weka obtemos melhores resultados de classificação na rede bayesiana. A grande desvantagem que pode surgir ao trabalharmos com intervalos definidos livremente pelo Weka é que surge uma maior dificuldade da parte do médico em interpretar os resultados. Pode surgir uma situação em que o resultado

obtido não seja conclusivo pois o intervalo pode abranger valores que misturem situações clínicas.

A segunda rede (Figura 9) que resulta dos dados dicretizados nos intervalos com significado médico, apresenta uma exactidão mais baixa (71,3%), mas com uma taxa de acerto razoável.

Ambos classificadores apresentam taxas relativamente altas de falsos positivos e negativos, devendo portanto ser utilizados apenas como indicadores fracos de sobrevida.

Em comparação com a rede manual obtida no GeNIe (nesta última obtivemos uma exactidão de 70%), consideramos as redes obtidas no Weka como de maior confiança.

Em relação à topologia das redes, ambas as redes se aproximam da rede construída manualmente pelo especialista (Figura 6). Como semelhança nas duas verificamos que o *outcome* status tem relação directa de pai-filho com as variáveis hx, size e hg. Ambas distinguem as variáveis sbp, dbp e wt, mas na rede manual estas são relacionadas com hx e nas redes do Weka com bm. Isto pode trazer ao especialista a curiosidade de verificar esta relação de bm com sbp, dbp e wt.

5.6 Modelo III – mini-TUBA

Utilizámos o miniTUBA para estudar relações causais entre um conjunto de variáveis clínicas dos doentes com cancro da próstata (variáveis na Tabela 2), durante um período de 5 anos no final dos quais utilizamos a sobrevivência como *outcome*.

Xiang et al (Xiang, Minter et al. 2007), ao disponibilizarem um *software* que pudesse responder à dúvida de como uma intervenção médica ou um tratamento pode influenciar a resposta de um doente ao longo do tempo, trouxeram uma mais-valia à comunidade científica. No seu artigo, Xiang et al demonstram como as redes bayesianas dinâmicas representam uma alternativa promissora para analisar dados clínicos.

As DBN's apresentam-se de forma semelhante às redes bayesianas simples, num gráfico acíclico.

E de forma semelhante, os gráficos não só permitem ao utilizador perceber quais as variáveis que influenciam outras variáveis, como são o suporte para o cálculo computacional das probabilidades condicionais que são necessárias para os resultados ou aprendizagem. A diferença reside na questão tempo, pois uma variável num determinado tempo t1 influencia as variáveis ligadas a si graficamente num tempo t2.

As redes bayesianas dinâmicas permitem também ciclos temporais entre as variáveis, permitindo ao utilizador interpretar as ligações como causas temporais. Desta forma a interpretação das relações entre as variáveis fornece mais informação clínica relevante.

5.7 Configurações de uma rede bayesiana dinâmica no mini-TUBA

5.7.1 Gestão das variáveis

Para podermos obter uma rede final utilizando o *software* mini-TUBA, foi necessário ajustar a base de dados. Numa rede dinâmica, onde podemos obter relações entre as variáveis ao longo do tempo, é necessário termos valores para essas mesmas variáveis em tempos distintos. Portanto foi necessário acrescentar à nossa base de dados mais 2 tempos clínicos, que foram os tempos escolhidos pelo especialista na construção do

modelo qualitativo como os indicados para recolha de dados com significância em doentes com cancro da próstata. Esses momentos são:

1º Momento (time_point 0): entrevista com o doente na primeira consulta (fase pré-diagnóstico)

2º Momento (time_point 1): após se obter os resultados dos exames auxiliares de diagnóstico (fase pós-diagnóstico)

3º Momento (time_point 2): após o doente ser sujeito a prostatectomia radical e após se obter o resultado de alguns exames realizados no pós-operatório (fase pós-operatório)

Ao criar estes 3 momentos, na base de dados surgiram *missing values* e o seu preenchimento foi aleatório (em variáveis como Hemoglobina ou Tensão Arterial) ou recorrendo a macros no *software Microsoft Office Excel* para o preenchimento dos valores lógicos (em variáveis como a Idade ou Status).

Podemos ver o exemplo de uma macro usada para preenchimento do valor idade:

```
Sub Altera_Idade()
For i = 2 To 1510

    Cells(i + 1, 3) =   Cells(i , 3)+2
    Cells(i + 2, 3) =  Cells(i , 3) +3
  i = i + 3

Next i
End Sub
```

Obtivemos uma base de dados com o aspecto que podemos ver na Tabela 11 onde a 1ª coluna identifica o número do doente, a 2ª coluna o momento em que foram recolhidos os dados dessa linha e da 3ª à 13ª colunas temos as variáveis. A 13ª coluna é o *outcome*. Cada conjunto de 3 linhas representa um doente e toda a informação recolhida sobre ele nos 3 momentos temporais.

