Revista de Matemática: Teoría y Aplicaciones 2003 **10**(1–2) : 37–56
cimpa – ucr – ccss issn: 1409-2433

ESTRATEGIA PARA LA SELECCIÓN DE MODELOS GRÁFICOS EN TABLAS DE CONTINGENCIA NO ESTRUCTURADAS

Adalberto González Debén * Karen Alfonso Sagué †

Recibido: 11 Agosto 2003

Resumen

Los modelos loglineales se utilizan con frecuencia con el objetivo de analizar los patrones de asociación entre las variables que conforman una tabla de contingencia multidimensional. Recientemente se ha popularizado la utilización de los modelos gráficos discretos, que son un subconjunto de los modelos loglineales jerárquicos cuya estructura de independencia condicional se representa de manera única por un grafo, lo que facilita la interpretación de los mismos. En este trabajo se describen y comparan cuatro métodos de selección de modelos gráficos discretos: hacia atrás, hacia delante, paso a paso y automático EH. Asimismo, se propone una estrategia general de selección de modelos.

Palabras clave: modelos gráficos, tablas de contingencia, métodos de selección de modelos.

Abstract

Loglinear models are frequently utilized in the analysis of association patterns between variables in a multidimensional contingency table. Discrete graphical models are fashionable. They are a subset of the hierarchical loglinear models and its conditional independence structures are represented by a graph. Because of that they can be more easily interpreted. In this work four models selection methods are described and compared: stepwise selection (backward, forward and in two steps) and the EH procedure.A strategy for model selection is proposed.

Keywords: graphical models, contingency tables, model selection methods.

Mathematics Subject Classification: 62H17

*Grupo de Estadística, Instituto de Cibernética, Matemática y Física, CITMA, Cuba, adal@cidet.icmf.inf.cu

†Departamento de Bioestadística, Instituto de Higiene, Epidemiología y Microbiología, MINSAP, Cuba, Kalfonso@heinsa.sld.cu

1 Introducción

La modelación gráfica es una forma de análisis multivariado en el cual se utilizan grafos para representar modelos. En ella se estudian los modelos gráficos, éstos son modelos probabilísticos para observaciones multivariadas cuya estructura de independencia se caracteriza de manera única por un grafo.

En la última década se han propuesto aplicaciones de los modelos gráficos que, por su importancia, han significado la creación de nuevas líneas de investigación. Dos ejemplos son el área de sistemas expertos [12], y el área de computación evolutiva [16].

Este trabajo consta de tres epígrafes además de la introducción y las conclusiones. En el segundo epígrafe se presentan brevemente las definiciones y resultados más importantes de los modelos gráficos discretos. En [1] aparecen estos resultados, y otros más que los complementan, con sus demostraciones. En el tercer epígrafe se presentan cuatro métodos de selección de modelos gráficos, se comparan entre sí y se propone una estrategia general para la selección de modelos en tablas de contingencia no estructuradas (donde no hay distinción entre variables explicatorias y de respuesta, sino que todas tienen el mismo status). En el cuarto epígrafe se ilustran los cuatro métodos y la estrategia a través de un ejemplo muy conocido de la literatura.

1.1 Conceptos básicos

Sea $K = \{1, \ldots, k\}$ el conjunto de índices de una tabla de contingencia y x_i el valor tomado por la i-ésima variable, entonces $x = (x_1, \ldots, x_k)$ denota una celda de la tabla y $X = (X_1, \ldots X_k)$ el vector aleatorio. La probabilidad asociada a una celda es $p(x) = Prob(X = x)$. Se define la función de masa de probabilidad de un vector aleatorio k dimensional X como $f_K(x) = p(x)$ donde p es la tabla de probabilidades y es tal que $\forall\ x$ $p(x) > 0$ y

$$\sum_x p(x) = 1.$$

En este trabajo no se considera el caso de celdas con ceros estructurales, o sea, $p(x) = 0$.

Por lo general se utilizan, para describir la forma en que se obtienen los datos, las distribuciones Poisson, multinomial o producto de multinomiales [3], [19].

Sea $X = (X_1, \ldots X_k)$ un vector aleatorio, si $a, b \subset \{1, \ldots, k\}$ son tales que $a \cap b = \emptyset$ y $a \cup b = \{1, \ldots, k\}$, se dice que $X = (X_a, X_b)$ es una partición del vector X donde $X_a = \{X_i, i \in a\}$ y $X_b = \{X_j, j \in b\}$.

Dado un vector aleatorio particionado $X = (X_a, X_b)$, el vector marginal X_a se define como $X_a = \{X_i, i \in a\}$. De igual forma x_a denota el valor del vector X_a en la tabla marginal y $p(x_a)$ es la probabilidad asociada a ella. Se cumple:

$$p(x_a) = \sum_{x_b} p(x_a, x_b)$$

Un problema muy frecuente en el análisis de datos categóricos es el de estudiar el grado de asociación que existe entre las variables. Es bien conocido el uso de los modelos loglineales para estos fines.

1.2 Modelos loglineales

Definición 1 *La expresión loglineal de una función de masa de probabilidad multinomial f_K es:*

$$log f_K(x) = \sum_{a \subseteq K} u_a(x_a)$$

donde la suma es en todos los subconjuntos a de $K = \{1, \ldots, k\}$ y las funciones u_a son funciones conocidas como términos de interacción.

