# Comparando algoritmos de aprendizaje estructural en redes bayesianas estáticas

Trabajo Final de Tópicos para Ciencias de la Computación

1st Diego Ramirez
*Universidad Peruana de Ciencias Aplicadas*
Lima, Perú

***Resumen*—El aprendizaje estructural es de vital importancia en para el correcto modelamiento de las redes bayesianas por lo tanto, el área ha sido ampliamente estudiada. Existen diversos enfoques para lograr el aprendizaje estructural en redes bayesianas por lo que en este artículo se compararó los algoritmos K2 y Chow Liu ante el dataset sintético Asia. Los resultadso demostraron que Chow Liu tiene un menor tiempo de ejecución sin embargo conlleva un gran número de reestricciones por lo tanto, se determina que K2 debe usarse cuando el enfoque principal es obtener el grafo que más se adecue a la data mientras que Chow Liu se puede aplicar a contextos en los que se prioriza la velocidad.**

***Index Terms*—Redes Bayesianas estáticas, Aprendizaje estructural, Inferencias exáctas, K2, Chow Liu, Variable Elimination**

## I. Introducción

Las redes bayesianas, también conocidas como redes de creencias, pertenecen a la familia de modelos gráficos probabilísticos [1]. Representan conocimiento previo sobre un cierto dominio, el cual está constituido por eventos aleatorios que en ciertos casos se encuentran relacionados entre sí. Estas dependencias pueden ser interpretadas en grafos acíclicos dirigidos que describen la dirección de las relaciones. "Los arcos indican la existencia de conexiones causales directas entre variables" [2]. Variables fuertemente relacionadas son expresadas en termino de sus probabilidades condicionales.

Estas estructuras relacionales son vitales para conocer e interpretar correctamente la data, por lo que el aprendizaje estructural es un concepto ampliamente estudiado en el área de modelos gráficos probabilísticos, sin embargo, a medida que los dominios se vuelven más complejos el espacio de búsqueda se amplía exponencialmente. Existen amplias investigaciones que aportan algoritmos de búsqueda tales como la de F. Glover [3] que propone la aplicación de Tabu Seach o la de A.S Hesar [4] que propone una búsqueda por Simulated Annealing de esta forma limitan el espacio de búsqueda mediante el uso de heurísticas sacrificando calidad por velocidad de procesamiento.

Las redes bayesianas pueden ser explotadas para predecir eventos futuros sin embargo, como se demuestra en el aporte de Cooper [5], "la inferencia probabilística usando redes de creencias generales son un proceso NP-Hard. En particular aquellas con variables no inicializadas" [5], lo que implica que no puede ser resuelto en tiempo polinomial. Diversos algoritmos se diseñaron con la finalidad de simplificar el proceso tales como Beleif Propagation por J. Pearl [6] y MAP propuesto por R. Bassett y J. Deride [7].

Como ya se mencionó, la investigación en estas áreas es amplia, por lo tanto, puede resultar complicado para nuevos investigadores diferenciar en que contextos sobresale tipo de algoritmo tanto de búsqueda como inferencia. En este artículo se realiza una comparación entre los algoritmos de aprendizaje estructural K2 de G. Cooper y E. Herskovit [8] contra Chow Liu [9] y, se comparan los resultados de las inferencias mediante el uso de Varaible Elimination [10]. De esta forma se busca establecer un punto de partida para aquellos interesados en ahondar en el tema.

Como aporte adicional se ha decidido ahondar brevemente en el tema de la carga ética que conlleva el desarrollo de modelos probabilísticos en machine learning tanto en la investigación como en el producción de software. En el área de la investigación es imperativo dar el credito intelectual a quien haya colaborado ya sea formando parte activa del equipo de investigación como aquellas personas que previamente han realizado articulos que sientan las bases para furutos trabajos. En caso de no seguir esta regla el trabajo y esfuerzo de una persona no es reconocido, a demás podria conllevar a robos intelectuales. Otra punto a tomar en cuenta es la seguridad de los datos. Miles de personas donan datos personales a investigadores con la finalidad de que se puedan usar en sus trabajos, estos datos son por lo general confidenciales y no protegerlos pone en riezgo la seguridad y privacidad del los individuos. Por último es importante notar que la veracidad de los datos presentados y las interpretaciones finales tienen que estar correctamente respaldadas en el artículos ya que otros investigadores pueden no tener los recursos para comporbar empíricamente los resultados y al asumirlos ciertos se esta generando un daño al progreso científico.

