Relatório de Experimentos

Pablo Dalbem de Castro RA 038499
George Barreto Pereira Bezerra RA 003030

Tema: Aprendizado em Redes Bayesianas

1. Introdução

Este relatório apresenta os resultados de testes experimentais realizados com o algoritmo K2 na tarefa de aprendizado de redes bayesianas. Todos os testes se basearam na capacidade do algoritmo em reproduzir uma rede bayesiana pré-definida através de amostragens produzidas por esta rede. Em outras palavras, uma rede bayesiana é utilizada como modelo para produzir uma determinada amostragem e através desta amostragem, o algoritmo K2 deve ser capaz de reproduzir a mesma rede. Será demonstrado aqui, porém, que segundo o critério utilizado para avaliar a qualidade de uma rede, o desempenho do algoritmo é totalmente dependente da representatividade da amostragem, isto é, não há compromisso direto em reproduzir a rede original, mas sim uma rede que represente a distribuição da amostragem em questão, a qual não corresponde necessariamente à verdade.

As Seções 2 e 3 deste relatório trazem introduções sobre o algoritmo K2 e o critério de verossimilhança, utilizado para avaliação da rede, respectivamente. A Seção 4 avalia a capacidade do algoritmo em reproduzir a rede original em função da quantidade de amostras disponíveis e a Seção 5 avalia o potencial do algoritmo K2 para encontrar a distribuição observada nas amostras, isto é, a sua capacidade de maximizar a verossimilhança. Por fim, a Seção 6 apresenta um balanço das conclusões obtidas ao longo do relatório e faz um questionamento sobre a utilidade prática das redes bayesianas como ferramenta de inferência de relações causais entre variáveis.

2. Algoritmo K2

O algoritmo K2 [Cooper & Herskovits, 1992] de inferência de redes bayesianas funciona de forma bastante simples. Ele faz uma busca "gulosa" no espaço de possíveis estruturas de rede, à procura daquela que maximiza um determinado critério de qualidade (no caso, a verossimilhança, que será discutida na próxima seção).

Inicia-se com uma rede sem conexões, isto é, considera-se as variáveis totalmente independentes umas das outras, e avalia-se a qualidade da rede em relação a uma dada amostragem. O próximo passo consiste em adicionar um arco à estrutura. Testam-se todas as possíveis estruturas que contém apenas um arco, avaliando cada uma, e armazenando aquela que maximiza o critério de qualidade. Se a rede com uma conexão apresentar maior qualidade que a rede sem conexões, a nova rede substitui a anterior. A partir daí, o processo se repete considerando agora redes com duas conexões. Se a rede com duas conexões for melhor que a rede de uma conexão, a primeira substitui a segunda. E assim

sucessivamente, até que uma rede com uma conexão a mais não seja capaz de aumentar o valor do critério de qualidade. Fica-se com a rede anterior, de maior qualidade.

3. Verossimilhança como Critério de Qualidade

A verossimilhança é uma medida estatística que estima a probabilidade de um determinado modelo reproduzir um conjunto de amostras observado. Ou seja, ela mede o quanto a densidade de probabilidade representada pelo modelo se aproxima da distribuição apresentada nos dados. Esta é uma medida bastante utilizada como critério de seleção de modelos (uma rede bayesiana é um modelo) quando não se possui nenhum conhecimento a priori, isto é, a única informação disponível a respeito do problema são as amostras.

Porém, a verossimilhança possui algumas desvantagens. Primeiramente, não há compromisso com a manutenção de simplicidade; muito pelo contrário, ela vai exatamente contra o princípio da "navalha de Occam" [Jacquett, 1944]. Entre dois modelos que expliquem os dados de maneiras aproximadamente semelhantes (isto é, com verossimilhanças aproximadamente iguais), o critério de máxima verossimilhança tenderá sempre a escolher aquele modelo de maior complexidade.

Como conseqüência disso vem o segundo problema: o modelo se torna excessivamente susceptível à qualidade da amostragem. Segundo a máxima verossimilhança, o modelo deve possuir quantas variáveis forem necessárias para melhor se adequar aos dados observados. Isto, porém, o torna muito específico para aqueles dados. Caso as amostras se distanciem ligeiramente da distribuição real – e isto geralmente vai ocorrer – a capacidade de previsão do modelo se torna bastante comprometida. O modelo se torna pouco tolerante ao ruído inerente à característica probabilística da amostragem. Se ganha em especificidade, mas perde-se muito em generalidade.