Definición 2 *Un modelo loglineal es jerárquico si de la restricción a cero de un término de interacción se puede inferir que todos los demás términos que lo contienen son también cero, es decir, si $u_a = 0$ entonces $u_t = 0$, $\forall t : t \supseteq a$.*

Definición 3 *Un término de interacción de la expresión loglineal de una función de densidad u_a se dice maximal si es diferente de cero y $\forall\ b \supset a$, $u_b = 0$.*

Los modelos jerárquicos pueden identificarse a partir de la lista de sus términos de interacción maximales. Esta lista es conocida como la fórmula del modelo y los elementos que la conforman son frecuentemente llamados generadores del modelo. Denotaremos un modelo con la letra M.

Definición 4 *Dado el modelo M y el subconjunto de variables a, el submodelo M_a es el modelo cuyo conjunto de generadores se obtiene de quitar todas las ocurrencias de los factores de a^c en el conjunto de los generadores de M y eliminar los términos redundantes.*

1.3 Estimación y pruebas de hipótesis

Sea $X = (X_1, \ldots X_k)$ un vector aleatorio con distribución multinomial. Tomemos una muestra aleatoria de tamaño N del vector X, y consideremos que estas observaciones son independientes. Se denota por $n(x)$ a la cantidad observada de individuos en la celda x.

Definición 5 *Dado un modelo jerárquico M, se dice que $\hat{p}$ es el estimador máximo verosímil de $p \in M$ si es la solución única del sistema:*

1. $\hat{p} \in M$
2. $\hat{p}(x_c) = \frac{n(x_c)}{N}$, $\forall c$ *generador de* M, *donde* $n(x_c) = \sum_{x_a} n(x_a, x_c)$

Se define la devianza de un modelo M_0 como:

$$G^2 = \sum_X n(x) \log(\frac{n(x)}{\hat{m}_0(x)})$$

donde $\hat{m}_0(x)$ es el estimador máximo verosímil de la frecuencia esperada en la celda x bajo M_0. Es conocido el resultado de que $G^2 \sim \chi^2(l)$, donde l son los grados de libertad que se calculan a través de la fórmula, [19]:

$$l = \#\text{ celdas de la tabla - }\#\text{ parámetros libres estimados}$$

Para la comparación de modelos anidados $M_0 \subseteq M_1$ se define la diferencia de las devianzas entre M_0 y M_1:

$$\Delta G^2 = G_0^2 - G_1^2 = 2\sum_x n(x) log(\frac{n(x)}{\hat{m}_0(x)}) - 2\sum_x n(x) log(\frac{n(x)}{\hat{m}_1(x)}) = 2\sum_x n(x) log(\frac{\hat{m}_1(x)}{\hat{m}_0(x)})$$

donde $\hat{m}_0, \hat{m}_1$ son los estimadores de las frecuencias esperadas bajo los modelos M_0 y M_1 respectivamente.

Bajo M_0, ΔG^2 tiene distribución χ^2 asintótica con grados de libertad dados por la diferencia de parámetros libres entre M_0 y M_1.

1.4 Selección de modelos

Muchas veces, en un problema concreto, se encuentran numerosos modelos que se ajustan a los datos por lo que el investigador se tiene que enfrentar a la tarea de seleccionar uno de ellos. Para esta labor, se cuentan con muchos criterios de lo que puede ser un buen modelo. Se considera que un modelo es adecuado cuando:

- No es rechazado por ninguna prueba de bondad de ajuste.
- Es fácilmente interpretable.
- No contiene términos que puedan ser excluidos del modelo sin causar un deterioro significativo del ajuste.
- No omite términos que mejoren significativamente el ajuste si son adicionados al modelo.
- Explica en gran medida la variabilidad de los datos.

En la práctica es muy difícil encontrar un modelo que cumpla simultáneamente los criterios anteriores. Por esta razón, la elección de un modelo está también determinada por el tipo de problema que se está resolviendo y lo que se pretende explicar de él.

2 Modelación gráfica

En este epígrafe daremos una introducción a los conceptos y propiedades fundamentales en los que se basa la modelación gráfica.

2.1 Independencia e independencia condicional de vectores aleatorios

Definición 6 *Sean X, Y vectores aleatorios, se dice que X y Y son independientes si y sólo si la función de densidad conjunta f_{XY} satisface:*

$$f_{XY}(x, y) = f_X(x) f_Y(y), \forall x, y.$$

donde f_X y f_Y representan las funciones de densidad marginales de X y Y respectivamente. Esta relación se representa con la notación $X \perp Y$.

Lema 1 (de reducción) *Sea (X, Y, Z) un vector aleatorio particionado tal que $X \perp (Y, Z)$, entonces se cumple que $X \perp Y$.*

Con esta propiedad se asegura que si $X \perp (Y, Z)$, entonces $X \perp Y$ y $X \perp Z$. Sin embargo, el recíproco no siempre es cierto.

Definición 7 *Sean X, Y, Z vectores aleatorios. Se dice que Y y Z son condicionalmente independientes dado X si y sólo si:*

$$f_{YZ/X}(y, z; x) = f_{Y/X}(y; x) f_{Z/X}(z; x), \forall x : f_X(x) > 0, \forall y, z.$$

Esta relación se representa con la notación $Y \perp Z|X$.

En el contexto de la independencia condicional, el lema de reducción tiene el enunciado:

Lema 2 *Si (X, Y, Z_1, Z_2) es un vector aleatorio particionado tal que $Y \perp (Z_1, Z_2/X)$ entonces $Y \perp (Z_1/X)$.*

Proposición 1 *Sea $X = (X_a, X_b, X_c)$ un vector aleatorio multinomial particionado, entonces $X_b \perp X_c/X_a$ si y sólo si todos los términos de interacción en la expresión loglineal asociado a una o más variables de b y c son cero.*

En el caso que a sea vacío, la proposición establece la independencia entre X_b y X_c. Para la independencia condicional de pares dado el resto, es decir, $X_i \perp X_j/V \backslash \{X_i, X_j\}$ se necesita que $u_{\{i,j\} \cup t} = 0$ siempre que $t \subseteq V \backslash \{X_i, X_j\}$.