## II. Estado del Arte

Diversos algoritmos de búsqueda han sido propuestos con anterioridad con la finalidad de encontrar métodos más eficientes para el aprendizaje estructural. Los principales grupos son los "Basados en Restricciones", "Basados en Puntaje"".

### *II-A. Algoritmos Restrictivos*

Los algoritmos de búsqueda por restricción aplican test de independencia condicional para encontrar el DAG que conlleva a las d-separaciones correspondientes. [11]Entre estos algoritmos se encuentran Max-Min Parents & Childs [12] que combina las ideas de técnicas de aprendizaje local, restrictiva, entre otros. Primero reconstruye el esqueleto de la red y luego realiza una búsqueda greedy por Hill Climbing utilizando el Puntaje Bayesiano [13] para orientar el sentido de las aristas. Los resultados de esta investigación demostraron que su propuesta se desempeña mejor que otros algoritmos restrictivos modernos.

### *II-B. Algoritmos basados en puntaje*

Los algoritmos basados en puntaje buscan encontrar un grafo que maximice la heurística retornada por una métrica de calidad. Sin embargo "esto posee considerable problemas desde que el espacio para toldas las posibles estructuras es al menos exponencial al número de variables". [14] En este grupo existen algoritmos como Tabu Search [3] y Simulated Annealing [4].

Número de posibles aristas: n(n-1)/2
Número de posibles estructuras: $2^{n(n-1)/2}$ para cada subconjunto de aristas
n: Número de variable(1)

Respecto a las métricas de calidad, existe una amplia variedad de aplicables tantos a redes estáticas como a dinámicas, que explotan diferentes propiedades tales como Entropía, Correlación, etc.

### *II-C. Bayesian Dirichlet*

La métrica BD propuesta por David Heckerman, Dan Geiger y David M. Chickering [13] combina conocimientos previos de modelos de dependencia y data estadística para aprender la estructura de la red en cuestión. Su propuesta establece ciertas suposiciones entre ellas que la data no ayuda a discriminar sobre la estructura de la red que según el artículo es la misma afirmación que hace la independencia condicional.

### *II-D. Bayesian information criterion*

La métrica BIC [16] es un criterio de selección fuertemente relacionado con AIC [15]. Ambos métodos favorecen grafos con menor cantidad de aristas al añadir una penalidad por agregar parámetros al modelo. Cabe resaltar que los resultados que se puedan obtener con BIC resultarán ser muy cercano a los que se obtengan por MDL [17] pero con el signo opuesto.

## III. Fundamento Teórico

### *III-A. Entropía*

La entropía mide la incertidumbre en una fuente de información. Esta definición quiere decir que aquella fuente de información donde exista gran variedad en los datos, tendrá mayor entropía debido a que la redundancia en los datos es menor. [18]

### *III-B. Métrica de calidad K2*

K2 fue propuesta por G.F. Cooper y E. Herskovits [8], se caracteriza por recibir una agrupación de varaible que determinan en gran medida el grafo resultante. Esto se debe a que en K2, a partir de una agrupación de varaibles V={$X_1$,$X_2$...,$X_n$}, solo podrá existir aristas de $X_i$ a $X_j$ cuando i $>$ j;

```
1. procedure K2;
2. {Input: A set of n nodes, an ordering on the nodes, an upper bound u on the
3.         number of parents a node may have, and a database D containing m cases.}
4. {Output: For each node, a printout of the parents of the node.}
5. for i := 1 to n do
6.    π_i := ∅;
7.    P_old := g(i, π_i); {This function is computed using equation (12).}
8.    OKToProceed := true
9.    while OKToProceed and |π_i| < u do
10.        let z be the node in Pred(x_i) − π_i that maximizes g(i, π_i ∪ {z});
11.        P_new := g(i, π_i ∪ {z});
12.        if P_new > P_old then
13.            P_old := P_new;
14.            π_i := π_i ∪ {z}
15.        else OKToProceed := false;
16.   end {while};
17.   write('Node:', x_i, 'Parents of this node:', π_i)
18. end {for};
19. end {K2};
```

Gráfico 1: Pseudocódigo de búsqueda K2 [8]

### *III-C. Métrica de calidad K2*

La métrica K2 recibe su nombre debido a su introducción con el algoritmo de busca de grado K2, más tarde sería extendida a la métrica Bayesiana de Dirichlet [13]. K2 funciona bajo las suposiciones de (1)El proceso que genero la data puede ser modelado de forma certera en una red bayesiana, (2)Dado un modelo, cada caso es independiente y (3)Los casos son completos. Bajo esas suposiciones llega a la conclusión que conlleva a la siguiente formula:

$$P(B_S, D) = P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

Formula 1: Formula K2 para el puntaje de estructuras [8]

Donde,

i: Cada variable en la estructura $B_S$.

j: Cada posible instanciación de los valores de variables padres de i.

k: Cada posible instanciación de los valores de i.