Em outras palavras, a máxima verossimilhança evidencia um dilema ingrato envolvendo seleção de modelos: ao aumentar a especificidade do modelo, reduzindo assim o *bias*, o critério termina por aumentar, como conseqüência inevitável, a sua susceptibilidade ao ruído (variância). (Veja *bias* × *variance dilemma* em [Forster, 2000].)

Uma solução muito mais adequada seria escolher um modelo cuja complexidade representa o ponto ótimo entre *bias* e variância, isto é, um ponto onde não é possível reduzir um sem aumentar o outro. Esta discussão, no entanto, não é o foco deste relatório. O leitor interessado deve se referir à literatura sobre seleção de modelos, onde esta questão é bastante debatida [Forster, 2000].

É possível encontrar na literatura critérios que procuram amenizar o problema da máxima verossimilhança. Os critérios BIC (*Bayesian Information Criterion*) [Schwartz, 1978] e AIC (*Akaike Information Criterion*) [Akaike, 1974], por exemplo, são medidas de qualidade bastante adotadas que introduzem um coeficiente de penalização da complexidade em conjunto com a verossimilhança no cálculo da qualidade do modelo. O resultado geralmente é mais interessante na prática do que o obtido com a máxima verossimilhança.

4. Descobrindo a Estrutura Original

Este experimento consiste em avaliar a capacidade do algoritmo K2 em descobrir a estrutura original de uma rede bayesiana em função do tamanho da amostragem. É importante observar que o tamanho do conjunto amostral em si, não é a variável mais relevante aqui. O objetivo principal é avaliar o potencial do algoritmo em função do nível de representatividade dos dados. Entretanto, dada a característica probabilística das amostras, a maneira mais direta de se obter amostras de maior representatividade é, logicamente, aumentando o número de amostras. Quanto maior o tamanho da amostragem, maior tende a ser a sua representatividade, de maneira assintótica. Assim, tendo infinitas amostras, a densidade de probabilidade dos dados é exatamente a densidade de probabilidade do modelo original, isto é, a verdade.

4.1. Experimento 1: Heckerman *et al.* (1997)

Este experimento foi retirado de [Heckerman, 1997]. Ele evidencia de forma bastante ilustrativa a dependência da rede obtida em relação ao tamanho do conjunto de dados. Partindo-se da rede da figura 1, onde são mostradas também as tabelas de probabilidade de cada variável, foram geradas amostras a serem apresentadas ao algoritmo K2.

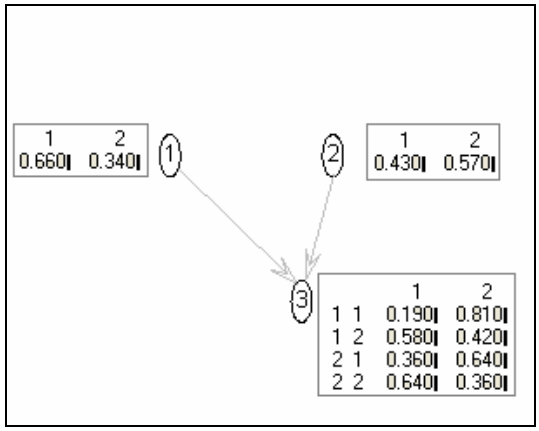


Figura 1. Rede bayesiada utilizada como modelo no experimento 1. Exemplo tirado de [Heckerman, 1997].

Como salientado anteriormente, o objetivo e observar se o algoritmo consegue convergir para a rede original partindo apenas dos dados. Cinco casos foram testados: 150, 250, 500, 1000 e 2000 amostras. Para cada situação, 20 conjuntos diferentes com o mesmo número de amostras foram gerados. Os resultados são mostrados na tabela 1. Apenas a relação entre as variáveis 1 e 3 foi avaliada no experimento, pois a relação entre as variáveis 2 e 3 é identificada corretamente pelo algoritmo com facilidade.

Tablela 1. Resultados do experimento 1. A tabela mostra as probabilidades de a variável v_1 causar v_3 e de v_1 e v_3 estarem relacionadas após 20 execuções do algoritmo K2 para cada situação.

Nº de amostras	**p(v_1 causa v_3)**	**p(v_1 causa v_3 ou v_3 causa v_1)**
150	0.05	0.2
250	0.15	0.4
500	0.45	0.85
1.000	0.85	1
2.000	0.85	1

Note na tabela que o desempenho do algoritmo é totalmente dependente do número de amostras. Como discutido na Seção 2, esta relação já é esperada, pois à medida que o número de amostras aumenta, mais próxima a verossimilhança se torna da verdade. Quando 500 dados são utilizados, é possível perceber que a relação de dependência entre as variáveis já se torna bem evidente, com 85% de probabilidade, porém não há distinção clara de que v_1 causa v_3. Apenas a partir de 1.000 amostras o algoritmo é capaz de identificar corretamente a relação de causalidade.