Tabela 11: Aspecto da base de dados utilizada no mini-TUBA com representação de 9 dos 502 doentes

Experimental_Unit_ID	Time_Point	age	wt	hx	sbp	dbp	hg	stage	sz	bm	dtime	status
1	0	75	76	0	15	9	13.8	3	0	0	0	1
1	1	77	76	0	15	9	13.8	3	0	0	0	1
1	2	80	68	0	15	9	13.8	4	2	0	60	1
2	0	54	116	0	13	7	14.6	3	0	0	0	1
2	1	56	116	0	13	7	12.6	3	0	0	0	1
2	2	59	108	0	13	7	12.6	4	42	0	1	0
3	0	69	102	1	14	8	13.4	3	0	0	0	1
3	1	71	102	1	14	8	11.4	3	0	0	0	1
3	2	74	94	1	14	8	11.4	4	3	0	40	0
4	0	75	94	1	14	7	17.6	3	0	0	0	1
4	1	75	94	1	14	7	15.6	3	0	0	0	1
4	2	80	86	1	14	7	15.6	4	4	0	20	0
5	0	67	99	0	17	10	13.4	3	0	0	0	1
5	1	67	99	0	17	10	13.4	3	0	0	0	1
5	2	72	91	0	17	10	13.4	4	34	0	60	1
6	0	71	98	0	19	10	15.1	3	0	0	0	1
6	1	71	98	0	19	10	13.1	3	0	0	0	1
6	2	76	90	0	19	10	13.1	4	10	0	24	0
7	0	75	100	0	14	10	13	3	0	0	0	1
7	1	75	100	0	14	10	11	3	0	0	0	1
7	2	80	92	0	14	10	11	4	13	0	46	0
8	0	73	114	1	17	11	12.6	3	0	0	0	1
8	1	73	114	1	17	11	12.6	3	0	0	0	1
8	2	78	106	1	17	11	12.6	4	3	0	60	1
9	0	60	110	0	12	8	14.6	3	0	0	0	1
9	1	60	110	0	12	8	14.6	3	0	0	0	1
9	2	65	102	0	12	8	14.6	4	4	0	60	1

5.7.2 Índice de *Markov*

O índice de *Markov* é uma das condições que podemos alterar antes do cálculo da rede dinâmica. O índice de *Markov* permite seleccionar o tempo entre as várias medições das variáveis, permitindo assim explorar diferentes relações causais ao longo de diferentes escalas de tempo. O índice de *markov* define-se por o intervalo de tempo entre o começo de um evento e o seu efeito. Por exemplo, se seleccionarmos *Markov lag*=1, a rede que obtemos dá-nos as relações entre as variáveis após "1 tempo" de intervalo. Este tempo pode ser considerado 1 minuto, 1 semana, 1 ano, etc., consoante a forma como foi realizada a colheita dos dados. Se seleccionarmos *Markov lag*=2 significa "2 tempos" entre as variáveis relacionadas entre si. O que significa que as relações que surgem na rede bayesiana são após ter ocorrido 2 intervalos temporais (2 anos ou 2 meses, etc…) entre o nó pai e o nó filho.

No nosso caso específico, a base de dados que utilizámos tem uma escala temporal no total de 5 anos e cada variável foi colhida de 2,5 em 2,5 anos. O objectivo é relacionar as variáveis após os 5 anos, isto é, poder verificar a influência das variáveis na sobrevivência após 5 anos, logo o *Markov lag* que nos interessa é de 2 pois pretendemos saltar dois tempos.

5.7.3 Discretização

Como para o cálculo das redes simples, foram utilizados dois processos de discretização das variáveis. No primeiro processo foram escolhidos os intervalos baseados em conhecimento científico (discretização em intervalos com significado médico, os mesmos das redes simples – Tabela 3) e o segundo processo para a discretização das variáveis foi utilizar um processo para discretização do mini-TUBA: discretização por intervalos. Escolhemos o mesmo número de intervalos que na primeira discretização, com intervalo de valores para cada divisão igual, mas o número de valores em cada divisão depende dos valores de cada variável (cálculo automático pelo mini-TUBA).

O mini-TUBA também permite o cálculo de uma rede bayesiana dinâmica sem discretização dos dados.

5.7.4 Algoritmo de aprendizagem

Foi seleccionado o *simulated annealing* em detrimento do *greedy learning*, por nenhum motivo específico, simplesmente porque é o que está por defeito.

5.7.5 Redes

O miniTuba ao efectuar os cálculos do seu algoritmo de forma a obtermos uma rede bayesiana dinâmica, baseia-se na base de dados introduzidos e nas configurações escolhidas pelo utilizador. O resultado é apresentado na forma de 10 redes bayesianas que o *software* denomina de *Top10 networks*. Estas 10 redes são as que apresentam melhores resultados calculadas através do algoritmo de simulated annealing. A 1ª rede do Top10 é considerada a que apresenta melhor probablidade de estar correcta, em relação às outras 9.

O gráfico de *Score Distribution of Top Ten Networks* mostra-nos a probabilidade relativa de cada rede e a probabilidade de a rede nº1 como a rede correcta dentro do Top10 (Figuras 11,13, e 15)

As figuras 10, 12, e 14 mostram-nos as redes nº1 obtidas com diferentes configurações: *Markov lag* 2 anos e com diferentes discretizações.

A Figura 10 representa as relações entre as variáveis com intervalo temporal de 5 anos. Por exemplo, o tamanho da próstata (sz) medido na primeira colheita de dados influencia o valor da hemoglobina (hg) do doente após 5 anos.

A Figura 11 apresenta o *Score Distribution of Top Ten Networks* que nos diz que a rede da Figura 10 obteve uma probabilidade de 0.10 de ser a melhor rede entre as outras 9.

A Tabela 12 apresenta-nos as probabilidades condicionais do Status que tem como único pai o nó Hx (História Familiar).

Podemos fazer a mesma leitura para a Figuras 12 e 13 – Tabela 13 e Figuras 14 e 15 – Tabela 14.

