

Universidade de São Paulo
Instituto de Matemática e Estatística
MAC 5789 - Laboratório de Inteligência Artificial

Exercício Programa 4: Redes bayesianas

Autor:

Walter Perez Urcia

São Paulo

Junio 2015

Resumo

Neste trabalho o objetivo foi analisar um conjunto de dados e construir manualmente tres redes bayesianas. Além disso, se dividirá o conjunto de dados em dados de treinamento e dados de test. Também se mostrará a construção de cada uma das redes bayesianas e suas assunções para estabelecer as diferentes relações entre as características dos dados. Por último, as redes serão comparados usando o Data-LogLikelihood.

Sumário

1	Dados de entrada	5
1.1	Características	5
1.2	Discretização de características	5
1.3	Dados de treinamento e dados de teste	6
2	Construção manual de redes bayesianas	7
2.1	Primeira rede	7
2.2	Segunda rede	8
2.3	Terceira rede	9
3	Experimentos e resultados	11
3.1	Considerações gerais	11
3.1.1	Formulas	11
3.1.2	Algoritmo	11
3.2	Experimento 1: Corretude do algoritmo	12
3.3	Experimento 2: Busca com dados reais	14
4	Dificuldades no projeto	17
5	Conclusões	17

Lista de Figuras

1	Primeira rede bayesiana	7
2	Segunda rede bayesiana	8
3	Terceira rede bayesiana	9
4	BIC Score (em milhares) para cada rede	12
5	Melhor rede para o modelo 1	13
6	Melhor rede para o modelo 2	13
7	Melhor rede para o modelo 3	14
8	BIC Score (em milhares) para os dados reais	15
9	Melhor rede encontrada para o conjunto de dados	15

1 Dados de entrada

Os dados de entrada foram extraídos da página Machine Learning Repository [2] e o conjunto de dados que será usado neste trabalho é Adulto [1]. Este conjunto tem 48842 instancias, mas algumas não tem sua data completa. Todos os dados no conjunto são de um censo no ano 1994 feito por Barry Becker para determinar se uma pessoa tem mais de 50 mil em dinheiro por ano.

1.1 Características

As características do conjunto de dados na página são:

1. Age (AG): a idade da pessoa
2. Workclass (W): tipo de emprego que a pessoa é
3. Fnlwgt (F): Não se da descrição sobre esta característica
4. Education (ED): Máximo nível de educação a pessoa tem
5. Education-num (EN): Máximo nível de educação a pessoa tem em forma numérica
6. Marital-status (M): Estado civil da pessoa
7. Occupation (O): A ocupação da pessoa
8. Relationship (RE): Relações de família como esposa ou filho
9. Race (RA): Descrição da raça da pessoa
10. Sex (S): sexo biológico
11. Capital-gain (CG): Ganhos de capital
12. Capital-loss (CL): Perdas de capital
13. Hours-per-week (H): Horas trabalhadas por semana
14. Native-country (N): País de origem da pessoa
15. Annual-income (AI): Diz se a pessoa tem mais de 50 mil por ano

Das características anteriores, não será usada a terceira (Fnlwgt) por não ter suficiente informação sobre ela. Por outro lado, algumas características tem valores discretos, mas algumas como *Age* e *Capital – gain* são valores contínuos e portanto tem que ser discretizados, o que será feito na subseção 1.2.

1.2 Discretização de características

As características numéricas são:

- Age (AG)
- Education-num (EN)

- Capital-gain (CF)
- Capital-loss (CL)
- Hours-per-week (H)

Para cada um de eles, se calculo sua mediana e foi mudado para uma variável booleana. Então para uma característica X_i , temos que $Median(X_i)$ é sua mediana e todos seus valores mudaram da forma $x_{ij} = (1 \text{ se } x_{ij} > Median(X_i) \text{ e } 0 \text{ no caso contrario })$.