Este exemplo traz à tona uma questão importante. Para um problema tão simples como este, são necessárias pelo menos 1.000 amostras para descobrir a estrutura original da rede. Isso inaceitável sob praticamente quaisquer circunstâncias em problemas reais. Quase sempre um número tão elevado de amostras em relação ao de variáveis não está disponível. O problema tende a se tornar ainda mais crítico quando o número de variáveis é aumentado. Segundo o princípio de *curse of dimensionality* [Bellman, 1961], o número de amostras necessárias para resolver um problema deste tipo aumenta exponencialmente com o número de variáveis. Ora, esta conclusão simplesmente elimina qualquer esperança de recuperar a estrutura verdadeira das relações causais em problemas complexos de mundo real, a exemplo das redes gênicas [Geard, 2004], onde o número de variáveis tende a ser grande e a quantidade de amostras é limitada.

No entanto, em situações onde nenhum conhecimento a priori é disponível, qualquer informação, mesmo que imprecisa, é considerada de grande relevância. Veja que com 500 amostras é possível descobrir que existe uma forte relação de causalidade entre as variáveis, mesmo que o sentido da relação não esteja definido. Infelizmente, essa condição não ajuda muito. 500 amostras é ainda um número muito alto dada a quantidade de variáveis. Passa-se de uma situação "extremamente inviável" para uma "muito inviável", o que não é de grande valia.

A despeito da dramaticidade da questão exposta acima, cabe lembrar que a relação entre as variáveis 2 e 3 é facilmente percebível pelo algoritmo, como descrito anteriormente. Mais uma vez, quando nenhum conhecimento a priori é sabido, ter certeza da relação de causalidade entre duas variáveis pode ser considerado de extrema importância, o que faz do algoritmo uma ferramenta útil.

4.2. Experimento 2: exemplo clássico da chuva.

Este é um exemplo clássico da literatura. A rede bayesiana consiste de 4 variáveis binárias, onde 1 significa *não* e 2 significa *sim*. A estrutura da rede e o significado lingüístico das variáveis são mostrados na figura 2.

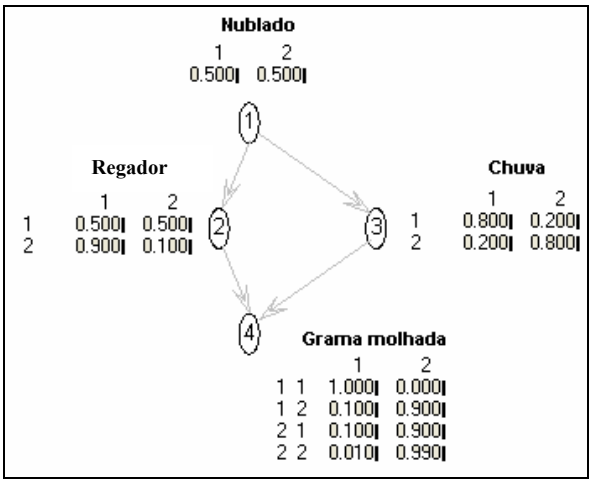
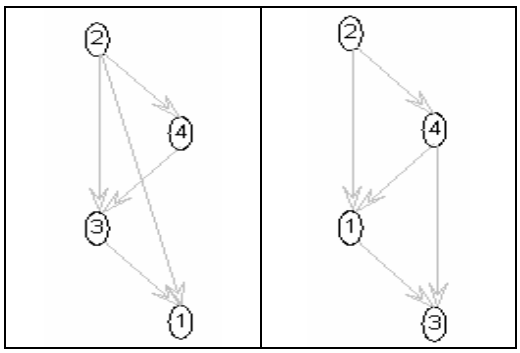


Figura 2. Exemplo clássico da chuva com 4 variáveis binárias. 1 significa *não* e 2 significa *sim*.

O algoritmo K2 foi utilizado para resolver o problema para 200, 1.000, 2.000, 10.000 e 50.000 instâncias. Para as 4 primeiras situações, o algoritmo oscilou entre duas estruturas, nenhuma delas exatamente a original, mostradas na figura 3a e 3b. Para 50.000 variáveis, o algoritmo encontrou apenas a estrutura mostrada na figura 3b.