De la proposición anterior se infiere que las propiedades de independencia e independencia condicional de un vector aleatorio multinomial están directamente relacionadas con la presencia o ausencia de términos de interacción en la expresión loglineal, en el caso de modelos jerárquicos lo anterior está determinado por los términos de interacción maximales de la expresión loglineal o, dicho de otra manera, de los generadores de dichos modelos.

2.2 Teoría de grafos y grafos de independencia

En este epígrafe, además de la definición de grafo de independencia, se introducen algunos conceptos de la teoría general de grafos que son necesarios para comprender la modelación gráfica.

Un grafo es un ente matemático que consta de dos conjuntos V y E, donde V es el conjunto de los vértices y E es el conjunto de las aristas entre ellos. Un grafo se denota de la forma $G = (V, E)$. Se dice que G es no direccionado si $(X, Y) \in E$ es equivalente a que $(Y, X) \in E$.

Definición 8 (Grafo de independencia) *Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio. Se dice que el grafo no direccionado $G = (V, E)$ es el grafo de independencia de X si $V = \{X_1, \ldots, X_k\}$ y $(X_i, X_j) \notin E$ si y sólo si $X_i \perp X_j/V \backslash \{X_i, X_j\}$.*

Como se puede observar, un grafo de independencia representa las relaciones de independencia e independencia condicional existentes entre las variables aleatorias presentes en el vector aleatorio.

Definición 9 *Sea el grafo $G = (V, E)$.*

1. *Se dice que $X, Y \in V$ son adyacentes si $(X, Y) \in E$. Esta relación se denota $X \sim Y$.*
2. *Una secuencia de vértices $X_0, \ldots, X_n$ es un camino de longitud n si y sólo si $X_i \sim X_{i+1}$ con $i = 0, \ldots, n-1$.*
3. *Un subconjunto de vértices separa a dos vértices X, Y si y sólo si cualquier camino que une estos vértices contiene al menos un vértice de dicho subconjunto.*
4. *Un subconjunto de vértices separa dos subconjuntos a y b si separa cada vértice $X \in a$ de $Y \in b$.*
5. *Dos vértices están conectados si existe un camino entre ellos. Un grafo se dice conexo si todo par de vértices está conectado.*
6. *Dado un subconjunto de vértices a, el grafo inducido por a, $G_a = (V', E')$ se obtiene de considerar a $V' = a$ y $E' = \{(X, Y) \in E : X, Y \in a\}$.*
7. *Sea $a \subset V$, se define como frontera de a al conjunto siguiente: $bd(a) = \{X_i \in V \backslash a : \exists X_j \in a \text{ tal que } X_i \sim X_j\}$*
8. *Un ciclo $X_0, \ldots, X_n$ se dice sin cuerdas si los únicos pares de vértices adyacentes son los sucesivos.*
9. *G se dice triangulado si no tiene ciclos sin cuerdas de longitud mayor o igual que cuatro.*

Lema 3 (de Separación) *Sea $X = (X_1, \ldots, X_k)$ un vector aleatorio con grafo de independencia $G = (V, E)$. Sean $a, b, c \subset V$ disjuntos tales que b y c están separados por a, entonces $b \perp c/a$.*

El inverso de este teorema también es cierto y se cumple en el sentido siguiente: si para $a, b, c \subset V$ se tiene que $a \perp b/c$ bajo cualquier función de densidad en el modelo, entonces c separa a a de b en el grafo de independencia. Con esto, las relaciones de independencia e independencia condicional pueden ser leídas directamente a través del grafo de independencia.

Muy vinculadas con el teorema de separación están las llamadas propiedades de Markov:

1. Propiedad de pares: Todo par de vértices no adyacentes X_i , X_j cumple

$$X_i \perp X_j / V \backslash \{X_i, X_j\}.$$

2. Propiedad Global: Para todos los subconjuntos disjuntos de vértices a, b, c; siempre que b y c estén separados por a se cumple que $b \perp c/a$.

3. Propiedad Local: Para todo vértice X_i, si $a = bd(X_i)$ y b es el resto de las variables, es decir, $b = V \backslash \{\{X_i\} \cup a\}$ entonces $X_i \perp b/a$.

Teorema 1 *Las tres propiedades de Markov son equivalentes.*

2.3 Modelos gráficos discretos

En este epígrafe se verán algunas de las características de los mismos y cómo se relacionan con la teoría de los grafos de independencia.

Definición 10 *Dado un vector aleatorio multinomial $X = (X_1, \ldots, X_k)$ con grafo de independencia $G = (V, E)$, se dice que el modelo asociado a G es gráfico si la expresión loglineal de la función de densidad de X:*

$$log f_K(x) = \sum_{a \subseteq K} u_a(x_a)$$

está sujeta a las restricciones:

- $\forall i, j \in K : (X_i, X_j) \notin E$ *, se cumple que* $u_a = 0, \forall a \subseteq K$ *tal que* $i, j \subseteq a$.
- $\forall i, j, m, n \in K : (X_i, X_j) \notin E$ *y* $(X_m, X_n) \in E$*, se cumple que* $u_a \neq 0 \forall a \subseteq K$ *tal que* $\{i, j\} \not\subset a$ *y* $\{m, n\} \subseteq a$.

