XXV Congreso Nacional de Estadística e Investigación Operativa
Vigo 4–7 de Abril del 2000

PROPAGACIÓN MONTE CARLO EN REDES BAYESIANAS USANDO VARIABLES CORRELACIONADAS

Antonio Salmerón[1], María Morales[1].

[1]Dpto. Estadística y Matemática Aplicada
Universidad de Almería

RESUMEN

En este trabajo proponemos un esquema aproximado para la estimación de las probabilidades a posteriori en una red bayesiana. El proceso se basa en la simulación mediante muestreo por importancia, en el que se introduce cierta dependencia entre las variables de la muestra con el objetivo de reducir la varianza de la estimación.

Palabras y frases clave: redes bayesianas, propagación de probabilidad, muestreo por importancia, Monte Carlo.
Clasificación AMS: 65C60, 68T37.

1. INTRODUCCIÓN

Una *red bayesiana* para un conjunto de variables $\{X_1, \ldots, X_n\}$, es un grafo dirigido acíclico donde cada nodo representa una variable aleatoria que tiene asociada una distribución de probabilidad condicionada a las variables que se corresponden con sus nodos padres en la red. La *propagación de probabilidad* consiste en obtener la probabilidad de ciertas variables de interés dado que se conoce el valor que toman algunas otras variables ($X_E = e$), es decir, $p(x'_k|e)$, para todo $x'_k \in U_k$, donde $k \in \{1, \ldots, n\}$ y U_k es el conjunto de posibles valores de X_k. Si el problema es suficientemente complicado, esas probabilidades condicionadas no podrán ser calculadas de forma exacta.

Para calcular la probabilidad de interés, dado que $p(x'_k|e)$ es igual a $p(x'_k, e)/p(e)$, y $p(e) = \sum_{x'_k \in U_k} p(x'_k, e)$, basta con calcular los valores $p(x'_k, e)$ para todo $x'_k \in U_k$, y normalizar después. Esta probabilidad conjunta podemos expresarla como

$$p(x'_k, e) \quad = \quad \sum_{x \in U_N} g(x), \tag{1}$$

donde $g(x) = \left(\prod_{i \in N} p_i(x_i|x_{F(i)})\right) \left(\prod_{j \in E} \delta_j(x_j; e_j)\right) \delta_k(x_k; x'_k)$ para todo $x \in U_N$ y $\delta_k(x_k; x'_k)$ es igual a 1 si $x_k = x'_k$ y 0 en otro caso.

2. MUESTREO POR IMPORTANCIA

Una forma de estimar dichas probabilidades es mediante la técnica de muestreo por importancia, que se basa en el uso de una función de muestreo auxiliar más fácil de simular que la distribución exacta. Concretamente, tomamos $p^* : U_N \to [0,1]$, verificando que $p^*(x) > 0$ para todo $x \in U_N$ tal que $g(x) > 0$. Entonces, podemos escribir la fórmula (1) como:

$$p(x'_k, e) = \sum_{\substack{x \in U_N, \\ g(x)>0}} \frac{g(x)}{p^*(x)} p^*(x) = E\left[\frac{g(X^*)}{p^*(X^*)}\right],$$

donde X^* es una v.a. con distribución p^*. Con esto, para estimar la probabilidad de interés podemos generar una muestra a partir de p^* y dar la media de los cocientes entre g y p^* como estimación.

Salmerón, Cano y Moral (1999) proponen un esquema para obtener las estimaciones de forma simultánea para todas las variables de interés, mediante un proceso de simulación en el que las variables se simulan de forma secuencial, obteniendo al final una configuración de todas las variables de interés, y usan *árboles de probabilidad* para representar de forma más eficiente las distribuciones de muestreo de cada variable.

3. VARIABLES ANTITÉTICAS

Nuestra propuesta se basa en la aplicación de una idea basada en el uso de *variables antitéticas*. Concretamente, en el proceso de simulación generamos simultáneamente dos configuraciones, de manera que cada vez que se va a generar un valor para una variable en una configuración usando un número aleatorio U, generamos otro valor a partir de $1-U$. de esta manera forzamos que la covarianza entre ambas variables sea negativa, con lo que la varianza total se reduce.

El algoritmo ha sido programado en lenguaje Java y se ha incorporado al software *Elvira*. Los resultados experimentales indican una reducción del tiempo de simulación y una disminución en el error en las estimaciones.

4. AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado por la CICYT bajo el proyecto TIC97-1135-C04-02.

5. REFERENCIAS

Rubinstein, R.Y. (1981). *Simulation and the Monte Carlo method.* Wiley.

Salmerón, A., Cano, A., Moral, S. (1999). "Importance Sampling in Bayesian Networks Using Probability Trees". Por aparecer en: *Computational Statistics and Data Analysis.*