(a) (b)

Figura 3. (a) Estrutura encontrada para 200, 1.000, 2.000 e 10.000 amostras. (b) Estrutura encontrada em todas as situações, inclusive a com 50.000 amostras.

O algoritmo K2 se mostrou incapaz de recuperar a estrutura original do problema, muito embora, tenha sido capaz de relacionar as variáveis com certa eficiência. Veja na figura 3a, que mesmo que o sentido das setas não esteja de acordo com o modelo original, a direção do relacionamento causal está correto, embora uma conexão adicional relacionando 2 e 3 tenha sido inserida. O mesmo acontece com a estrutura da figura 3b, sendo que a conexão adicional relaciona 1 com 4.

Para analisar os resultados obtidos, vamos assumir que a amostragem com 50.000 pontos é suficientemente grande para representar a distribuição verdadeira adequadamente, isto é, vamos considerar que mesmo com infinitas amostras o resultado seria o mesmo da figura 3b. Sendo assim, duas questões merecem observação especial (por conveniência, essas questões serão forçosamente tratadas separadamente aqui):

1. Porque o algoritmo introduziu uma conexão a mais na rede, sendo que as variáveis em questão não estão diretamente relacionadas?

2. Porque não foi possível determinar com exatidão o sentido das relações causais, dado que a representatividade da amostragem é a máxima possível?

Nesta seção analisaremos apenas a primeira questão. A segunda será discutida na análise das Seções 4.3 e 5.

Uma possível explicação para o resultado destacado na questão 1 é a seguinte. Um modelo com mais variáveis pode explicar com igual ou maior precisão um fenômeno qualquer do que um modelo semelhante, mas com uma variável a menos. Se o modelo com menos variáveis explica perfeitamente o fenômeno, então o modelo com mais variáveis pode explicar perfeitamente também, basta considerar o valor da variável adicional como nulo. Diz-se que esses modelos são "modelos aninhados" (*nested models*), segundo a teoria de seleção de modelos.

Seguindo este raciocínio agora no contexto das redes bayesianas, se uma rede com 4 arcos explica bem um conjunto de dados, uma rede com 1 ou mais arcos além desses 4 pode explicar os mesmos dados de forma igual ou melhor. Ou seja, estas redes são modelos aninhados. Como o critério de máxima verossimilhança não penaliza a complexidade, o modelo mais complexo tenderá sempre a ser o escolhido (essa particularidade foi descrita na Seção 3), sendo, portanto, esta a razão para as redes encontradas possuírem uma conexão extra.

Não se pode desconsiderar também que o algoritmo K2 pode estar realizando uma busca ineficiente, isto é, talvez a rede original, ou uma outra rede qualquer, possua uma verossimilhança maior que a da rede encontrada. Dessa forma, a explicação dada acima não se aplica necessariamente.

4.3. Experimento 3: exemplo da gravidez.

Esta rede bayesiana representa uma relação causal que determina a probabilidade de uma mulher estar grávida ou não, dado o estado de uma série de variáveis. Estes dados foram encontrados em http://www.cs.huji.ac.il/labs/compbio/Repository/. A rede possui 6 variáveis, sendo a primeira com 7 valores discretos e as outras binárias. A figura 4 mostra a rede juntamente com as tabelas de probabilidade de cada variável.

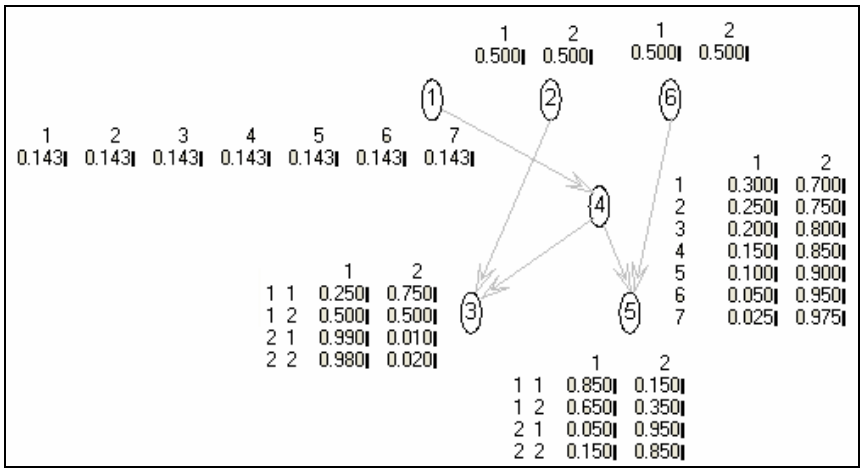
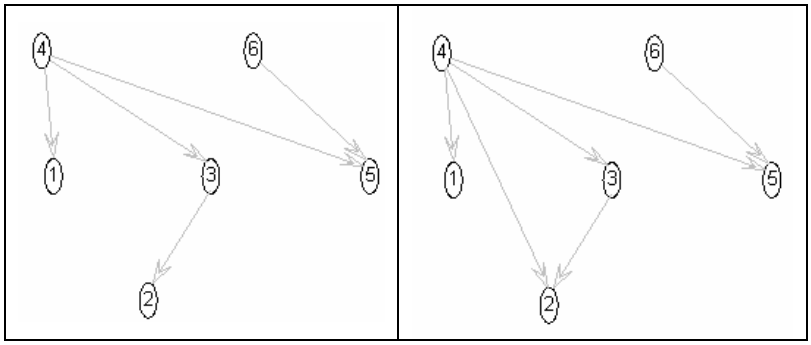