1.3 Dados de treinamento e dados de teste

Para poder testar e comparar as redes bayesianas que serão construídas nas seguintes seções temos que dividir os dados em dois conjuntos. O primeiro será usado para treinar e calcular todas probabilidades da rede bayesiana. Enquanto o segundo conjunto será usado para testar cada uma das redes bayesianas construídas. Para este trabalho os dados serão divididos em um 65% para treinar e um 35% para testar as redes. Os mesmos percentagens serão usados quando dados sejam gerados baseados nos modelos feitos manualmente.

2 Construção manual de redes bayesianas

2.1 Primeira rede

A Figura 1 mostra a rede bayesiana construída manualmente.

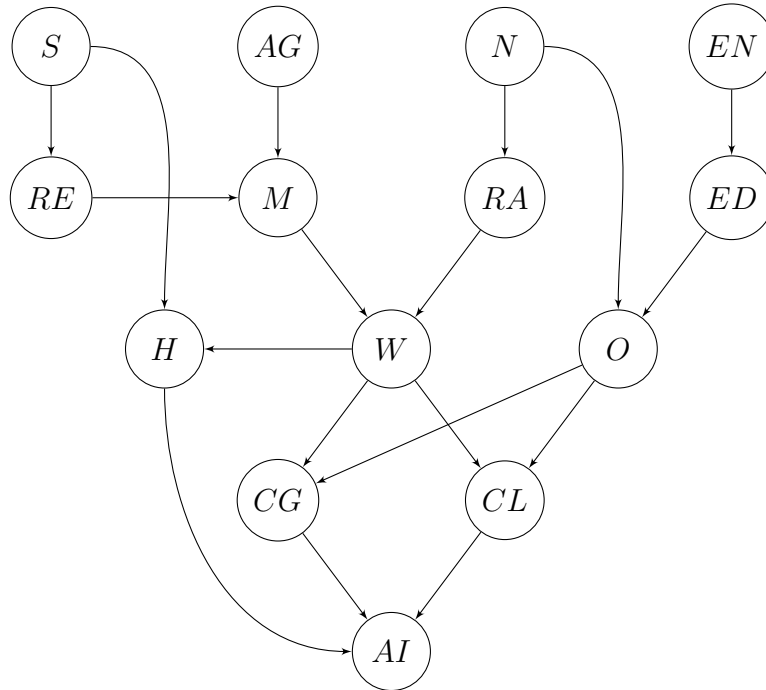


Figura 1: Primeira rede bayesiana

Para construir esta rede se levou em consideração as seguintes suposições:

- Sexo, idade e país de origem não tem como causa alguma das outras características porque na vida real não dependem de nada
- O país de origem da pessoa influi na raça, por exemplo é mais seguro que uma pessoa com rasgos asiáticos tenha um país de origem asiático
- O país de origem influi na ocupação da pessoa porque não existem todas as ocupações em todos os países do mundo (e.g. astronauta)
- O estado civil e a raça da pessoa influem no tipo de trabalho que tem (e.g. pessoas casadas podem ter trabalhos privados)
- O tipo de trabalho e o sexo influem nas horas de trabalho por semana porque existem trabalhos só para mulheres que não precisam muitas horas
- O tipo de trabalho e a ocupação influem nos ganhos e perdas de capital diretamente (e.g. algumas ocupações ganham mais dinheiro que outras em alguns casos, dependendo do tipo de trabalho)

- Para os ingresos anuais da pessoa, temos em consideração os ganhos, perdas de capital e as horas de trabalho semanais (e.g. uma pessoa que perde muito, tem poucos ingresos anuais, da mesma forma para alguém que ganha muito)

2.2 Segunda rede

Na seguinte figura se mostram todas as relações de dependencia que tem cada característica.

