Análise do Algoritmo Amostragem de Gibbs para Cálculo das Probabilidades Marginais em Redes Bayesianas

Thales Lange[1], Raimundo C. G. Teive[1]

[1]Mestrado em Computação Aplicada – Universidade do Vale do Itajaí (UNIVALI)
Caixa Postal 88.122-000 – São José – SC – Brasil

{thalesl, rteive}@univali.br

Abstract. *This paper analyzes and discusses the obtained results with the development of the Gibbs Sampling Algorithm, which is used to calculate the marginal probabilities in Bayesian Networks. To allow this discussion, firstly, the algorithm used in the development is detailed. Finally the obtained results with the tests with this algorithm are presented. In these tests three analysis criteria were taken into account: precision of marginal probabilities, influence of sample amount in the precision and the sample amount influence in the computational performance. Results have proved that the efficiency of this algorithm depends strongly on the number of samples. Besides, its application is more appropriated to small networks.*

Resumo. *Este artigo analisa e discute os resultados obtidos com a implementação do algoritmo Amostragem de Gibbs, algoritmo passível de utilização na resolução das probabilidades marginais em Redes Bayesianas. Para proporcionar essa discussão, detalha-se a abordagem algorítmica empregada na implementação, para assim apresentar os resultados obtidos com os testes. Este teste oferecem subsídios para análise de três critérios: precisão das probabilidades marginais, influência do número de amostras na precisão e no desempenho computacional. Resultados comprovam que a eficiência deste algoritmo depende fortemente do número de amostras, tendo sua aplicação mais apropriada para redes pequenas.*

1. Introdução

Redes Bayesianas caracterizam-se como técnica de Inteligência Artificial (IA) do paradigma simbólico, que promove o formalismo para o raciocínio em circunstâncias de crenças parciais e condições de incerteza, geralmente quando existe conhecimento incompleto/parcial do domínio do problema. [Pearl 1988]

Com intuito de utilizar Redes Bayesianas, percebe-se essencial o uso de algoritmos que calculem as probabilidades marginais (com ou sem presença de evidências), pois os cálculos das probabilidades marginais são de fundamental importância na aplicabilidade das Redes Bayesianas. [Jensen 2001] Ainda como motivador, destaca-se a abstração do modelo matemático/algorítmico.

A necessidade de abstração do modelo matemático/algorítmico surge em razão da complexidade NP-Difícil para obtenção das probabilidades marginais [Russell e Norvig 2004] e da necessidade de evitar que os usuários de Redes Bayesianas necessitem conhecer o funcionamento dos algoritmos, ou mesmo evitar que os algoritmos restrinjam as possibilidades de modelagem das Redes Bayesianas.

2. Rede Bayesiana utilizada nos ensaios

Com objetivo de facilitar o entendimento do algoritmo Amostragem de Gibbs e de servir como caso de uso para realização de ensaios de precisão das probabilidades marginais e de desempenho computacional, apresenta-se na Figura 1, a Rede Bayesiana para análise do merecimento do crédito, disponível como exemplo didático na documentação do GeNIe [GeNIe 2009]. Como motivadores para sua escolha, destacam-se: pequena quantidade de variáveis; o tema comum; topologia mista de conexões convergentes e divergentes; e média aproximada de três estados por variável.

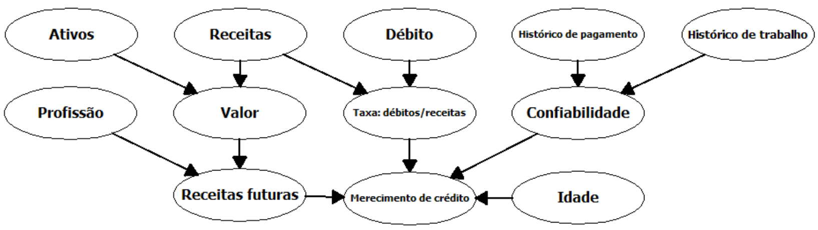

Figura 1. Rede Bayesiana para análise do merecimento do crédito.