Figura 4. Exemplo da gravidez. Rede bayesiana com 6 variáveis, sendo a primeira com 7 valores discretos e as outras binárias.

Para amostragens com 1000 e 2000 pontos, o algoritmo oscilou entre dois tipos de estruturas, mostradas nas figura 5a e 5b. A rede da figura 5a corresponde exatamente à mesma estrutura relacional do exemplo original, sendo que o sentido dos arcos é diferente. Já a figura 5b mostra uma rede igual à da figura 5a, porém com um arco a mais, correspondendo assim a um modelo aninhado. Para amostragens com 4000 e 8000 dados, apenas a estrutura da figura 5b foi encontrada, quando não uma estrutura ainda mais complexa.



(a) (b)

Figura 5. (a) Rede encontrada com direções das relações causais semelhantes ao modelo original. (b) Rede encontrada com uma conexão adicional.

Para este problema o algoritmo parece ter obtido um desempenho relativamente bom. Ele foi capaz de encontrar a estrutura da rede original em termos de relacionamento de variáveis, mesmo sendo esta rede mais complexa que as anteriores. Entretanto, parece que o problema da complexidade adicional provocada pela medida de qualidade da rede persiste. Este resultado reforça o fato de que uma medida que penalize a complexidade de um modelo pode ser mais adequada que simplesmente a máxima verossimilhança.

Vale ressaltar também que assim como no experimento 2, o sentido das relações causais não pôde ser recuperado adequadamente (este resultado está relacionado à questão 2, levantada na Seção 4.2), embora as redes encontradas possuam arcos exatamente entre as mesmas variáveis. Redes deste tipo, com conexões entre as mesmas variáveis não importando o sentido, são ditas equivalentes de Markov [Heckerman, 1997]. Embora não seja uma relação universal, redes equivalentes de Markov muitas vezes apresentam a mesma densidade de probabilidade (isto é, são equivalentes de distribuição [Heckerman, 1997]). Isto significa que, caso duas redes possuam exatamente a mesma distribuição, não há condições de se distinguir entre as duas na ausência de conhecimento a priori. Ou seja, em muitas situações, será impossível para um algoritmo qualquer de inferência de redes bayesianas recuperar exatamente a mesma rede que gerou os dados, mesmo que a amostragem seja infinita, pois há outros modelos que representam os mesmos dados com a mesma parcimônia e a mesma eficiência, sendo, portanto totalmente equivalentes em termos de complexidade e distribuição.

Mais uma vez, convém considerar que este não é necessariamente o caso aqui. É possível que o algoritmo esteja simplesmente selecionando uma rede ruim. Esta justificativa será abordada na próxima seção.

5. K2 como Algoritmo de Maximização

Como comentado anteriormente, o algoritmo K2 é um algoritmo de busca. Ele tenta encontrar a estrutura de rede que maximiza a verossimilhança em relação a um conjunto de dados. Os experimentos realizados na Seção 4 mostraram que nem sempre a rede encontrada corresponde ao modelo original, muitas vezes porque o número de dados utilizados não é suficientemente representativo. Além disso, o critério de máxima verossimilhança influencia o resultado de forma a encontrar modelos menos parcimoniosos. Mas e quanto à eficiência do algoritmo em si? Será que o K2 encontra sempre a rede com verossimilhança máxima dentre todas as possíveis ou ele converge para um máximo local? Em outras palavras, o fato das redes encontradas não terem sido exatamente as procuradas é resultado apenas da falta de representatividade dos dados ou a eficiência do algoritmo K2 também influencia no resultado?