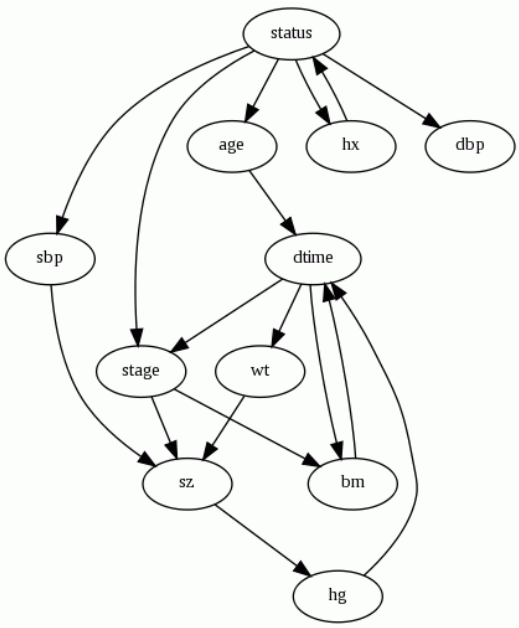

Figura 10: Rede Bayesiana com Markov lag 2 e sem discretização das variáveis (rede 1 do Top10)

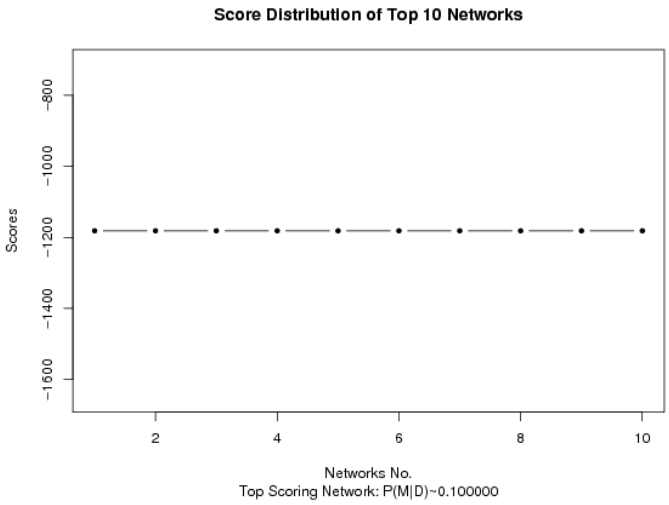

Figura 11: Score atribuido às 10 melhores redes (Top10)

Tabela 12: Probabilidades condicionais do nó Status

Time t		Time t+2	
hx	status	0	1
0	0	0.5000	0.5000
1	0	0.5000	0.5000
0	1	0.6289	0.3711
1	1	0.8047	0.1953

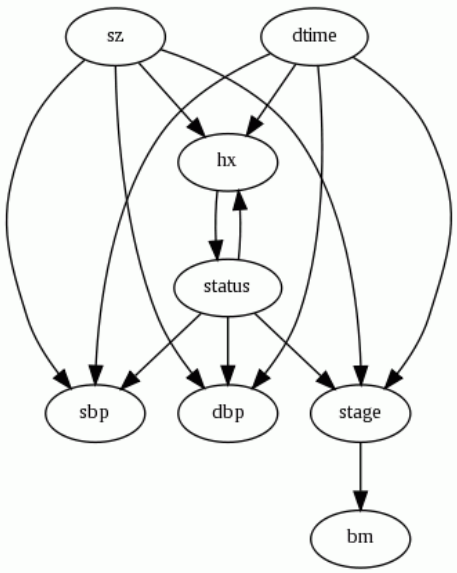

Figura 12: Rede Bayesiana com Markov lag 2 e com discretização das variáveis por binning

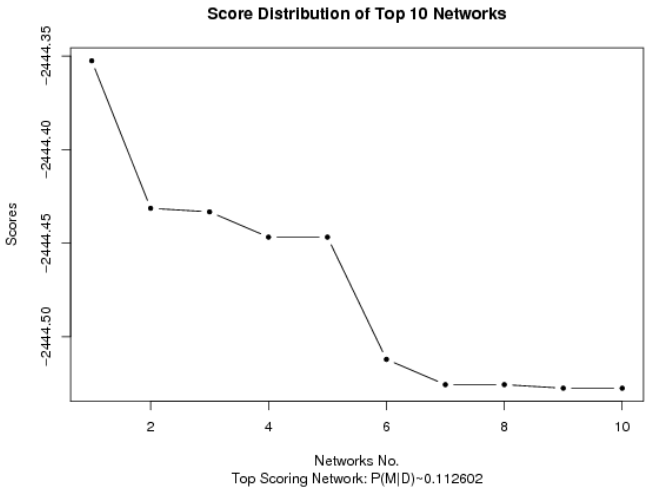

Figura 13: Score atribuido às 10 melhores redes (Top10)

Tabela 13: Probabilidades condicionais do nó Status

Time t		Time t+2	
hx	status	< 0.500	> 0.500
< 0.500	< 0.500	0.5000	0.5000
> 0.500	< 0.500	0.5000	0.5000
< 0.500	> 0.500	0.6289	0.3711
> 0.500	> 0.500	0.8047	0.1953