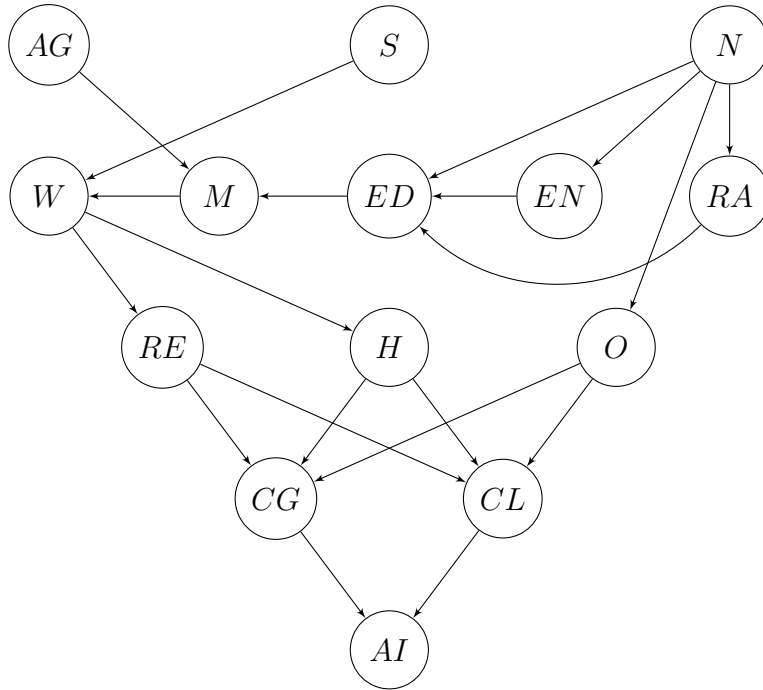


Figura 2: Segunda rede bayesiana

Tendo em consideração as seguintes suposições podemos construir esta rede:

- A educação, raza e a ocupação dependem do pais de origem da pessoa por as oportunidades que podem existir li
- O estado civil depende da idade e a educação que tem a pessoa (e.g. não muitos jovens se casam, gente que consegue nível universitario se casa)
- A educação depende da raza da pessoa em alguns lugares do mundo (com muito racismo)
- O sexo e o estado civil influi no tipo de trabalho pela mesma razão da rede anterior
- Dependendo o tipo de trabalho pode necessitar mais horas de trabalho e também influi no tipo de relação que a pessoa tem (e.g. solteiro, casado)
- Os ganhos e perdas de capital dependem das horas trabalhadas na semana porque mais horas trabalhadas podem significar mais ganhos

- Da mesma forma, uma melhor ocupação pode ganhar o perder mais que outras
- A relação também influi porque se gasta mais quando a pessoa está uma relação
- Os ingresos anuais dependem de quanto a pessoa ganha e perde capital

2.3 Terceira rede

Por último, a Figura 3 mostra a última rede bayesiana construída manualmente.

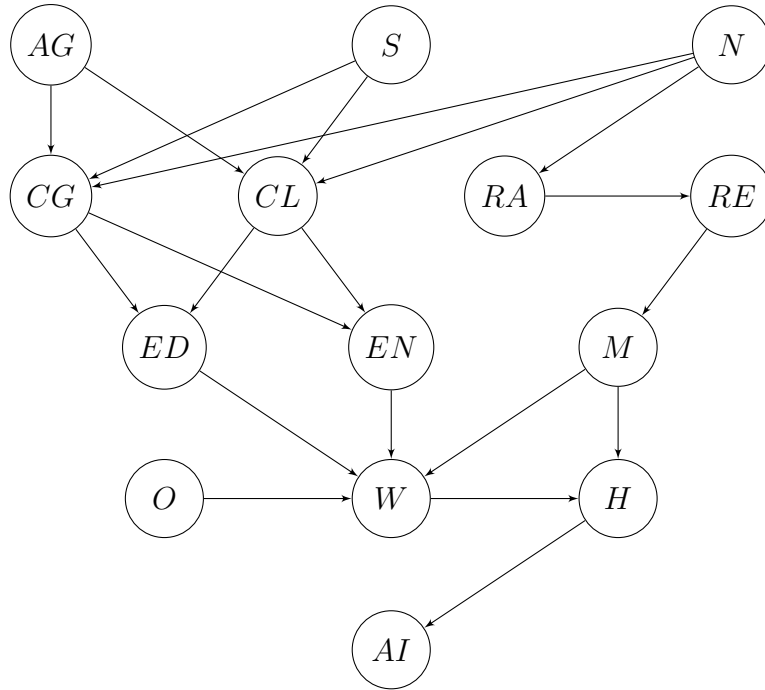


Figura 3: Terceira rede bayesiana

Para esta última rede se levou em consideração o seguinte:

- A idade, o sexo e o país de origem influem nos ganhos e perdas de capital porque uma pessoa mais jovem não tem muitas perdas, uma mulher pode ter muitas perdas por muitas compras e o país de origem pode ter uma crise que gere muitas perdas as pessoas.
- Dependendo dos ganhos e perdas de capital da pessoa pode conseguir um nível de educação maior
- O tipo de trabalho é afetado pela educação, o estado civil e a ocupação porque, por exemplo, algumas ocupações não são comumente trabalhadas de forma privada
- A ocupação da pessoa não depende de nada porque pode trabalhar em qualquer coisa que ele quizer

- As horas de trabalho semanais tem que ver com o estado civil (e.g. pessoas casadas trabalham mais que as solteiras) e o tipo de trabalho (e.g. trabalhos privados o independentes não tem um cronograma feito)
- Os ingresos anuais só dependem das horas trabalhadas semanalmente (e.g. se trabalha mais horas, pode ganhar mais)

3 Experimentos e resultados

3.1 Considerações gerais

3.1.1 Formulas

Para cada uma das redes se calculou todas suas probabilidades usando a propriedade seguinte:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i \mid Pa(X_i))$$

Onde $Pa(X_i)$ são todas características que são pais de X_i na rede bayesiana. Além disso, a probabilidade condicional pode ser calculada da seguinte forma:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Mas para calcular todas as probabilidades, se tem que avaliar cada variável X_i com os valores que aparecem os dados de entrada sendo a fórmula da seguinte forma:

$$P(A = a \mid B = b) = \frac{N(A = a \cap B = b) + \alpha_{ab}}{N(B = b) + \alpha_b}$$

Para este trabalho os valores de α foram uma estimação com base nos dados calculada com a formula:

$$\alpha_a = \frac{1}{|\Omega_a|}$$

Na formula anterior, se a fosse um conjunto de variáveis, o fator debaixo da fração muda para uma produtora de Ω_i .

3.1.2 Algoritmo

A seguinte lista mostra os parâmetros usados para a busca da melhor rede bayesiana e os valores que foram considerados para todos os experimentos:

- Número de ordens topológicas geradas: 100
- Número máximo de pais: O número que tem a rede feita manualmente e 3 para a busca com os dados reais
- Número de re-starts: 10

Então para construir uma rede foram geradas ordens topológicas e para cada característica se fez uma busca dos melhores pais que pode ter dada essa ordem topológica (os possíveis pais só foram as características anteriores a ela). O número de vezes que foi feito a busca dos melhores pais é igual ao número de re-starts.

3.2 Experimento 1: Corretude do algoritmo

Para poder testar que o algoritmo faz o que diz foi testado com os modelos feitos manualmente e especificados na seção 2. O objetivo do experimento foi usar o algoritmo para fazer uma busca de redes bayesianas e que a rede usada para gerar os dados seja encontrada entre as melhores. Caso contrario, o algoritmo não está correto.

Usando as probabilidades calculadas para cada rede feita manualmente foram gerados dados de treinamento e de teste com a mesma proporção especificada em 1.3 tendo em total 30 mil dados para cada uma.

Na tabela 1 mostra o parâmetro Data Log-Likelihood com os dados de teste gerados para cada uma e também o valor do parâmetro usando os dados reais. Os valores são

Rede	Real Data	Synthetic Data
1	-163794.41	-123773.44
2	-178879.85	-116475.45
3	-180462.16	-132307.87

Tabela 1: Data Log-Likelihood para cada rede

diferentes entre os dois conjuntos de dados porque a quantidade de dados é diferente, mas a proporção entre os valores é similar. A figura 4 mostra as mudanças do valor do parâmetro BIC Score em cada iteração do algoritmo.