O objetivo desta seção é avaliar o potencial do algoritmo K2 como algoritmo de maximização. Para isso, serão comparadas as verossimilhanças das redes originais com as redes encontradas. Se a rede encontrada possui uma verossimilhança maior que a da rede original significa que o algoritmo está fazendo o seu papel em maximizar o critério de qualidade. Caso contrário, o algoritmo não está fazendo a busca de maneira adequada, e a sua ineficiência tem uma boa parcela de responsabilidade nos resultados encontrados.

5.1 Quando a Verossimilhança Não Corresponde à Verdade

Quando a distribuição dos dados observados não corresponde exatamente à densidade de probabilidade do modelo original, é possível que exista uma outra estrutura de rede bayesiana capaz de representar os dados com uma maior verossimilhança. Neste caso, o compromisso do algoritmo de busca é de encontrar esta outra rede e não a rede original que gerou os dados. Utilizando os mesmos modelos da Seção 4, foram avaliadas as verossimilhanças das redes originais e das redes encontradas quando a representatividade dos dados não é máxima.

Para o experimento 1 da Seção 4.1, comparamos a verossimilhança da rede encontrada com a da rede original quando o número de amostras é 150 e 250, valores em que as duas redes diferem e a representatividade dos dados é baixa. A tabela 2 mostra os resultados obtidos em 20 diferentes amostragens para cada situação. Os valores da tabela são negativos porque a verossimilhança é medida em logaritmo.

Tabela 2. Desempenho médio do algoritmo K2 para o problema do experimento 1 em 20 amostragens. A tabela mostra a média da verossimilhança da rede original e da rede encontrada e também a porcentagem de vezes em que a rede encontrada pelo algoritmo foi melhor que a rede original.

N^o de amostras	**Média da veross. da rede original**	**Média da veross. da rede encontrada**	**Rede encontrada melhor que a orginal (%)**
150	-304,5946	-302,8409	100%
250	-498,2075	-499,3341	90%

Para o problema do experimento 2, foram utilizadas amostragens com 200 e 1000 dados. Os resultados obtidos em 20 amostragens diferentes são mostrados na tabela 3.

Tabela 3. Desempenho médio do algoritmo K2 para o problema do experimento 2 em 20 amostragens. A tabela mostra a média da verossimilhança da rede original e da rede encontrada e também a porcentagem de vezes em que a rede encontrada pelo algoritmo foi melhor que a rede original.

Nº de amostras	**Média da veross. da rede original**	**Média da veross. da rede encontrada**	**Rede encontrada melhor que a orginal (%)**
200	-395,3814	-397,2511	30%
1000	-1.960,4	-1.957,1	20%

Para o experimento 3, foram testadas situações com 500 e 1000 pontos. A tabela 4 apresenta os resultados.

Tabela 4. Desempenho médio do algoritmo K2 para o problema do experimento 3 em 20 amostragens. A tabela mostra a média da verossimilhança da rede original e da rede encontrada e também a porcentagem de vezes em que a rede encontrada pelo algoritmo foi melhor que a rede original.

Nº de amostras	**Média da veross. da rede original**	**Média da veross. da rede encontrada**	**Rede encontrada melhor que a orginal (%)**
500	-2,2728	-2,2645	100%
1000	-4,5151	-4,5120	100%

Os resultados desta análise são um pouco contraditórios. Para os experimentos 1 e 3, o algoritmo K2 se comportou extremamente bem, encontrando em quase todas as situações uma rede que maximiza a verossimilhança. No experimento 2, no entanto, o desempenho do algoritmo foi bastante ineficiente. A rede original possui quase sempre uma verossimilhança maior que a da rede encontrada. Isso significa que o algoritmo K2 deveria ter sido capaz de recuperar a rede original ou então alguma outra com maior verossimilhança.

Comecemos então analisando o experimento 2. Como dito na Seção 2, algoritmo K2 é um algoritmo guloso. Uma vez seguindo em uma direção, ele não poderá voltar atrás, convergindo assim para um ótimo local. É perfeitamente possível que para um dado problema a introdução de um determinado arco *a* seja melhor em termos de qualidade do que a de qualquer outro arco, mas que dois outros arcos *b* e *c* em conjunto e na ausência de *a* produzam uma estrutura ainda melhor. A questão é que o algoritmo decidirá pelo inicialmente pelo arco *a*, sendo então incapaz de encontrar a melhor estrutura, isto é, aquela que contém *b* e *c*.