Diferentes modelos pueden tener el mismo grafo de independencia, el modelo gráfico asociado a un grafo es el modelo jerárquico maximal que admite el grafo de independencia.

Definición 11 *Un grafo se dice completo si todos sus vértices son adyacentes y un subconjunto de vértices es completo maximal si es completo y cualquier otro subconjunto de vértices que lo contiene no lo es.*

Proposición 2 *Un modelo jerárquico es gráfico si y sólo si los generadores del mismo coinciden con los subconjuntos de vértices completos maximales de su grafo de independencia.*

La caracterización anterior facilita la construcción del grafo de independencia de un modelo gráfico a partir de su fórmula. La construcción se basa en conectar todos los pares de vértices que aparecen en un mismo generador.

Como hay una correspondencia biunívoca entre modelos gráficos y grafos de independencia, se puede analizar la estructura de independencia condicional de un modelo gráfico utilizando solamente su grafo de independencia asociado, e interpretarlo a través de las propiedades de Markov. La ventaja sobre los modelos loglineales jerárquicos es que para interpretar estos últimos hace falta, además, considerar los términos de interacción que faltan en el modelo para que sea gráfico.

2.4 Colapsabilidad

La propiedad de colapsabilidad es muy útil pues mediante su uso es posible descomponer un problema complejo en subproblemas más sencillos, y analizar sin pérdida de información las tablas marginales.

La noción de colapsabilidad es mayormente utilizada, en el contexto de tablas de contingencia, cuando se estudia si las medidas de asociación marginal se mantienen inalterables en las tablas parciales [3], [19]. Este concepto no está directamente vinculado con un modelo determinado y es conocida como colapsabilidad paramétrica.

Aquí utilizaremos otro concepto de colapsabilidad, donde sí se tiene en cuenta el modelo; por lo que es conocida como colapsabilidad del modelo [4],[11].

Definición 12 *Un modelo M se dice colapsable sobre el subconjunto de variables X_a con $a \subset \{1, \ldots, k\}$ si $\forall x_a$, observación del vector aleatorio X_a, se cumple:*

$$\hat{p}\left(x_a\right) = \hat{p}_a\left(x_a\right)$$

donde $\hat{p}_a\left(x_a\right)$ es el estimador máximo verosímil de la distribución marginal $p\left(x_a\right)$ basada en los datos marginales.

En este epígrafe se expone la caracterización de un modelo gráfico por su grafo de independencia para, mediante éste, identificar cuando el modelo es colapsable sobre un subconjunto de variables.

Proposición 3 *Un modelo gráfico es colapsable sobre a si para cualquier componente conexa de a^c, b se tiene que $bd(b)$ es completo.*

A través de esta proposición es posible leer, del grafo de independencia de un modelo gráfico, cuándo y cómo se puede colapsar el modelo.

Existen algunos modelos a los que se les puede aplicar de manera sucesiva la propiedad de colapsabilidad hasta lograr expresiones explícitas para las estimaciones máximo verosímiles de los parámetros. Son los llamados modelos que se pueden descomponer.

En el siguiente epígrafe se expone un resultado que relaciona esta característica con la estructura triangulada del grafo de independencia.

2.5 Modelos que se pueden descomponer

Dentro de la clase de los modelos gráficos está la clase de los modelos que se pueden descomponer. Ellos cuentan con propiedades que los hacen muy importantes en la modelación gráfica, entre ellas:

1. Los estimadores máximo-verosímiles se pueden calcular de forma directa.
2. Son fáciles de interpretar.
3. Pueden sugerir el mecanismo estocástico por el cual fueron generados los datos.

Una definición clásica de modelo que se puede descomponer es que el grafo asociado sea completo o reducible a dos componentes que se puedan descomponer.

El siguiente Teorema es unos de los resultados más importantes que se han obtenido a través de la teoría de la modelación gráfica, pues da una caracterización de modelo que se puede descomponer teniendo en cuenta solamente el tipo de grafo que representa a dicho modelo.

Teorema 2 *Un vector aleatorio X se puede descomponer si y sólo si su grafo de independencia G es triangulado.*

Proposición 4 *Para todo par de modelos anidados que se pueden descomponer existe una secuencia de modelos de este tipo, basada en la exclusión de aristas del grafo de independencia, que va del modelo más complejo al más simple y viceversa (existe una secuencia de modelos que se pueden descomponer basada en la inclusión de aristas del grafo de independencia que va del modelo más simple al más complejo).*

La clase de los modelos loglineales de interacción contiene a la clase de modelos loglineales jerárquicos, ésta contiene a la clase de los modelos gráficos discretos, la que a su vez, contiene a la clase de los modelos que se pueden descomponer. En la tabla 1 se muestran las cantidades de modelos de cada tipo para tablas de contingencia de dimensiones uno, dos, tres, cuatro y cinco [11].

TIPOS DE MODELOS	DIMENSIONES				
	1	2	3	4	5
Interacción	2	8	128	32 768	2 147 483 648
Jerárquicos	2	5	19	167	7 580
Gráficos	2	5	18	113	1 450
Se pueden descomponer	2	5	18	110	1 233

Tabla 1: Ver Tabla 3, página 50, Lauritzen (1989).

Esta diferencia en cantidades se hace mayor a medida que aumenta la cantidad de variables del problema. En resumen, para fines de selección de modelos resulta conveniente, en una etapa inicial, trabajar con las clases más pequeñas.

En los problemas que se presentan habitualmente en la práctica de la estadística puede parecer que no es tan importante esta diferencia; por lo que la principal utilidad de la modelación gráfica en esta área sigue siendo la interpretabilidad de los modelos [20],[7].