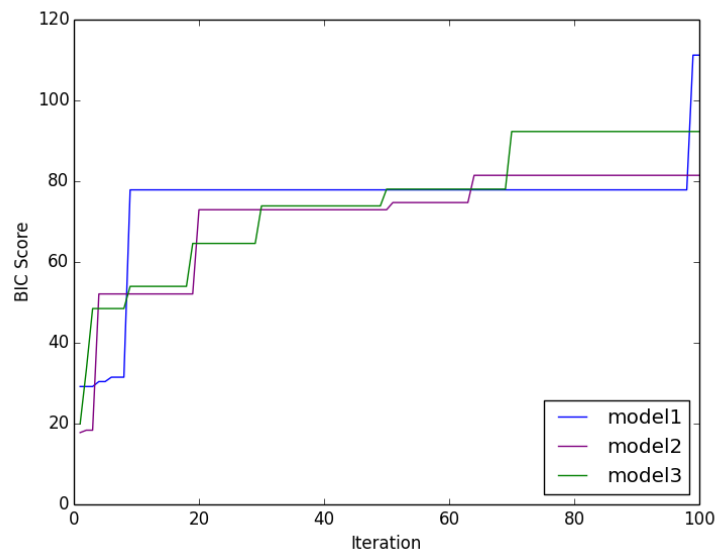


Figura 4: BIC Score (em milhares) para cada rede

Por último, o algoritmo encontrou a rede bayesiana usada para gerar os dados em todos os casos, mas nenhuma foi a melhor encontrada, mas sem muito cercana.

As figuras 5, 6 e 7 mostram as redes que o algoritmo encontrou para cada conjunto de dados.