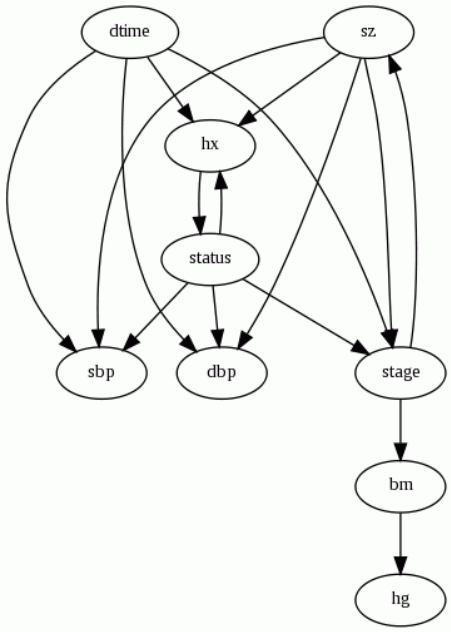

Figura 14: Rede Bayesiana com Markov lag 2 e com dados discretizados em intervalos com significado médico

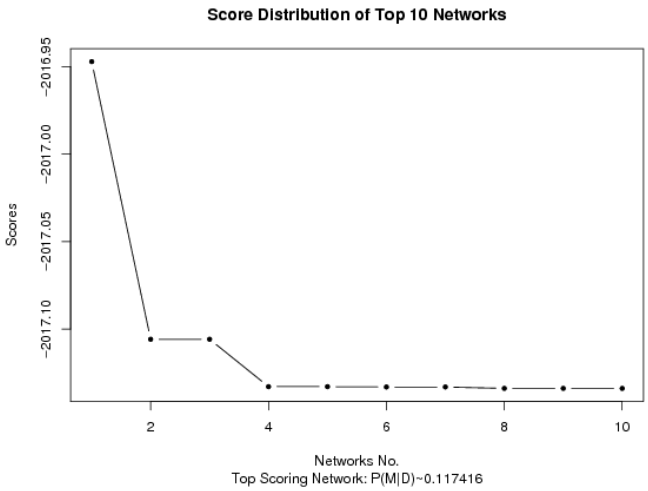

Figura 15: Score atribuido às 10 melhores redes (Top10)

Tabela 14: Probabilidades condicionais do nó Status

Time t		Time t+2	
hx	status	< 0.500	> 0.500
< 0.500	< 0.500	0.5000	0.5000
> 0.500	< 0.500	0.5000	0.5000
< 0.500	> 0.500	0.6289	0.3711
> 0.500	> 0.500	0.8047	0.1953

5.7.6 Resultados e conclusões

Ao fazer o cálculo de uma rede bayesiana dinâmica, o mini-Tuba apresenta como resultados as 10 melhores redes para a base de dados introduzida e a sub-rede que se conserva nas 10 melhores redes (sub-redes em anexo).

Ao compararmos as sub-redes conservadas nas três redes bayesianas com *Markov lag* =2 (com discretização dos dados diferente entre as três) verificamos que a que mantém mais ligações semelhantes nas 10 melhores redes é a rede bayesiana dinâmica em que não foi realizada qualquer discretização dos dados, seguindo-se a rede com discretização por *binning.* Por este motivo podemos concluir que a rede que menos varia no cálculo das melhores 10 redes é a rede sem discretização (Figura 10), que permite maior estabilidade nas relações das variáveis, seguindo-se a rede com discretização por *binning* (Figura 12).

As duas redes com discretização dos dados (Figura 12 e 14) são muito semelhantes relativamente à topologia. Podemos concluir que a discretização dos dados em intervalos de significância médica só poderá trazer vantagens se a discretização pelo *software* criar intervalos que não misturem diferentes situações clínicas (como hipotensão e hipertensão, por exemplo).

Ainda em relação à topologia podemos referenciar a relação do *outcome* status directamente com: hx, sbp, dbp e stage nas três redes dinâmicas. Em comparação com a rede manual evidenciamos a relação com hx, que é a que se mantém. Podemos concluir que a variável hx pode ter uma relação importante em termos médicos com o status.

Se formos comparar a probabilidade de cada rede bayesiana dinâmica ser a melhor (na Figura *Score Distribution of Top Ten Networks*), podemos observar que todas possuem probabilidades muito baixas (entre 10% e 13%), o que significa pouca exactidão nos resultados apresentados. Mas mesmo em resultados baixo podemos verificar que, assim como nas redes bayesianas simples, parece que optando pela discretização automática realizada pelos programas informáticos (permitir liberdade ao *software* para organizar os dados) conseguimos obter resultados com maior exactidão.

5.8 Modelo IV – Redes bayesianas e regras

Programação Lógica Indutiva (*Inductive Logic Programming*, ILP) é uma técnica popular para aprendizagem relacional. Difere de outras técnicas de aprendizagem em pelo menos 2 aspectos: (1) provê ao utilizador uma explicação sobre o conceito que foi aprendido e sobre a classificação de novos exemplos, pois utiliza uma linguagem de primeira ordem para representar o classificador, (2) a linguagem baseada em primeira ordem permite que se definam conceitos que relacionam informação entre indivíduos. Por exemplo, num banco de dados com informações sobre exames de doentes, é possível relacionar informação relacional entre exames de um mesmo doente.

A tarefa básica de ILP (Lavrac and Dzeroski 1994) é, dado um conjunto de exemplos (observações), que podem ser positivos ou negativos, e suas características (atributos, *background knowledge*) e um conjunto de restrições representadas através de uma linguagem lógica de primeira ordem, encontrar um modelo, também descrito em uma linguagem lógica de primeira ordem, de forma que este modelo represente o maior número possível de exemplos positivos e o menor número possível de negativos.