Nos outros experimentos isto não aconteceu. O algoritmo encontrou uma solução melhor que a original (embora não saibamos se existe uma outra solução melhor que a encontrada), indicando que a sua busca foi eficiente. Imagina-se, pois, que as superfícies de busca no espaço de estruturas seja menos "acidentado" para estes problemas. Se elas realmente possuírem menos ótimos locais que a superfície de busca do experimento 2, torna-se muito mais fácil para um algoritmo guloso encontrar a melhor solução.

Através dos testes realizados, não é possível generalizar a conclusão de que o algoritmo é uma técnica boa ou ruim de maximização; conclui-se apenas que ele não é ótimo. É necessário avaliar o desempenho de outros algoritmos no problema do experimento 2 para realizar uma análise comparativa.

5.2 Quando a Verossimilhança é a Verdade

Quando o número de amostras é suficientemente grande, pelo menos para os problemas simples analisados na seção 4, é aceitável esperar que não exista outra rede a não ser a original (ou então a sua equivalente de distribuição) que explica melhor os dados observados, isto é, que a verossimilhança é uma medida da verdade. Neste experimento tentaremos avaliar se em situações desse tipo, a rede encontrada pelo K2, quando difere da rede original, é uma equivalente de distribuição. Isto significa dizer que o algoritmo foi competente o suficiente para encontrar a melhor solução (ótimo global), mesmo que a rede não seja exatamente a esperada.

O primeiro teste foi realizado para a rede do experimento 2, na situação em que o número de amostras é 50.000. Espera-se que esse número de amostras seja suficientemente grande para representar fielmente o modelo verdadeiro. O segundo teste foi feito com a rede do experimento 3, também para 50.000 amostras, quando o algoritmo encontra a mesma rede da figura 5b. Os resultados obtidos são mostrados na tabela 5.

Tabela 5. Verossimilhança do modelo original e da rede encontrada pelo algoritmo K2 para os experimentos 2 e 3 com 50.000 amostras.

Experimento	Verossimilhança do modelo original	Verossimilhança da rede encontrada
2	-9.5158e+004	-9.8990e+004
3	-2.2267e+005	-2.2267e+005

No primeiro teste, a verossimilhança da rede obtida (figura 3b) é menor do que o da rede original. Isto significa que o algoritmo não teve um bom desempenho, pois as redes não são equivalentes de distribuição. No segundo teste, entretanto, a rede encontrada (figura 5b) e a rede original, embora diferentes, possuem exatamente a mesma verossimilhança, ou seja, são equivalentes de distribuição. Se a distribuição dos dados for realmente suficientemente representativa, o algoritmo foi capaz de encontrar o ótimo global.

Resta, no entanto, uma dúvida. As redes do segundo teste, embora equivalentes de distribuição, possuem complexidades diferentes. É possível que a rede obtida seja então um modelo aninhado da rede da figura 5b, de menor complexidade, pois elas são praticamente iguais, com exceção de um arco que não pertence à estrutura original. Avaliando a verossimilhança da rede da figura 5b, o valor obtido foi 2.2284e+005, ligeiramente diferente do obtido com as outras redes. Isto descarta a hipótese de que os modelos são aninhados, porém, se um critério como BIC ou AIC, que penalizam a complexidade, tivesse sido utilizado, a estrutura original teria sido privilegiada por ser menos complexa, dando uma maior chance ao algoritmo de recuperá-la.

6. Discussão

Os métodos de inferência de redes bayesianas são realmente úteis como ferramenta de descoberta das relações causais entre variáveis e de modelagem de distribuição em problemas complexos de mundo real? Refiro-me mais especificamente a problemas em que o número de variáveis tende a ser grande e a quantidade de amostras é bastante limitada. As redes têm utilidade prática para este tipo de problema?

Embora as análises feitas neste relatório sejam insuficientes para responder com precisão a esta pergunta, baseado nos resultados obtidos é possível arriscar um palpite coerente.