3. Descrição do algoritmo Amostragem de Gibbs

O algoritmo Amostragem de Gibbs é uma forma particularmente conveniente da Cadeia de Markov Monte Carlo (CMMC). Em função de ser uma variante CMMC, as cadeias de Markov presentes em cada variável da Rede Bayesiana devem ser ergódicas. Em essência, uma variável é ergódica, quando todo estado é acessível a partir de qualquer outro estado e não pode existir nenhum ciclo estritamente periódico. [Russell e Norvig 2004]

Na Figura 2 ilustra-se o fluxograma do algoritmo Amostragem de Gibbs implementado. Observa-se como principal atividade do fluxograma, o cálculo P (Variável | Cobertura de Markov).

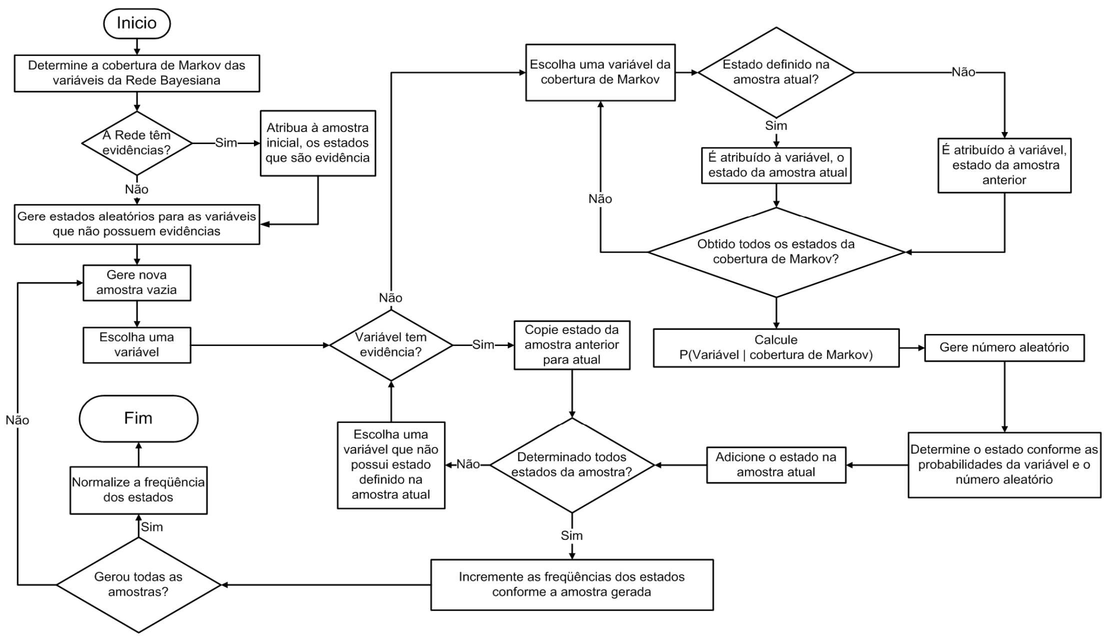

Figura 2. Fluxograma do algoritmo amostragem de Gibbs

A cobertura de Markov de uma variável A é composta pelos pais da variável A, os filhos Y da variável A e os demais pais das variáveis Y. [Russell e Norvig 2004]. Para exemplificar, enumera-se a cobertura de Markov da variável Taxa: débitos/receitas (variável da Rede Bayesiana da Figura 1): Receitas; Débitos; Merecimento de crédito; Confiabilidade; Receitas Futuras ; e Idade.

Para calcular a P (Variável | Cobertura Markov), utiliza-se a Equação 1, onde a_i representa um estado da variável A, mb(A) representa as variáveis da cobertura de Markov da variável A e Y_j são as variáveis filhos de A.

$$P\left(a_i | mb(A)\right) = P\left(a_i | pais(A)\right) * \prod P\left(y_i | pais\left(Y_j\right)\right) \qquad \text{Equação 1}$$

Ainda em relação ao fluxograma da Figura 2, percebe-se outra característica importante, o conceito de amostra. A amostra representa um vetor com tamanho igual ao número de variáveis presentes na Rede Bayesiana, que tem a finalidade de armazenar os estados sorteados em cada iteração. Para exemplificar, na Tabela 1 ilustra-se um possível estado de uma amostra. Observe que conforme o fluxograma da Figura 2, o algoritmo somente avança para a próxima iteração depois de definido todos os estados da amostra.