$P(B_S)$: Es la probabilidad de ocurrencia de la estructura $B_S$.

$r_i$): Es la cardinalidad sobre la variable i.

$N_{ij}$): Conteo del número de instancias de combinación

de valores j para la variable i.

En caso se cumpla j = 0 entonces se marginaliza j.

$N_{ijk}$): Conteo del número de instancias de combinación de valores j con la combinac para la variable i.

Sin embargo, a nivel computacional resulta complicado realizar las productorias propuestas ya que los números a comparar serian muy pequeños. Por lo tanto, se logró reformular la propuesta a una sumatoria de logaritmos.

$$P(B_S, D) = \sum_{i=1}^{n} \sum_{j=1}^{q} log(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!}) + \sum_{k=1}^{r_i} log(N_{ijk}!)$$

Formula 2: Formula logarítmica K2 para el puntaje de estructuras [8]

Donde,

i: Cada variable en la estructura $B_S$.

j: Cada posible instanciación de los valores de variables padres de i.

k: Cada posible instanciación de los valores de i.

$r_i$): Es la cardinalidad sobre la variable i.

$N_{ij}$): Conteo del número de instancias de combinación de valores j para la variable i.

En caso se cumpla j = 0 entonces se marginaliza j.

$N_{ijk}$): Conteo del número de instancias de combinación de valores j con la combinac para la variable i.

### *III-D. Kruskal*

El algoritmo de Kruskal [19] permite encontrar el árbol de expansión mínima expansión sin embargo la heurística que se usara el máximo peso. Dado un grafo inicial G, un conjunto de aristas e y sus respectivos pesos p se ordena las aristas en base al peso de cada una, luego se inicia un loop en el que se evalúa la arista de mayor, pero, en caso no genere un ciclo se añadirá al grafo y se continuará la siguiente iteración hasta que se cuente con n-1 aristas donde n es el número de nodos

### *III-E. Mutual Information*

"En teoría de la probabilidad, y en teoría de la información, la información mutua o transinformación de dos variables aleatorias es una cantidad que mide la dependencia mutua de las dos variables, es decir, mide la reducción de la incertidumbre (entropía) de una variable aleatoria, X, debido al conocimiento del valor de otra variable aleatoria Y". [20]

$$I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) log \frac{P(x_i|y_j)}{P(x_i)})$$

Formula 3: Test de independencia condicional Mutual Information [20]

### *III-F. Chow Liu*

El algoritmo Chow Liu [9] para arboles de expansión(spanning trees en ingles) logra expresar una distribución n-dimensional para una red de n variables al encontrar un árbol de dependencias que concuerde con la data de aprendizaje. El algoritmo inicial al generar un grafo no dirigido en el cual cada nodo tenga una arista a todos los demás nodos, luego genera un peso para cada arista donde el peso esta dado por el resultado de la función Mutual Information. Teniendo los pesos de cada arista ($x_i$,$y_j$) solo queda generar el árbol a partir de un algoritmo de expansión máxima tal como Kruskal.

### *III-G. Variable Elimination*

Variable Elimination "es un algoritmo de inferencia exacta en modelos graficos probabilísticos, tales como redes bayesianas." [10] Dado un grafo de dependencias G, un conjunto de distribuciones D y una inferencia Q(X—e) donde X es una variable perteneciente de la red G y e es un conjunto de evidencias perteneciente a G se descompone por regla de la cadena la inferencia requerida. Una vez descompuesta, se realiza una suma de productos donde la suma es la marginalización de cada variable no observada. Una vez terminado el proceso de marginalización, se tendrá Q(X,e) por lo que se tiene que dividir sobre la probabilidad de Q(e),dado que Q(X—e)= Q(X,e)/Q(e).

$$P(V|e) = \frac{1}{Z} \prod_{i=1}^{n} P(X_i|\pi_i)$$

Formula 4: Inferencia de variable V con la evidencia e [10]

Donde,

Z: Es una constante de normalización

V: Es una variable a inferir

e: Es la evidencia conocida

## IV. MÉTODO

Con la finalidad de comparar los algoritmos de búsqueda, se desarrolló un motor de redes bayesianas en C++ y se implemento diversas funcionalidades tales como leer archivos, calcular diversas probabilidades, aplicar hiper parámetros, encontrar el grado de independencia condicional entre dos variables por medio de Mutual Information y Pearson, encontrar el grafo según diversos algoritmos como K2, Fuerza Bruta, Chow Liu aplicando la métricas K2, por último el motor permite realizar inferencias en base a un modelo entrenado. Con este motor se va a realizar comparaciones en base a diversos datasets sintéticos. Las comparaciones abarcaran, tiempo de ejecución, complejidad del modelo y resultados de inferencias.