A cada passo da construção do classificador em ILP, precisamos calcular a qualidade de uma determinada regra aprendida. Normalmente, a qualidade de uma regra é medida através de métricas tipo *accuracy, m-estimate* etc. Mas podemos também calcular a qualidade de uma regra através de sua contribuição quando adicionada a uma rede bayesiana. Em outras palavras, a cada geração de uma nova regra, adicionamos esta regra a uma rede bayesiana e calculamos se a adição desta regra maximiza a verossimilhança (*likelihood*) do classificador bayesiano. Se melhorar, esta regra passa a fazer parte da rede. Se a regra não melhorar a qualidade do classificador, esta é descartada e uma nova regra é gerada. O método utilizado neste trabalho, que constrói o classificador a medida em que se aprende novas regras, gera uma rede bayesiana da classe TAN (*Transition Augmented Network*), onde cada nó pode ser descendente de no máximo 2 outros nós (Davis, Burnside et al. 2005). Como mencionado anteriormente, redes TAN são muito populares porque permitem representar ligações mais complexas entre variáveis e não têm uma complexidade muito alta para sua construção.

Em nosso experimento, utilizamos o sistema ILP *Aleph* para gerar as regras e um programa externo, escrito em Java por Davis, para construir a rede bayesiana TAN. Em nossa metodologia utilizamos validação cruzada com 5 *folds* (5-*fold cross-validation)* com uma validação cruzada interna de 4 *folds* para aprender os parâmetros da rede. Utilizamos os 502 exemplos do banco de dados UCI sobre câncer de próstata e obtivemos os resultados por *fold* conforme pode ser visto na Tabela 16.

Table 15: *Score* da melhor rede TAN encontrada por *fold*

Fold	*Score*
1	68,5
2	71,8
3	67,6
4	67,2
5	68,7
Média	68,76

As probabilidades encontradas por este método não estão muito distantes das encontradas pelos outros métodos puramente probabilísticos estudados nas outras secções. Porém temos a vantagem adicional de saber quais são os atributos mais importantes, em forma de regras, que melhoram a probabilidade de classificação. Como exemplo, mostramos, na Figura 14, as regras que foram consideradas relevantes para a construção da melhor rede TAN em um dos *folds*.

```
malignant(A) :-   bm(A,no).
malignant(A) :-   dtime(A,' >=15').
malignant(A) :-   hx(A,no).
malignant(A) :-   weight(A,' >88'),   dbp(A,' <9').
```

Figure 16: Regras utilizadas na rede TAN final em um dos *folds*

Os atributos bm, dtime, hx e uma combinação de peso com dbp mostraram-se relevantes neste "fold" para a construção da melhor rede.

Na Figura 15, apresentamos uma das regras utilizadas durante o treino para construir a rede final. Esta regra nos diz que a pressão diastólica menor do que 9 e o *doubling time of PSA* menor do que 3 são razoavelmente indicativos de morte em doentes com cancro da próstata. Estas características estão ausentes, neste *fold*, em todos os doentes que sobreviveram e aparecem em 21 dos doentes que não sobreviveram. A probabilidade desta regra é de 70,9%.

```
malignant(A) :-   dbp(A,' <9'), dtime(A,' <3').
```

Figure 17: Regra encontrada durante a construção da rede

Outras regras semelhantes foram encontradas em cada *fold*. Estas regras podem ser examinadas pelos especialistas, juntamente com as probabilidades de acerto/erro dadas pela rede bayesiana construída para auxiliar o processo de decisão clínica. Neste processo, o médico especialista pode orientar o doente A, por exemplo, fazer uma biopsia ou algum outro exame complementar.

A observação das regras pode fornecer ao especialista relações entre variáveis (e os intervalos de valores) que podem parecer medicamente incongruentes mas que trazem informação nova e a investigar em mais pormenor.

6 Conclusões

Nesta dissertação foram exploradas as aplicações de algoritmos de redes bayesianas a uma base de dados médica, de forma a obter prognósticos em doentes com cancro da próstata. Mais concretamente, foram criadas redes em que os dados sofreram diferentes processamentos de discretização e redes bayesianas estáticas e dinâmicas, para comparação de exactidão e resultados. Foram explorados alguns programas informáticos que permitem trabalhar bases de dados e observada a sua prestação.
A opção pela utilização do GeNIe, do Weka e do mini-TUBA enquanto *software* de *data-mining* revelou-se adequada dado que facilitou em grande parte todas as análises efectuadas no contexto deste trabalho. A utilização do Aleph e do SAYU não foi complexa, porém, como não apresentam interfaces amigáveis e gráficas, tornou o processo mais trabalhoso.

O estudo feito através de diferentes programas informáticos permitiu constatar que os resultados podem ser concordantes mas apresentam variações na tabela de probabilidades condicionais consoante lidamos com uma rede manual, estática ou dinâmica. Apesar de concordantes, a exactidão conseguida nas redes varia, e aquela que apresenta resultados com uma melhor confiança é a rede obtida no Weka com discretização automática, utilizando o algoritmo do *software* (87,9%). As redes dinâmicas apresentam resultados mais interessantes em termos médicos mas uma baixa probabilidade de ser a melhor rede.