En el contexto de los sistemas expertos, se aprovecha la estructura modular de los modelos gráficos para salvar el inconveniente de la gran complejidad y volumen de cálculo que involucra este tipo de problemas [14], [10] . En el área de computación evolutiva, la utilidad mayor de los modelos gráficos y en particular, de los que se pueden descomponer, es de tipo algorítmica [15],[2].

3 Selección de modelos gráficos

En este epígrafe se discuten algunos métodos de selección de modelos gráficos discretos. Los métodos que se estudian son:

- Selección hacia atrás.
- Selección hacia adelante.
- Selección en dos pasos.
- Selección automática EH.

El método de selección en dos pasos aparece descrito en [20]. Los métodos de selección automática (hacia delante, hacia atrás y EH) están implementados en el paquete de programas MIM [7].

La búsqueda se puede restringir al subconjunto de los modelos que se pueden descomponer. Sin embargo, se pueden encontrar también modelos interesantes aunque no se puedan descomponer, o incluso, que no sean gráficos; de ahí que la utilidad mayor de estos métodos de selección sea en una etapa exploratoria inicial.

3.1 Método de selección hacia atrás

Se parte de un modelo inicial, generalmente complicado y consistente con los datos, por lo que frecuentemente se toma el saturado. En cada paso se van eliminando sucesivamente las aristas menos importantes. Para ello se utiliza la prueba χ^2 basada en la diferencia entre devianzas de dos modelos sucesivos: el modelo que contiene la arista contra el que no la contiene. El procedimiento termina cuando todas las aristas que están presentes en el modelo son importantes.

Las razones para la no eliminación de una arista son:

- Aristas fijas en el modelo, esto es: las variables de ambos extremos de la arista son consideradas explicatorias.
- Respeto al principio de coherencia, es decir: si en un determinado paso se rechaza la exclusión de una arista, ya no se vuelve a considerar su eliminación en los pasos subsiguientes.
- Restricción al subconjunto de los modelos que se pueden descomponer: en ningún paso se consideran aquellas aristas cuya exclusión dé como resultado un modelo que no se puede descomponer.

De forma resumida, el algoritmo consiste en lo siguiente:

1. A partir de un modelo inicial M, se considera su grafo de independencia y a cada una de las aristas presentes en él se le aplica la prueba de hipótesis:

   H_0: M sin la arista.
   H_A: M

2. Se elimina la arista menos importante (la de mayor valor de p, $p > \alpha$).

3. Se ajusta el modelo resultante y se toma como modelo inicial.

4. Se repiten los pasos anteriores hasta que todas las aristas presentes sean importantes ($p \leq \alpha$).

Del algoritmo anterior se desprende que la búsqueda se realiza en los modelos consistentes con los datos, lo cual garantiza que el escogido tenga un buen ajuste. En general, el procedimiento simplifica los modelos, pero como casi siempre se parte de un modelo complicado, esta simplificación es relativa.

3.2 Método de selección hacia adelante

En este método el modelo inicial es sencillo y por lo general inconsistente con los datos. La idea básica es ir añadiendo las aristas que resulten importantes. Es usual comenzar con el modelo de independencia entre las variables. Al igual que en el método anterior, en éste también se usa la prueba χ^2 basada en la diferencia entre devianzas de dos modelos sucesivos. El procedimiento termina cuando ninguna de las aristas por incorporar mejora significativamente el ajuste.

Las razones para la no inclusión de una arista son:

- Principio de coherencia. Funciona como en el caso anterior, pero con el rechazo de la inclusión.

- Restricción a modelos que se pueden descomponer. Igual que en el método anterior, pero con la inclusión.

De modo más resumido, los pasos del algoritmo son:

1. En el modelo inicial M, se prueba, para cada arista susceptible de ser elegida, las hipótesis:

   H_0: M
   H_A : M con la arista

2. Se adiciona al modelo la arista más importante (la de menor valor de $p \leq \alpha$).

3. Se ajusta el modelo resultante y se considera como modelo inicial.

4. Se repite el procedimiento hasta que ninguna arista a incluir sea significativa ($p > \alpha$).

Es fácil ver que este método realiza la búsqueda en el conjunto de modelos no consistentes con los datos. El criterio de parada no considera el buen ajuste global, sino solamente que la inclusión de una arista mejore el ajuste; esto hace que muchas veces el modelo seleccionado no se ajuste.

3.3 Método de selección en dos pasos

Este método es una combinación de los dos anteriores. En él se eliminan del modelo saturado todas las aristas que no son importantes y luego, partiendo del modelo simplificado resultante, se le añaden todas las aristas importantes.

La selección en dos pasos cuenta con las siguientes etapas:

1. A partir del modelo saturado se realizan todas las pruebas de exclusión de una arista.

   H_0: Modelo saturado sin una arista.

   H_A: Modelo saturado.

2. Se excluyen del modelo saturado todas las aristas que no resulten importantes según esta prueba ($p > \alpha$). Sea M el modelo resultante.

3. Con el modelo resultante se realizan todas las pruebas de inclusión de las aristas susceptibles a incluir:

   H_0: M

   H_A: M con la arista

   Se adicionan al modelo M todas las aristas que resulten importantes a través de la prueba de hipótesis ($p < \alpha$).

En resumen, la desventaja de los tres métodos anteriores radica en que pueden perderse modelos interesantes que ofrezcan un buen ajuste. Al mismo tiempo, para un mismo problema, cada uno de ellos puede dar como resultado modelos diferentes. En el caso de que todos coincidan, puede considerarse que el modelo resultante es un buen candidato a ser elegido.