Tabela 1. Exemplo de uma amostra

Ativos	Receitas	Taxa: débitos/receitas	Ativos	...	Merecimento do crédito
Poucos	s0_30000	Favorável	Ø	E_n	Ø

Ao término dos sorteios dos estados das variáveis de uma iteração do algoritmo, contabiliza-se numa outra estrutura de dados, a frequência de ocorrência dos estados.

Ao final de todas as iterações, obtém-se a quantidade de vezes que cada estado foi sorteado (frequência). Com isso, possibilita-se a obtenção das probabilidades marginais através da normalização da frequência dos estados de cada variável.

4. Análise do algoritmo Amostragem de Gibbs

Esta seção tem o objetivo de analisar e discutir os resultados obtidos com a implementação do algoritmo Amostragem de Gibbs. A análise do algoritmo divide-se nos seguintes tópicos: precisão das probabilidades marginais e a influência do número de amostras tanto na precisão como no desempenho computacional.

4.1. Precisão das probabilidades marginais

Para a Rede Bayesiana do merecimento do crédito da Figura 1, pode-se observar na Tabela 2, que o algoritmo apresenta probabilidades marginais próximas das probabilidades exatas, obtidas no software Netica com o algoritmo *Junction Tree*. Para o ensaio dos resultados, utilizaram-se 5000 amostras e a existência das seguintes evidências: $\text{Valor}_{\text{Baixo}}$, $\text{Taxa: débitos/receitas}_{\text{Favorável}}$ e $\text{Merecimento do crédito}_{\text{Negativo}}$.

Ainda em relação à Tabela 2, observa-se a presença de erro percentual significativo para o estado Saudáveis. As simulações demonstram que geralmente os maiores erros percentuais encontram-se em estados com pequenos valores de probabilidade. Observa-se que em termos de erro absoluto, o estado Muito rentável apresenta erro mais crítico do que estado Saudáveis, por mais que o estado Saudáveis apresente erro percentual maior, ao menos no ponto de vista dos autores.

Tabela 2. Probabilidades marginais e erro para duas variáveis

Nome da variável	Nome do estado	Probabilidade (Gibbs)	Probabilidade (Juntion Tree - Netica)	Erro (%)	Erro absoluto
Ativos	Saudáveis	0,0062	0,0043	**43,58%**	0,0019
	Médios	0,3746	0,3643	2,83%	0,0103
	Poucos	0,6192	0,6314	-1,93%	-0,0122
Profissão	Muito rentável	0,2598	0,2433	6,79%	**0,0165**
	Rentável	0,3268	0,3341	-2,18%	-0,0073

4.2. Influência do número de amostras

Com objetivo de explorar a influência do número de amostras na precisão numérica das probabilidades marginais, na Figura 3 apresenta-se a média do erro de todos os estados e o erro máximo de cada simulação para cada quantidade de amostras. Observa-se o uso do módulo dos erros, para elaboração do gráfico.

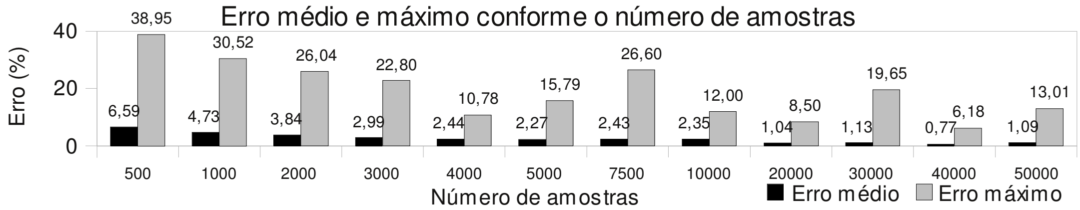

Figura 3. Erro médio e máximo conforme o número de amostras

Conforme se percebe na Figura 3, o acréscimo da quantidade de amostras diminui o erro médio e máximo. Entretanto, o aumento dessa quantidade influencia o desempenho computacional, uma vez que ocorre acréscimo da quantidade de cálculos. Para explorar essa consequência, na seguinte subseção apresenta-se a influência do número de amostras no desempenho computacional.