## V. EXPERIMENTOS

Los experimentos se realizan ante el dataset sintético Asia (Lauritzen and Spiegelhalter 1988), este cuenta con 8 variables y 5000 instancias. Asia modela un estudio realizado respecto a diversas enfermedades pulmonares y sus diferentes relaciones

con eventos tales como fumar, visitas a Asia, etc. Sobre este dataset se aplica un hiper-parámetro de Dirichlet equivalente a uno con la finalidad completar el dataset en caso se necesite. Posteriormente se aplicará el algoritmo Chow Liu para encontrar el grafo de dependencias y se realizará un conjunto de inferencias para verificar la capacidad del motor.

## VI. Resultados

Comparación de modelos

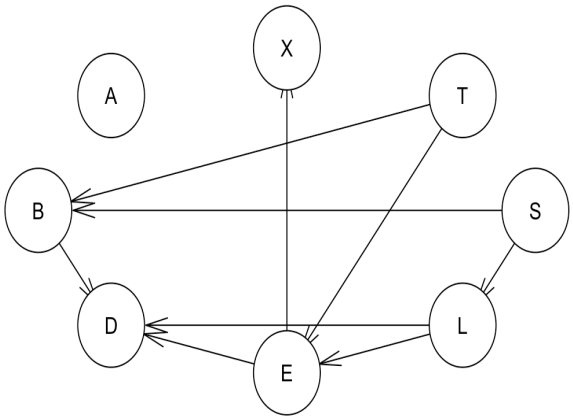


Gráfico 2: Modelo de dependencias obtenido por K2

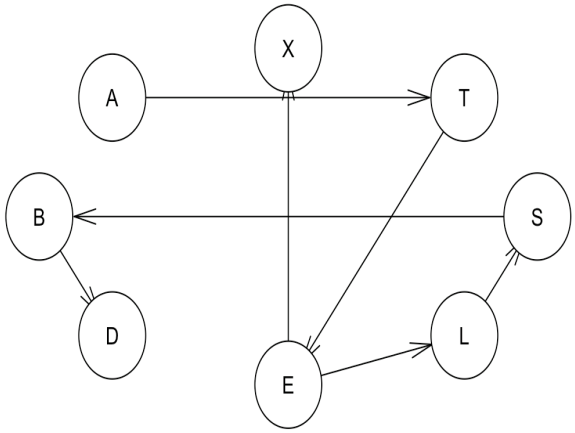


Gráfico 3: Modelo de dependencias obtenido por Chow Liu

| | Num aristas | Nodos indepen | Max Num padres |
|---|---|---|---|
| K2 | 9 | 1 | 2 |
| CL | 7 | 0 | 1 |

Tabla 1: Comparación de complejidad de los modelos

Como se puede apreciar en el Gráfico 2 y Gráfico 3, los modelos presentan características similares, K2 cuenta con 9 aristas mientras que Chow Liu presenta 7, respecto a los nodos independientes K2 cuenta con un nodo independiente mientras que Chow Liu no presenta nodos independientes. Por último K2 presenta un máximo de dos padres por nodo mientras que Chow Liu presenta uno. Sin ebmargo, ahi terminan las similitudes ya que si observamos cuidadosamente, ambos algoritmos solo coinciden en tres aristas lo que representa menos de la mitad de las aristas condicieron.

| | T(segundos) |
|---|---|
| K2 | 1.2390s |
| CL | 0.4598s |

Tabla 2: Comparación de tiempo de ejecución de los modelos

Respecto al tiempo de ejecución, se puede apreciar que el algoritmo K2 obtuvo casi tres veces más tiempo de ejecución que Chow Liu. Esto se debe a que K2 tienen que compara diversos modelos mientras que Chow Liu solo observa relaciones de entre variables sin observar el modelo como un todo.