3.4 Método de selección automática EH

Esta es una forma diferente de selección de modelos. Está basada en un algoritmo de búsqueda en el que se escogen los modelos más simples consistentes con los datos. Durante la búsqueda se ajusta una sucesión de modelos y se clasifican como aceptados o rechazados.

En este algoritmo se respeta el criterio de coherencia, que establece lo siguiente para dos modelos anidados $M_0 \subseteq M_1$:

- Si M_1 es rechazado, o sea, es inconsistente con los datos, entonces M_0 también es rechazado.
- Si M_0 es aceptado, o sea, es consistente con los datos, entonces M_1 también es aceptado.

Este criterio sirve para agilizar el proceso. La idea consiste en que si se acepta un modelo M, entonces se aceptan todos los demás modelos que contienen a M, y por lo tanto no es necesario considerarlos (de forma similar ocurre en el caso contrario).

En un paso cualquiera de este algoritmo, se ajusta un modelo que se adiciona a la lista de modelos aceptados o rechazados según sea el caso. La clase de los modelos con la que se está trabajando queda de esta forma dividida en tres conjuntos disjuntos. El primero contiene a todos aquellos que tienen como submodelo a un modelo aceptado, por lo que son considerados consistentes con los datos. Este conjunto es llamado conjunto de modelos débilmente aceptados. El otro conjunto contiene a todos los que son submodelos de uno rechazado; es decir, son inconsistentes con los datos. Este es el conjunto de los modelos débilmente rechazados. El último de los conjuntos contiene a los modelos que todavía no han sido clasificados. Este es el conjunto de los modelos indeterminados.

El proceso termina cuando el conjunto de los modelos indeterminados queda vacío. Al final del proceso de búsqueda, se obtiene un listado de modelos minimales aceptados, donde todos son factibles. Con esto, el investigador tiene más opciones para escoger el modelo apropiado según el problema que esté resolviendo.

Aunque mediante el criterio de coherencia se agiliza la búsqueda, el método en general es algo lento debido a que considera en la búsqueda una gran cantidad de modelos, muchos de los cuales no se pueden descomponer.

3.5 Comparación entre los métodos

Como se puede apreciar, todos los métodos de selección de modelos expuestos anteriormente tienen como objetivo común encontrar modelos sencillos consistentes con los datos. Sin embargo, su diferencia estriba en el hecho que no utilizan los mismos algoritmos de búsqueda ni el mismo criterio para determinar la consistencia con los datos. Debido a esto, no necesariamente se selecciona el mismo modelo aplicando los diferentes métodos.

Una característica que hace poco apreciado el método de selección hacia adelante es que, en determinadas ocasiones, se escoge un modelo que realmente no se ajusta a los datos.

El método de selección hacia atrás no tiene esta desventaja, pues, al trabajar sobre modelos consistentes con los datos, siempre se garantiza la selección de un modelo que se ajusta; aunque puede resultar excesivamente complejo.

En la selección automática EH se utilizan pruebas de bondad de ajuste global más que pruebas entre modelos sucesivos y se pueden seleccionar varios modelos.

Cuando se seleccionan varios modelos, las interpretaciones de los mismos pueden diferir. Entonces se hace difícil la selección de uno para explicar el problema, a menos que se utilice conocimiento sustantivo. Cuando no tiene sentido hablar de un único modelo, se prefiere mencionar el conjunto de modelos posibles.

De modo general, no se puede decir que un método sea mejor que otro. En todo caso hay que estar conscientes de la envergadura del problema que se investiga para tener noción del tipo de modelo que se necesita para la explicación del problema.

3.6 Estrategia para la selección de modelos

Cuando el costo computacional no es relevante, debe tenerse en cuenta los siguientes principios:

1. No debe utilizarse un único método de selección automática.

2. Cuando los diferentes métodos coinciden, el modelo seleccionado es un buen candidato.

3. Cuando se ajustan modelos diferentes, se debe mencionar el conjunto de modelos posibles e interpretarlos haciendo énfasis en sus similitudes y diferencias.

4. Se pueden hacer combinaciones de métodos; por ejemplo: considerar como modelo mínimo en la selección automática EH al modelo que resulta en el paso hacia atrás del método de selección en dos pasos.

Estos principios, de hecho, constituyen una estrategia de trabajo para cada tipo de problema particular. Es conveniente, comenzar la búsqueda en la clase de modelos que se pueden descomponer; si no es suficiente, pasar a la clase de los modelos gráficos discretos, luego a la de modelos jerárquicos loglineales y por último a la de modelos loglineales de interacción.

En otro tipo de problemas, donde se tienen muchas variables y no se busca un modelo con fines explicativos sino, por ejemplo, para generar datos como parte de algoritmos más complejos, se recomienda trabajar únicamente con la clase de modelos que se pueden descomponer.

4 Ejemplo: factores de riesgo de enfermedades coronarias

En este epígrafe, (para comparar los diferentes métodos de selección de modelos gráficos) se utiliza un ejemplo de la literatura, [20], [7]. Se empleó el paquete de programas MIM [7]. En todos los casos se utilizó un nivel de significación $\alpha = 0.05$.

Este ejemplo consiste en una tabla de contingencia en la cual se cruzan seis variables categóricas que se consideran factores de riesgo para enfermedades del corazón. Los datos provienen de un estudio prospectivo hecho en Checoslovaquia a una muestra de 1841 empleados de una fábrica de automóviles. En la tabla 2 se muestran los valores observados.