Este estudo permitiu tirar conclusões médicas acerca do prognóstico de doentes com cancro da Próstata. E como possibilidade de interpretação médica nas diferentes redes, deixamos um exemplo:

Se fizermos uma leitura do Status de um doente sujeito a cirurgia da próstata após 5 anos e compararmos com a existência ou não de história familiar de cancro da Próstata, podemos concluir:

Rede manual (foi utilizado o GeNIe para ler as tabelas de probabilidades condicionais):

> Se o doente tem história familiar de cancro da Próstata, a probabilidade de estar morto ao fim de 5 anos do diagnóstico é de 75%. Se o doente não tem

história familiar de cancro da Próstata a probabilidade de estar morto ao fim de 5 anos do diagnóstico é de 66%.

Se compararmos com a **Rede Bayesiana estática** (foi utilizado o WEKA para ler as tabelas de probabilidades condicionais):

Se o doente tem história familiar de cancro da Próstata, a probabilidade de estar morto ao fim de 5 anos do diagnóstico é de 56% para um estadio do tumor 2 ou 3 e de 40% para um estadio do tumor 4. Se o doente não tem história familiar de cancro da Próstata a probabilidade de estar morto ao fim de 5 anos do diagnóstico é de 44% para um estadio do tumor 2 ou 3 e de 60% para um estadio do tumor 4.

Rede Bayesiana com regras:

A melhor rede obtida com SAYU ressalta como atributos mais relevantes o bm, o dtime, o hx e uma combinação de peso com pressão diastólica. Esta rede tem uma probabilidade média de aproximadamente 70% de prever casos de sobrevivência ou não, através dos atributos mencionados.

Redes Bayesianas Dinâmicas:

Rede 1 (Figura 10): Se o doente tem história familiar de cancro da Próstata, a probabilidade de estar morto ao fim de 5 anos do diagnóstico é de 80%. Se o doente não tem história familiar de cancro da Próstata a probabilidade de estar morto ao fim de 5 anos do diagnóstico é de 63%.

Rede 2 (Figura 12): Exactamente os mesmos resultados do que na rede 1

Rede 3 (Figura 14): Exactamente os mesmos resultados do que na rede 1

Como podemos verificar, a história familiar de cancro da próstata pode influenciar o prognóstico do doente, mas não é condição para que esse prognóstico seja a morte.

Esta leitura é possível de fazer para várias variáveis e permite ao médico tirar importantes conclusões.

Importa ainda referir que devido à grande dificuldade em ter disponível uma base de dados com a informação necessária para o cálculo das redes, esta mesma base de dados teve que ser alterada para o cálculo das redes dinâmica, logo os resultados não

são adequados para uma interpretação médica rigorosa. Este facto revelou-se como o maior entrave ao desenvolvimento de um sistema de apoio à decisão clínica assente, principalmente, em redes bayesianas dinâmicas.

Verificou-se que a melhor técnica para discretização dos dados é deixar liberdade aos programas (*software*) para o fazerem, mas trazendo a desvantagem de que pode dificultar a interpretação médica dos resultados quando os intervalos dos valores sobrepuserem diferentes situações clínicas. Este trabalho permite abrir uma nova janela de investigação, pois coloca em causa os intervalos de valores utilizados como valores padrão na área médica, considerando que pode ser importante uma revisão dos mesmos.

De um modo sucinto, pode afirmar-se que este trabalho apresenta uma critica construtiva das potencialidades das redes bayesianas em geral e que estas podem ser um instrumento de mais-valia para o suporte à decisão clínica no prognostico de sobrevivência em doentes com cancro da próstata, assim como para comparação com outras tabelas de prognóstico já existentes (desenvolvidas através de outras técnicas de inferência que não as redes bayesianas).

7 Trabalho futuro

O crescimento desmesurado de casos de cancro e com o aumento exponencial de mortes tem preocupado as populações e as grandes autoridades e coordenadoras da saúde para as nações unidas (*World Health Organization*).

O aproveitamento das grandes bases de dados e dos sistemas de informação em saúde para uma melhoria dos cuidados é um trunfo que emerge na actualidade.

O desenvolvimento de uma aplicação que permita apoio às decisões clínicas por forma a, num futuro, podermos reduzir o numero de mortes por cancro ou aumentar a esperança de vida dos doentes (que deriva da escolha do tratamento a que o doente é sujeito), é uma mais valia dos sistemas de informação em saúde que deve ser explorada.

As bases de dados informatizadas, com informação de doentes com cancro sujeitos a tratamentos são fundamentais para a evolução do conhecimento científico, e o tratamento destes dados e a sua aplicação no apoio à decisão clínica é necessário.

As DBN's são um método de prognóstico estudado e desenvolvido (Ghahramani 1998; Verduijn, Peek et al. 2007; van Gerven, Taal et al. 2008), mas não aplicado no contexto de prognóstico do cancro. A aplicação numa base de dados deste contexto pode trazer grandes vantagens para a decisão médica e para os doentes.

O objectivo, no futuro, é desenvolver uma aplicação que utilize uma DBN desenhada para apoiar as decisões médicas no tratamento do cancro, mas que permita aprendizagem da rede através da introdução de novos dados pelo profissional de saúde. Esta aprendizagem realizada individualmente em cada meio hospitalar permite que cada aplicação esteja orientada para a sua realidade, aproximando mais os resultados da sua realidade específica.