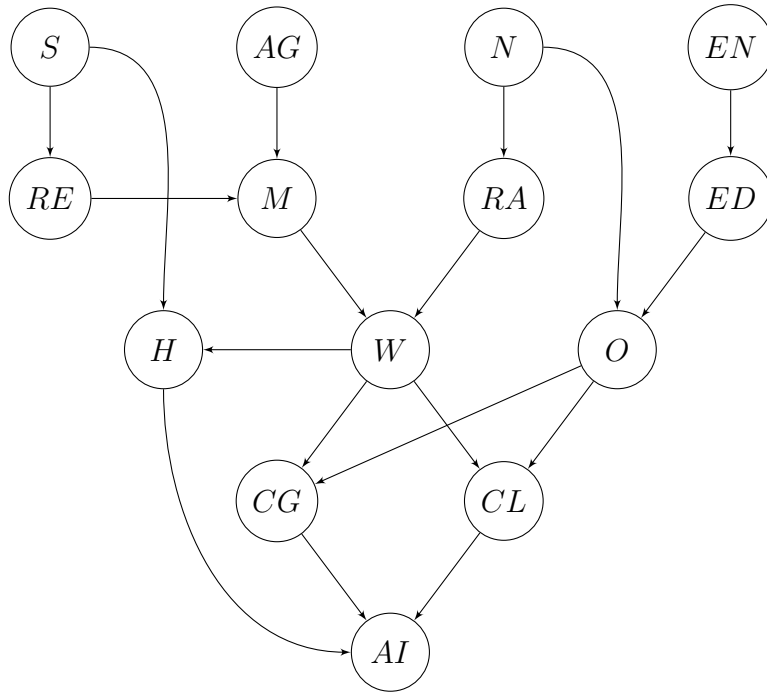


Figura 5: Melhor rede para o modelo 1

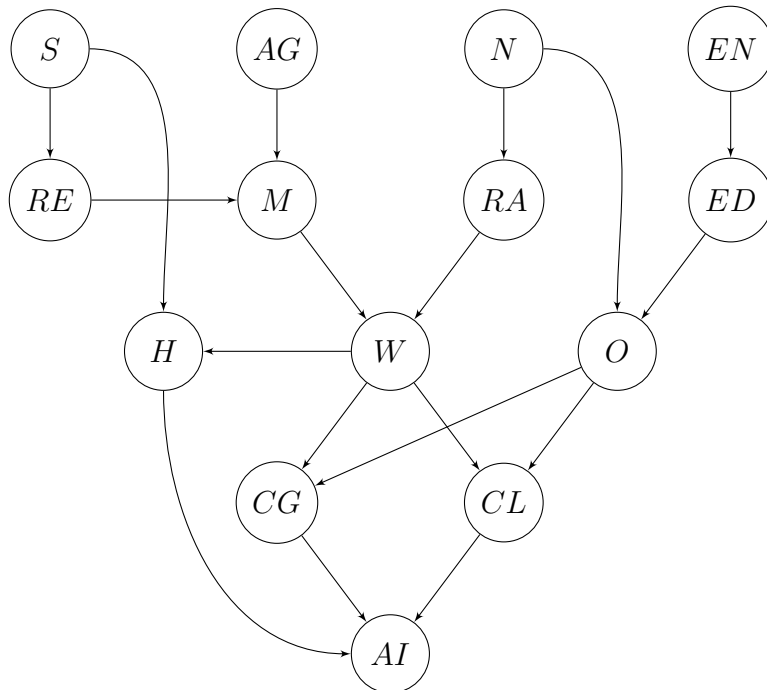


Figura 6: Melhor rede para o modelo 2

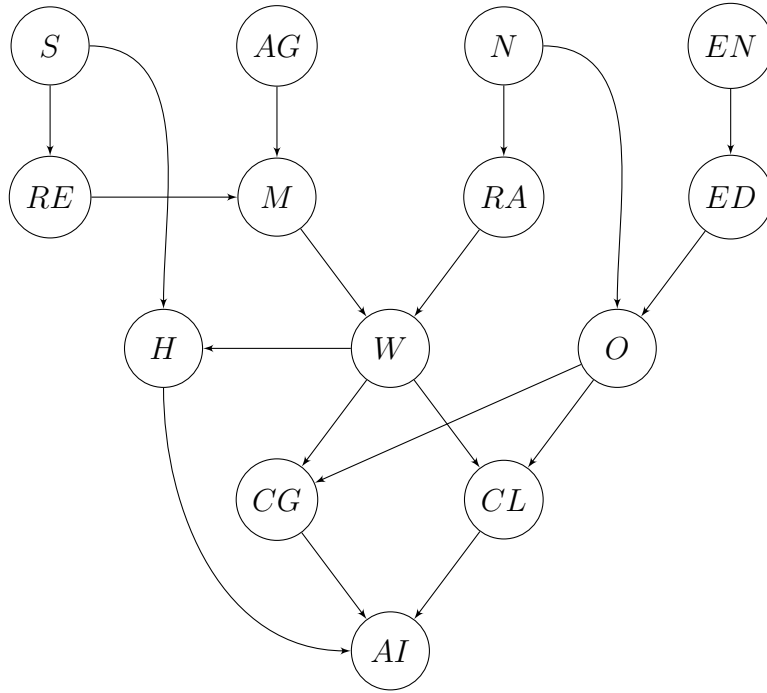


Figura 7: Melhor rede para o modelo 3

Portanto, fica demonstrado que o algoritmo faz a busca corretamente dado um conjunto de dados.

3.3 Experimento 2: Busca com dados reais

Finalmente o algoritmo foi executado com o conjunto de dados originais para fazer a busca da melhor rede. A figura 8 mostra o melhor valor do BIC Score para cada uma das iterações do algoritmo.