Foi visto que o algoritmo K2 depende de uma quantidade de amostras excessivamente grande – considerando as restrições impostas pelos problemas em foco – para chegar a uma rede que explique perfeitamente os dados (experimentos 1 e 3) e que em algumas situações, nem com um número infinito de amostras é possível recuperar a densidade de probabilidade original (experimento 2) – este último caso deve ser considerado à parte, já que o resultado está relacionado a uma limitação específica do algoritmo que talvez possa ser atenuada com o uso de heurísticas mais eficientes. Conforme discutido na Seção 4, uma rede com apenas 3 variáveis precisa de 1000 amostras para compor um conjunto de dados representativo. Segundo o princípio de *curse of dimensionality*, uma rede com mais variáveis deve ter o seu conjunto de dados acrescido exponencialmente para que esta representatividade se mantenha. No entanto, na prática o princípio não se confirmou. Para o experimento 3, envolvendo uma rede com 6 variáveis, com o mesmo número de amostras foi possível encontrar uma rede equivalente à original. Talvez o problema não seja tão crítico assim. Parece que a natureza do modelo é a grande determinante neste caso. A questão é que, se todas as relações causais são bastante intensas, isto é, suas conseqüências são observadas com grande probabilidade, um número relativamente pequeno de amostras é necessário para compor uma amostragem representativa. Mas se nestes mesmos termos uma das conexões é relativamente fraca, o conjunto amostral deve ser consideravelmente maior para incluir também os eventos menos prováveis de forma significativa. Ora, geralmente não é de estrita relevância ter acesso a esses pormenores, dado que um modelo aproximado contendo apenas as relações causais mais intensas seguramente possuirá robustez suficiente para explicar e generalizar a maioria dos fenômenos. É, portanto, de fundamental importância que as redes geradas revelem as conexões mais intensas, e para isso, não é necessário um conjunto amostral muito extenso.

Existe um outro ponto que merece destaque, e se refere às redes equivalentes de distribuição. A análise da Seção 5.2 mostrou que em algumas situações existem redes bayesianas com estruturas diferentes, mas que possuem exatamente a mesma densidade de probabilidade. Como argumentado em [Heckerman, 1998], nesses casos é impossível para qualquer algoritmo fazer a distinção entre os modelos baseado apenas nos dados. Isso leva então a um questionamento: o quão diferente podem ser duas redes equivalentes de distribuição e com que freqüência essa particularidade pode ocorrer? Primeiramente, se duas redes equivalentes de distribuição podem apresentar estruturas completamente diferentes, a escolha arbitrária pelo modelo errado pode trazer conseqüências desastrosas quando se está interessado nas relações causais, e não na distribuição em si. Esta, no entanto, não foi a situação observada nos experimentos. Segundo, se a ocorrência de redes equivalentes é freqüente, passa-se a não ter confiança alguma nos resultados encontrados, a não ser que a afirmativa primeira esteja errada. Esta é uma questão especial que deve ser investigada com cautela.

Falta comentar sobre o desempenho da abordagem proposta. Os testes mostraram que o algoritmo K2 utilizando como critério de qualidade a máxima verossimilhança deixou a desejar em várias circunstâncias. Em particular os experimentos realizados na Seção 5, deixaram claro que o algoritmo converge para ótimos locais com uma certa freqüência, sendo esta uma das razões pelas quais a estrutura original dos modelos não é recuperada. Além disso, foi visto que o critério de máxima verossimilhança tende a valorizar redes

mais complexas, o que leva a conexões não existentes na rede original e reduz a aplicabilidade prática dos modelos gerados.

Voltemos então à pergunta inicial. A abordagem aprendizado de redes bayesianas pode ajudar a resolver problemas complexos? A conclusão final deste relatório, embora ainda carente de embasamento em investigações mais profundas, é que sim. Com o uso de uma abordagem mais sofisticada, isto é, com heurísticas de busca mais eficientes e critérios de seleção de modelos mais consistentes, a tarefa de aprendizado redes bayesianas sem conhecimento a priori pode ajudar a encontrar as relações mais intensas entre as variáveis mesmo na ausência de um conjunto de amostras muito representativo, gerando por sua vez modelos que podem ajudar a entender os eventos associados a problemas de mundo real.

Referências

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

[Bellman, 1961] Bellman, R., *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.

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

[Forster, 2000] Forster, Malcolm R., "Key Concepts in Model Selection: Performance and Generalizability" *Journal of Mathematical Psychology*, 44, 205-231, 2000.

[Geard, 2004] Geard, Nicholas, "Modelling Gene Regulatory Networks: Systems Biology to Complex Systems", *ACCS Draft Technical Report*, 2004.

[Heckerman, 1997] Heckerman D. *et al.*, "A Bayesian Approach to Causal Discovery", Technical Report MSR-TR-97-05, 1997.

[Jacquette, 1944] Jacquette, D. "Ockham's Razor. Philosophy of Mind", Engleswoods Cliffs, N.J., Prentice Hall, pp. 34-36, 1994.

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