Há ainda muito trabalho a ser realizado, sendo os mais importantes: (1) a integração de ferramentas deste tipo na rotina médica, (2) melhoramento dos algoritmos de aprendizagem das redes, (3) desenvolvimento de algoritmos para o refinamento de redes previamente criadas. Uma das limitações deste trabalho foi a não existência de uma ferramenta que, dada uma rede já criada, construísse uma nova rede com base na

já existente. A construção de tal ferramenta não é um processo trivial e estava fora do escopo desta dissertação.

Anexo I – Sub-redes que se conservam no Top10 das redes bayesianas dinâmicas calculadas no mini-Tuba

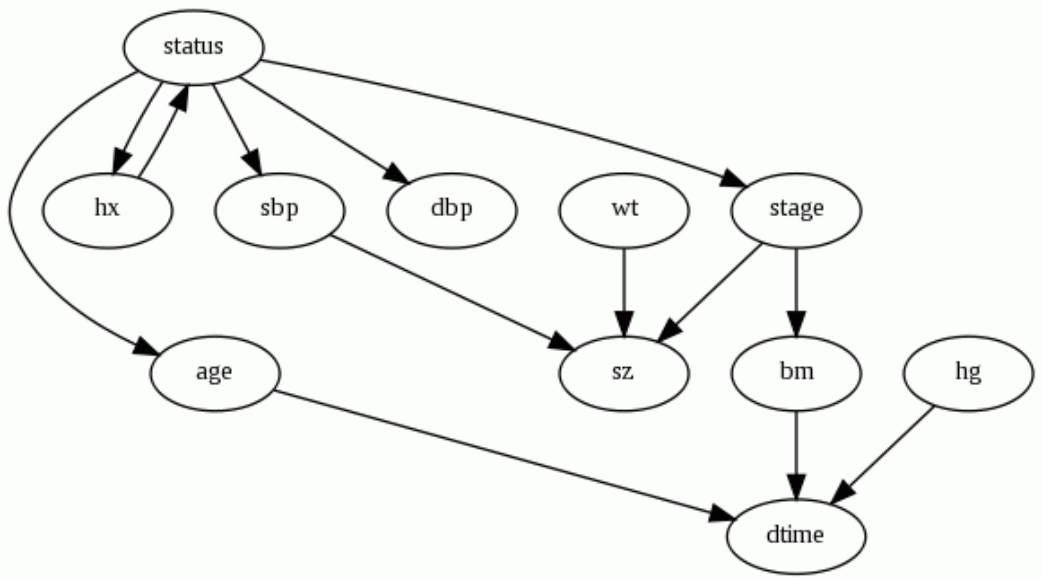

Figure 18: Sub-rede que se conserva nas 10 melhores redes com as seguintes configurações: Markov lag 2 e sem discretização das variáveis (Figura10)

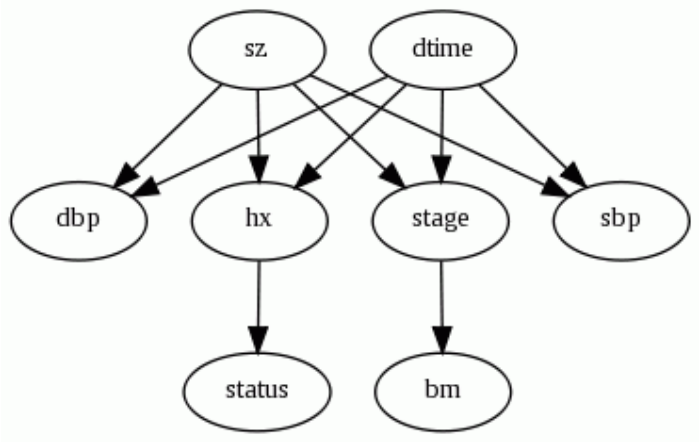

Figure 19: Sub-rede que se conserva nas 10 melhores redes com as seguintes configurações: Markov lag 2 e com discretização das variáveis por binning (Figura10)

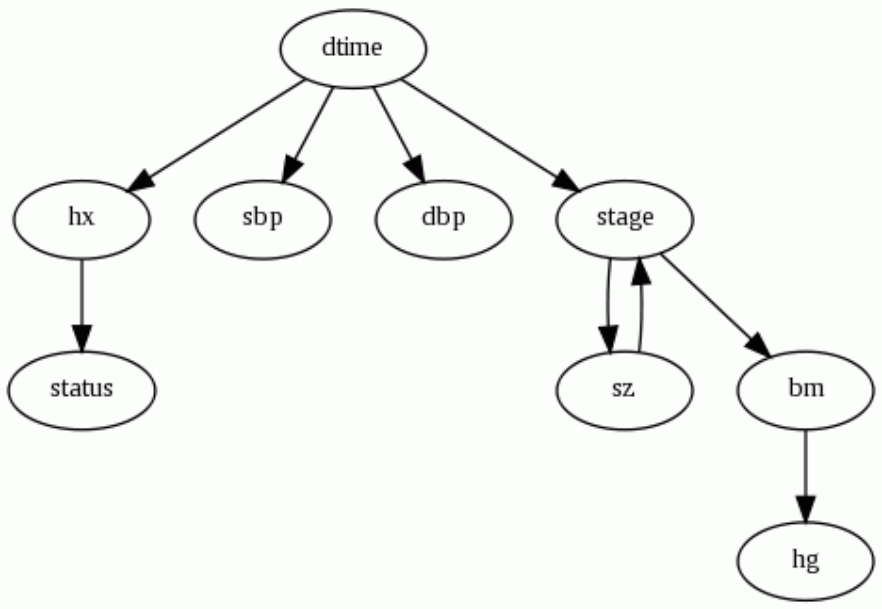

Figure 20: Sub-rede que se conserva nas 10 melhores redes com as seguintes configurações: Markov lag 2 e com dados discretizados em intervalos com significado médico

Referências

Andrews, D. and A. Herzberg (1985). Data : a collection of problems from many fields for the student and research worker, New York, Springer-Verlag.