4.2.1. Influência do número de amostras no desempenho computacional

Com intuito de demonstrar o aumento do esforço computacional e a diminuição do erro médio, na Figura 4 apresenta-se o gráfico do erro percentual médio e tempo de execução versus o número de amostras. Observa-se o aumento logarítmico do tempo de execução conforme o aumento do número de amostras, entretanto o erro médio não diminui na mesma proporção. Percebe-se ainda, que com maiores quantidades de amostras, o algoritmo entra numa região de saturação, em outras palavras, o erro pouco diminui com o aumento da quantidade de amostras.

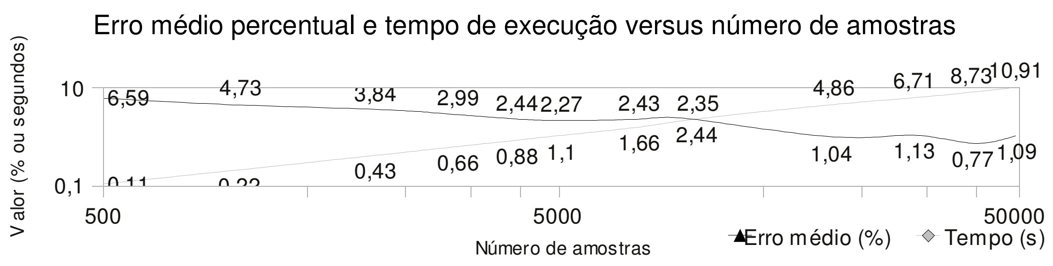

Figura 4. Erro médio e máximo conforme o número de amostras

No caso de uso desse algoritmo, constata-se necessário o uso adequado da quantidade de amostras, pois se o número for insuficiente, os erros associados tornam os resultados insatisfatórios, por outro lado busca-se não utilizar grandes quantidades de amostras, pois o esforço computacional pode coibir o seu uso, sem proporcionar melhorias significativas na precisão numérica dos resultados. Quando empregado esse algoritmo, deseja-se utilizar uma quantidade de amostras que apresente o melhor custo/benefício na relação precisão dos resultados e esforço computacional.

Entretanto, com os resultados apresentados no gráfico da Figura 4, admite-se que o algoritmo amostragem de Gibbs não proporciona a solução mais adequada para os cálculos das probabilidades marginais em Redes Bayesianas, pois além de se tratar de um algoritmo com solução aproximada, o fator desempenho torna proibitivo o uso de Redes Bayesianas de médio/grande porte.

Ainda em relação ao critério de desempenho computacional, constataram-se os resultados semelhantes aos resultados de Yuan e Druzdzel (2004), que também afirmaram que o algoritmo amostragem de Gibbs apresenta desempenho computacional insatisfatório na resolução de probabilidades marginais em Redes Bayesiana de grande porte ou mesmo na comparação com outros algoritmos aproximados e/ou exatos.

5. Conclusões

Com a implementação e realização de ensaios com o algoritmo Amostragem de Gibbs, permitiu-se mensurar o erro médio e máximo das probabilidades marginais. Verificou-se a influência do número de amostras na precisão numérica das probabilidades marginais, como também sua influência no desempenho computacional. Como principais conclusões obtidas com os resultados dos ensaios, destacam-se:

- a quantidade de amostras influencia a precisão dos resultados;
- erro percentual significativo para pequenos valores de probabilidade marginal;
- elevado número de amostras não proporcionam benefícios significativos à precisão dos resultados, uma vez que o algoritmo entra numa região de saturação; e
- desempenho computacional insatisfatório do algoritmo, que torna proibitivo o seu uso em Redes Bayesianas de maior porte ou em sistemas com restrição de tempo de resposta.

Referências

GeNIe 2.0 (2009) “GeNIe Reference Manual”, Decision Systems Laboratory, University of Pittsburgh, United States of America.

Jensen, F. V. (2001) “Bayesian Networks and Decision Graphs”.Springer-Verlag, 1ª edição, New York.

Pearl, J. (1988) “Probabilistic reasoning in intelligent systems: networks of plausible inference”, Elsevier, 2ª edição, United States of America.

Russell, S. J. e Norvig, P. (2004) “Inteligência Artificial”, Elsevier, 2ª edição, Rio de Janeiro.

Yuan, C. e Druzdzel, M. J. (2004) “Importance Sampling Algorithms for Bayesian Networks: Principles and Performance”, Proceedings of 19th Annual Conference on Uncertainty in Artificial Intelligence.