Las variables que se midieron son las siguientes:

A: fumador (sí, no)
B: trabajo mental extenuante (sí, no)
C: trabajo físico extenuante (sí, no)
D: presión sanguínea (< 140, > 140)
E: razón de lipoproteínas (< 3, > 3)
F: antecedentes familiares (sí, no)

En la tabla 3 se muestran los resultados obtenidos con cada uno de los métodos considerados; a saber: selección hacia atrás, selección hacia adelante, método automático EH y selección en dos pasos.

				B	**No**		**Si**	
F	**E**	**D**	**C**	**A**	**No**	**Si**	**No**	**Si**
No	< 3	< 140	No		44	40	112	67
			Sí		129	149	12	23
		> 140	No		35	12	80	33
			Sí		109	67	7	9
	> 3	< 140	No		23	32	70	66
			Sí		50	80	7	13
		> 140	No		24	25	73	57
			Sí		51	63	7	16
Si	< 3	< 140	No		5	7	21	9
			Sí		9	17	1	4
		> 140	No		4	3	11	8
			Sí		14	17	5	2
	> 3	< 140	No		7	3	14	14
			Sí		9	16	2	3
		> 140	No		4	0	13	11
			Si		5	14	4	4

Tabla 2: Ver Tabla 8.5.1, página 261, Whittaker (1990).

Métodos de Selección	Modelo seleccionado	G^2	Grados de libertad	p
Hacia Atrás	ADE, ACE, ABC, BF	51.3587	46	0.2718
Hacia Adelante	AC, BC, BE, DE, BF	95.0341	52	0.0003
Automático	ADE, ACE, BC, F	62.0779	49	0.0994
EH	ADE, AC, BC, BE, F	63.0128	50	0.1023
En dos pasos	ACE, ADE, BC, BF	57.3463	48	0.1672

Tabla 3: Modelos seleccionados para los datos de la tabla 2.

Como se puede observar, los métodos no dan como resultado el mismo modelo. Por esta razón se hace necesario un análisis para escoger el, o los modelos más adecuados.

En primer lugar, el modelo que se obtiene con el método de selección hacia adelante no se ajusta a los datos. Esto implica que no tiene sentido utilizarlo para analizar la asociación existente entre las variables.

El segundo modelo, que se ajusta mediante el método de selección automática EH (ADE, AC, BC, BE, F), no se puede descomponer. Esto se ve fácilmente a partir del criterio que asegura que un modelo se puede descomponer si y sólo si su grafo es triangulado. En la figura 1 aparece el grafo de independencia asociado a este modelo. Se observa la existencia de un ciclo sin cuerdas de longitud cuatro: A, C, B, E, A. Por razones de facilidad de interpretación se decidió descartarlo también.

Figura 1: Grafo de independencia del modelo ADE, AC, BC, BE, F.

Modelos	Modelos Colapsados	G^2	Grados de Libertad	P
ADE, ACE, ABC, BF	ADE, ACE, ABC	14.1672	16	0.5863
ADE, ACE, BC, F	ADE, ACE, BC	20.1549	18	0.3242
ADE, ACE, BC, BF	ADE, ACE, BC	20.1549	18	0.3242

Tabla 4: Modelos resultantes de eliminar la variable F de los modelos de la tabla 3.

Los restantes modelos que quedan por analizar, se pueden descomponer y ofrecen un buen ajuste. En las figuras 2, 3 y 4 se muestran los grafos de independencia correspondientes.

Obsérvese que la estructura de independencia condicional de cada uno de ellos es muy similar. En primer lugar, en los tres casos aparecen los generadores ADE y ACE. Por otro lado, los tres modelos son colapsables sobre el conjunto formado por las variables $\{A, B, C, D, E\}$. Esto se debe a que F es independiente del resto de las variables en un caso (figura 3) y, en los otros dos, F es condicionalmente independiente del resto de las variables dado B; por lo que en todos lo casos se cumple que $bd\ (F)$ es completo. La consecuencia práctica es que se puede reducir el análisis a la tabla marginal resultante. Esto se puede hacer de dos maneras: marginalizando los modelos ya ajustados o eliminando la variable F y repitiendo el proceso de selección de modelos. El primer procedimiento está completamente justificado por la teoría. El segundo resulta de la experiencia de que no hay un único modelo factible, aunado a las facilidades que dan los paquetes de programas y al carácter exploratorio inicial de este tipo de análisis.

En la tabla 4 aparecen los modelos que se obtienen de eliminar la variable F.

Los dos últimos modelos coinciden y además son casos particulares del primer modelo. Como no hay mucha diferencia en el ajuste de ambos modelos, seleccionamos el más sencillo: ADE, ACE, BC.

En la tabla 5 se muestran los resultados obtenidos con la tabla marginal resultante de eliminar la variable F

El análisis en este caso es similar al de la tabla 3. Aquí se obtiene el mismo modelo (ADE, ACE ,BC.) En la figura 5 se muestra el grafo asociado.

Utilizando las propiedades de Markov, se puede deducir la estructura de independencia condicional del modelo seleccionado:

Figura 2: Grafo de independencia del modelo ADE, ACE, ABC, BF.

Figura 3: Grafo de independencia del modelo ADE, ACE, BC, F.

Figura 4: Grafo de independencia del modelo ADE, ACE, BC, BF.

Métodos de Selección	Modelo Seleccionado	G^2	Grados de Libertad	p
Hacia atrás	ADE, ACE, BCE	13.72	16	0.6192
Hacia adelante	DE, BE, BC, AC	57.84	22	0.0000
Automático	ACE, ADE, BC	20.15	18	0.3242
EH	ADE, AC, BC, BE	21.09	19	0.3319
En dos pasos	ACE, ADE, BC	20.15	18	0.3242

Tabla 5: Modelos seleccionados con cada método para los datos de la tabla Marginal (quitando la variable F).