Barry, M. J., F. J. Fowler, et al. (1992). "The American-Urological-Association Symptom Index for Benign Prostatic Hyperplasia." Journal of Urology **148**(5): 1549-1557.

Chickering, D., D. Heckerman, et al. (1997). A Bayesian Approach to Learning Bayesian Networks with Local Structure. In Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence, Providence, RI.

Chow, C. K. and C. N. Liu (1968). "Approximating Discrete Probability Distributions with Dependence Trees." Ieee Transactions on Information Theory **It14**(3): 462-+.

Colozza, M., F. Cardoso, et al. (2005). "Bringing molecular prognosis and prediction to the clinic." Clin Breast Cancer **6**: 61-76.

Cruz, J. A. and D. S. Wishart (2007). "Applications of machine learning in cancer prediction and prognosis." Cancer Inform **2**: 59-77.

Davis, J., E. Burnside, et al. (2005). Learning Bayesian networks of rules with SAYU. Proceedings of the 4th international workshop on Multi-relational mining. Chicago, Illinois, ACM**:** 13-13.

Dawes, R. M., D. Faust, et al. (1989). "Clinical versus actuarial judgment." Science **243**(4899): 1668-1674.

Gawande, A. (2002). Complications: A surgeon's notes on an imperfect science, New York: Holt.

GeNIe (Visited 2010). "http://genie.sis.pitt.edu/."

Ghahramani, Z. (1998). Learning Dynamic Bayesian Networks. Adaptive Processing of Sequences and Data Structures, International Summer School on Neural Networks, "E.R. Caianiello"-Tutorial Lectures, Springer-Verlag.

Gleason, D. F. and G. T. Mellinger (1974). "Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging." J Urol **111**(1): 58-64.

Graham, A. R., S. H. Paplanus, et al. (1990). "A Diagnostic Expert System for Colonic Lesions." American Journal of Clinical Pathology **94**(4): S15-S18.

Howell, D. (1999). Fundamental Statistics for the Behavioral Sciences, Duxbury Press.

Kononenko, I. (1993). "Inductive and Bayesian Learning in Medical Diagnosis." Applied Artificial Intelligence **7**(4): 317-337.

Lahdesmaki, H. and I. Shmulevich (2008). "Learning the structure of dynamic Bayesian networks from time series and steady state measurements." Machine Learning **71**(2-3): 185-217.

Lavrac, N. and S. Dzeroski (1994). Inductive Logic Programming: Techniques and Applications. . New York, Ellis Horwood.

Mark, H., F. Eibe, et al. (2009). "The WEKA data mining software: an update." SIGKDD Explor. Newsl. **11**(1): 10-18.

Meehl, P. E. (1996). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence, Northvale, NJ: Jason Aronson. (Original work published 1954).

Partin, A. W., L. A. Mangold, et al. (2001). "Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium." Urology **58**(6): 843-848.

Pontes, J. E., H. Ohe, et al. (1984). "Transrectal ultrasonography of the prostate." Cancer **53**(6): 1369-1372.

Revett, K., P. S. Magalhães, et al. (2006). Data mining a prostate cancer dataset using rough sets, IEEE CS Press.

Sim, I., P. Gorman, et al. (2001). "Clinical decision support systems for the practice of evidence-based medicine." Journal of the American Medical Informatics Association **8**(6): 527-534.

Srinivasan, A. (2001). The Aleph Manual. http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph_toc.html.

Teeter, A. E., J. C. Presti, et al. (2009). "Does early prostate-specific antigen doubling time (ePSADT) after radical prostatectomy, calculated using PSA values from the first detectable until the first recurrence value, correlate with standard PSADT? A report from the Shared Equal Access Regional Cancer Hospital Database Group." Bju International **104**(11): 1604-1609.

Thompson, I. M., D. K. Pauler, et al. (2004). "Prevalence of prostate cancer among men with a prostate-specific antigen level <= 4.0 ng per milliliter." New England Journal of Medicine **350**(22): 2239-2246.

van Gerven, M. A. J., B. G. Taal, et al. (2008). "Dynamic Bayesian networks as prognostic models for clinical patient management." Journal of Biomedical Informatics **41**(4): 515-529.

Verduijn, M., N. Peek, et al. (2007). "Prognostic Bayesian networks I: Rationale, learning procedure, and clinical use." Journal of Biomedical Informatics **40**(6): 609-618.

Verduijn, M., P. M. J. Rosseel, et al. (2007). "Prognostic Bayesian networks II: An application in the domain of cardiac surgery." Journal of Biomedical Informatics **40**(6): 619-630.

WHO (visited 2010). World Health Organization - http://www.who.int/cancer/en/.

Wolberg, W. H., W. N. Street, et al. (1994). "Machine Learning Techniques to Diagnose Breast-Cancer from Image-Processed Nuclear Features of Fine-Needle Aspirates." Cancer Letters **77**(2-3): 163-171.

Xiang, Z., R. M. Minter, et al. (2007). "miniTUBA: medical inference by network integration of temporal data using Bayesian analysis." Bioinformatics **23**(18): 2423-2432.