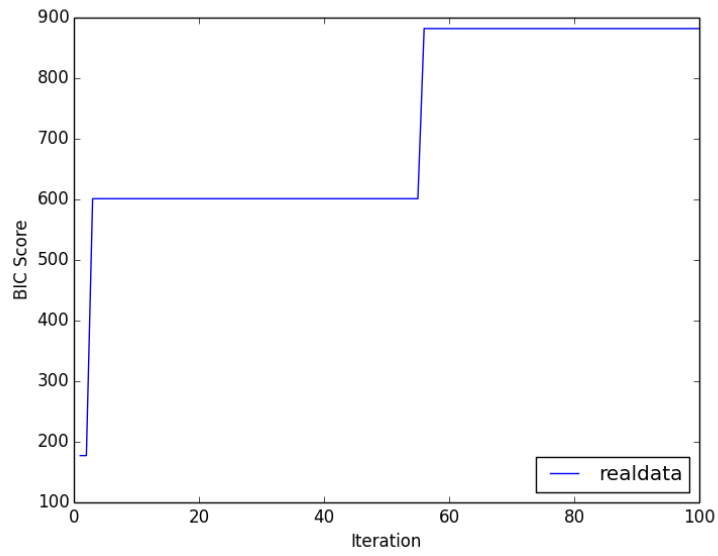


Figura 8: BIC Score (em milhares) para os dados reais

Se são comparados os valores com os que tinha na subseção 3.2, estes valores são muito maiores o que quer dizer que possivelmente tem uma maior quantidade de aristas no grafo. Ao final da execução do algoritmo, foi encontrada a rede mostrada na figura 9.

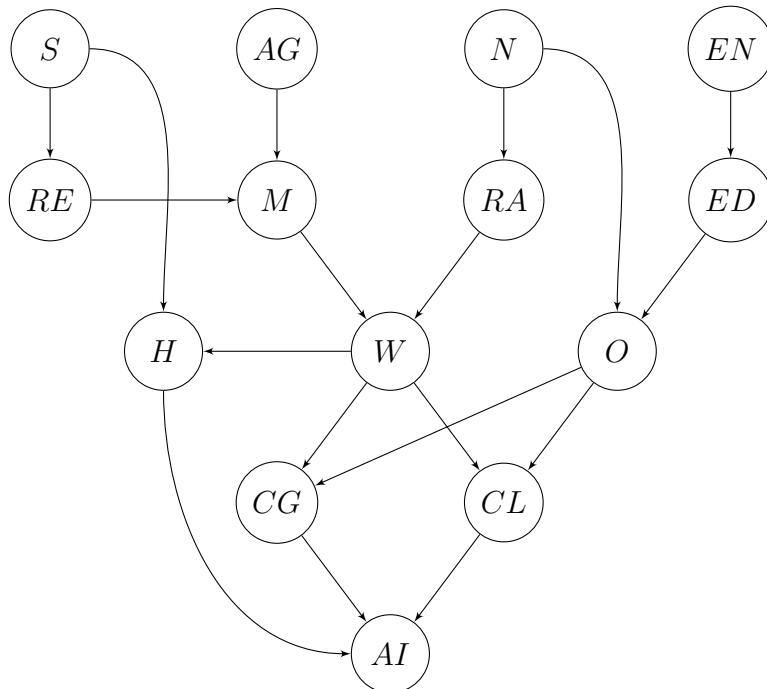


Figura 9: Melhor rede encontrada para o conjunto de dados

O valor do parâmetro Data Log-Likelihood para esta rede é $DLL = -156842.48$ que é só um pouco maior à rede 1 feita manualmente.

4 Dificuldades no projeto

Durante a execução do projeto se tiveram algumas dificuldades mencionadas a continuação:

- Muitos conceitos novos
- Cálculo de probabilidades com base nos dados de treinamento
- Cálculo de Data-Loglikelihood
- Entendimento do algoritmo para a geração de dados usando um modelo e suas probabilidades
- Foram precisadas muitas otimizações no algoritmo para não computar a mesma coisa muitas vezes

5 Conclusões

Pode-se concluir em geral que:

- A primeira rede é a melhor rede manual que foi construída comparando seu Data Log-Likelihood com as outras duas
- O algoritmo desenvolvido para a busca e construção de redes bayesianas usando um conjunto de dados é correto
- O algoritmo encontrou uma rede melhor que as feitas manualmente com Data Log-Likelihood igual a -156842.48

Referências

- [1] Barry Becker. Adult Data Set, 1996.
- [2] Center for Machine Learning and Intelligent Systems. UCI Machine Learning Repository, 2007.