Figura 5: Grafo de independencia del modelo ADE, ACE, BC.

- $B \perp / \{A, D, E\} / C$
- $C \perp D / \{E, A\}$

En resumen, el trabajo mental (B) es condicionalmente independiente del hábito de fumar (A), la presión (D) y la razón de liproproteínas (E) dado el trabajo físico (C). Además, el trabajo físico es condicionalmente independiente de la presión sanguínea dados el hábito de fumar y la razón de lipoproteínas. Si se quisiera predecir la presión sanguínea a partir de los demás factores de riesgo, sólo serían necesarios el hábito de fumar y la razón de lipoproteínas.

Este es un caso en el que los diferentes métodos de selección no coinciden, como sucede frecuentemente en la práctica. Por esta razón, cuando el costo computacional no es lo más importante, se recomienda probar con los diferentes métodos disponibles para que el análisis de los resultados sea más completo.

5 Conclusiones

La importancia fundamental de los modelos gráficos discretos está dada por las posibilidades de interpretación que brindan, pues como existe una relación biunívoca entre ellos y el grafo de independencia asociado, se facilita la interpretación de los mismos utilizando las propiedades de Markov e identificando las propiedades de colapsabilidad y de poderse descomponer.

No puede decirse de manera absoluta que ninguno de los métodos de selección de modelos gráficos estudiados sea el mejor para el tratamiento de tablas de contingencia no estructuradas. En este trabajo se formuló una estrategia general que debe ser adaptada a cada caso concreto.

La estrategia propuesta también puede servir como una etapa inicial en la búsqueda de un modelo más complejo, ya sea en la clase de los modelos jerárquicos loglineales, o la de los modelos jerárquicos de interacción.

En otro tipo de problemas, donde se tienen muchas variables y no se busca un modelo con fines explicativos sino, por ejemplo, para generar datos como parte de algoritmos más complejos, se recomienda trabajar únicamente con la clase de modelos que se pueden descomponer.

Agradecimientos:

Los autores agradecen al Dr. Jesús E. Sánchez García, por revisar y corregir varias versiones preliminares de este trabajo, y a la M. en C. Elva Díaz Díaz, no sólo por haber realizado la oponencia, sino por sus acertadas sugerencias, que permitieron que quedara mejor.

Referencias

[1] Alfonso, K. (1999) *Análisis de Tablas de Contingencia Vía Modelos Gráficos*. Tesis de Licenciatura en Matemática, Universidad de La Habana.

[2] Acid, S.; Campos, L.M. (1999) "Fast algorithms for learning simplified graphical models", in: *Proceedings of the Second Symposium on Artificial Intelligence*. Editorial Academia, La Habana.

[3] Agresti, A. (1990) *Categorical Data Analysis*. Wiley, New York.

[4] Asmussen, S.; Edwards, D. (1983) "Collapsability and response variables in contingency tables", *Biometrika* **70**(3): 566–578.

[5] Darroch, J. N.; Lauritzen, S. L.; Speed, T. P. (1980) "Markov fields and log-linear models for contingency tables", *Annals of Statistics* **8**: 522–539.

[6] Edwards, D. (1990) "Hierarchical interaction models (with discussion)", *J.Royal Stat. Soc.* B **52**: 3–20, 51–72.

[7] Edwards, D. (1995a) *Introduction to Graphical Modelling*. Springer Texts in Statistics, New York.

[8] Edwards, D. (1995b) "Graphical Modelling", in: J. Krazanowski (Ed.) *Recent Advances in Descriptive Multivariate Analysis*. Clarendon Press, Oxford.

[9] Edwards, D.; Kreiner, S. (1983) "The analysis of contingency tables by graphical models", *Biometrika* **70**: 553–565.

[10] Larrañaga, P.; Etxeberría, R.; Lozano, J.A.; Sierra, B.; Inza, I.; Peña, J. (1999) "A review of the cooperation between evolutionary computation and probabilistic graphical models", in: *Proceedings of the Second Symposium on Artificial Intelligence*, Editorial Academia, La Habana.

[11] Lauritzen, S.L. (1989) "Lectures on contingency tables", (3rd edn). Technical Report R-89-29, Institute for Electronic Systems, Aalborg University.

[12] Lauritzen, S.L.; Spiegelhalter, D.J. (1988) "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)", *J. Royal Stat. Soc.* B **50**: 157–224.

[13] Lauritzen, S. L.; Wermuth, N. (1989) "Graphical model for associations between variables, some of which are qualitative and some quantitative", *Annals of Statistics* **17**: 31–57.

[14] Lauritzen, S.L. (1996) *Graphical Models*. Oxford Science Publications, New York.

[15] Mühlenbein, H.; Mahning, T. (1999) "The Factorized distribution algorithm for additively decomposed functions", in: *Proceedings of the Second Symposium on Artificial Intelligence*, Editorial Academia, La Habana.

[16] Mühlenbein, H.; Mahning, T.; Ochoa, A. (1999) "Schemata, distribution and graphical models in evolutionary optimization", *Journal of Heuristics* **15**: 215–244.

[17] Wermuth, N.; Lauritzen, S.L. (1983) "Graphical and recursive models for contingency tables", *Biometrika* **70**: 537–552.

[18] Wermuth, N.; Lauritzen, S.L. (1990) "On substantive research hypotheses, conditional independence graphs and chains models (with discussion)", *J.Roy. Stat. Soc.* B **52**: 21–72.

[19] Wickens, T.D. (1989) *Multiway Contingency Tables Analysis for the Social Sciences*. LEA, New Jersey.

[20] Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.