Comparación de infernicias

Inferencia 1 P(T$\|B = yes, A = no, S = no, X = yes$)

| | Resultado | Probabilidad |
|---|---|---|
| K2 | No | 0.7625 |
| CL | No | 0.9609 |

Tabla 3: Comparación resultados de en inferencia 1

Inferencia 2 P(S$\|B = no$)

| | Resultado | Probabilidad |
|---|---|---|
| K2 | No | 0.7103 |
| CL | No | 0.7103 |

Tabla 4: Comparación resultados de en inferencia 2

Como se puede observar, en la primera inferencia se obtiene la misma conclusión sin ebmargo, la probabilidad de acierto varia en 0.2 aproximadamente esto se debe la falta de dependencia de T respecto a A en el grafo resultante de K2. Por otro lado en la Tabla 4 podemos observar que los resultados son iguales esto se debe a que en ninguno de los grafos S depende de B.

## VII. Conclusiones

En base a los experimentos realizado, se ha logrado determinar que Chow Liu es más rapido que K2 en cuanto tiempo de ejecuión, sin embargo Chow Liu tiene limitaciones severas tales como no permitir nodos independientes, como es el caso de K2. Respecto a las inferencias se puede concluir que en el caso epxlorado los resultados finales no varian mucho sin embargo cabe resaltar que Chow Liu tiene una dirección de aristas muy diferente a la presentada por K2 por lo que se puede asumir que otras inferencias pueden obtener resultados distinos a los observados. Se logró determinar empíricamente que K2 debe usarse cuando el enfoque principal es obtener el grafo que más se adecue a la data mientras que Chow Liu se puede aplicar a contextos en los que se prioriza la velocidad.

## Referencias

[1] F. Ruggeri, R. Kenett and F. Faltin, Encyclopedia of statistics in quality and reliability. Chichester, England: John Wiley, 2007.

[2] P. Larranaga, M. Poza, Y. Yurramendi, R. Murga and C. Kuijpers, ”Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 9, pp. 912-926, 1996.
[3] F. Glover, ”Tabu Search—Part I”, ORSA Journal on Computing, vol. 1, no. 3, pp. 190-206, 1989.
[4] A.S Hesar, ”Structure learning of Bayesian belief networks using simulated annealing algorithm”, Middle-East J. Sci. Res. 18, pp. 1343–1348, 2013.
[5] G. Cooper, ”The computational complexity of probabilistic inference using bayesian belief networks”, Artificial Intelligence, vol. 42, no. 2-3, pp. 393-405, 1990.
[6] J. Pearl, ”Belief Propagation in Hierarchical Inference Structures”, UCLA-ENG-CSL-8211, 1982.
[7] R. Bassett and J. Deride, ”Maximum a posteriori estimators as a limit of Bayes estimators”, Mathematical Programming, 2018.
[8] G. Cooper and E. Herskovits, .A Bayesian method for the induction of probabilistic networks from data”, Machine Learning, vol. 9, no. 4, pp. 309-347, 1992.
[9] C. Chow and C. Liu, .Approximating discrete probability distributions with dependence trees”, IEEE Transactions on Information Theory, vol. 14, no. 3, pp. 462-467, 1968.
[10] Zhang, N.L., Poole, D, .A Simple Approach to Bayesian Network Computations”, pp. 171–178
[11] S. Triantafillou and I. Tsamardinos, Score based vs constraint based causal learning in the presence of confounders. Heraklion, 2016, pp. 1-3.
[12] I. Tsamardinos, L. Brown and C. Aliferis, ”The max-min hill-climbing Bayesian network structure learning algorithm”, Machine Learning, vol. 65, no. 1, pp. 31-78, 2006.
[13] D. Heckerman, D. Geiger and D. Chickering, Machine Learning, vol. 20, no. 3, pp. 197-243, 1995.
[14] D. Margaritis, Learning Bayesian Network Model Structure from Data. Pittsburgh, 2003, pp. 18-23.
[15] H. Akaike, .A new look at the statistical model identification”, IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, 1974.
[16] .An Introduction to Bayesian Analysis”, Springer Texts in Statistics, 2006.
[17] G. Schwarz, .Estimating the Dimension of a Model”, The Annals of Statistics, vol. 6, no. 2, pp. 461-464, 1978.
[18] C. Shannon, .A Mathematical Theory of Communication”, Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, 1948.
[19] J. Kruskal, .On the shortest spanning subtree of a graph and the traveling salesman problem”, Proceedings of the American Mathematical Society, vol. 7, no. 1, pp. 48-48, 1956.
[20] D. Marinescu and G. Marinescu, Classical and quantum information. Burlington, MA: Academic Press, 2012.