Centro de Investigación en Matemáticas, A.C.

CIMAT

Modelos Gráficos Clásicos y Generalizados para el Análisis de Datos Binarios y su Aplicación en Datos de Accesos de Internet

Tesis

Que para obtener el Grado de
Maestría en Ciencias
con especialidad en
Computación y Matemáticas Industriales

Presenta

Jesús Emeterio Navarro Barrientos

Director de Tesis:
Dr. Johan Jozef Lode van Horebeek

Guanajuato, Gto. México
12 de Diciembre de 2001

Modelos Gráficos Clásicos y Generalizados para el Análisis de Datos Binarios y su Aplicación en Datos de Accesos de Internet

CIMAT

Tesis que para obtener el Grado de Maestro en Ciencias con especialidad en Computación y Matemáticas Industriales presenta:

Jesús Emeterio Navarro Barrientos

Director de Tesis: Dr. Johan Jozef Lode Van Horebeek

Centro de Investigación en Matemáticas, A. C.
Guanajuato, Gto., México

12 de diciembre de 2001

Resumen

En este trabajo de Tesis se presenta la labor realizada en el área de Modelos Gráficos para el análisis de Datos Binarios. Se muestra un resumen de técnicas utilizadas comunmente para el análisis de datos discretos, entre ellas los modelos gráficos clásicos. Se muestran aspectos teóricos de modelos gráficos clásicos y generalizados. Se propone la utilización de un algoritmo para encontrar independencias implicadas en modelos gráficos generalizados y dos métodos para encontrar los estimadores de máxima verosimilitud en modelos gráficos generalizados. Se propone también un nueva representación de los modelos gráficos generalizados y su utilización en el análisis de accesos realizados a hojas Web en un servidor. Los modelos gráficos que se muestran se obtuvieron utilizando una herramienta denominada $\mathcal{MOGG}$, la cual ha sido desarrollada también dentro de este trabajo de tesis.

Índice General

0.1 Prefacio

Esta tesis es el resultado de los estudios de posgrado realizados en el Centro de Investigación en Matemáticas, A.C., Guanajuato, México, correspondientes al programa de Maestría en Ciencias con Especialidad en Ciencias de la Computación y Matemáticas Industriales.

Aportaciones de la tesis:

1. Presentamos un panorama general de modelos para trabajar con datos discretos como aparecen en la literatura.

2. Desarrollamos un ambiente de software para analizar y procesar datos discretos tanto con modelos gráficos como para modelos gráficos generalizados. Se tiene la siguiente funcionalidad: obtener modelos gráficos a partir de tablas de contingencia o datos muestrales, selección por pasos (Stepwise selection) y selección por pasos en reversa (Backwise selection) de modelos gráficos clásicos, selección por pasos de modelos gráficos generalizados, Etiquetado (label) de aristas y segmentos en modelos, predicción de variables en modelos generalizados, etc.

3. Presentamos un esquema numérico para estimar modelos gráficos generalizados.

4. Como caso de estudio, realizamos el análisis de accesos a las páginas WWW del CIMAT.

El desarrollo del presente trabajo de tesis se resume en los siguientes capítulos.

Capítulo 1.- Se da una introducción al tema que trata este trabajo de tesis, asi como el estado del arte de los métodos utilizados para representar datos discretos y sus relaciones.

Capítulo 2.- Se explica en detalle en que consition los modelos gráficos clásicos, su representación, la extensión a modelos gráficos generalizados y en 2.4.1 se propone un algoritmo para encontrar independencias implicadas para modelos gráficos generalizados.

Capítulo 3.- Se describen los métodos y algoritmos utilizados para construir modelos gráficos clásicos y generalizados, además en 3.3.1 y 3.3.2 se proponen dos procesos para realizar la estimación en modelos gráficos generalizados.

Capítulo 4.- Se muestra como caso de estudio el análisis de un conjunto de visitas a algunos grupos de páginas del servidor de Internet del Centro de Investigación en Matemáticas. Se muestran varios modelos obtenidos y un breve análisis de los resultados. Se presentan también otros resultados obtenidos de algunos ejemplos encontrados en la literatura.

Capítulo 5.- Se describen las conclusiones del trabajo de tesis y el trabajo futuro propuesto.

Apéndice.- Se muestra el manual de usuario del programa $\mathcal{MOGG}$..

Capítulo 1

Introducción

1.1 Introducción

En los ultimos años el problema de encontrar una representación gráfica de un conjunto de datos que facilite el entendimiento de las relaciones y patrones que hay en ellos, ha sido la tarea que muchas personas han querido resolver. Aunque el conjunto de datos puede ser muy variado y puede ir desde datos cuantitativos o continuos, hasta datos categóricos o cualitativos, nos enfocaremos en los datos categóricos. Para este tipo de datos existen muchos menos métodos desarrollados a diferencia de los métodos para datos continuos o cuantitativos.

Dado que los conjuntos de datos por sí solos no pueden la mayoría de las ocasiones dar una idea de las relaciones, dependencias, o patrones, se necesitan de métodos o algoritmos que de alguna manera descubran esa información que se encuentra en los datos pero que a simple vista no puede ser detectada.

Podemos agrupar las preguntas de interés que existen comunmente en la práctica de la siguiente manera:

1. Clasificación.- ¿Este cliente podrá cumplir con el pago de un credito?
2. Segmentación y Agrupamiento (Clustering).- ¿Qué grupos de clientes tengo?
3. Descripción de conceptos.- ¿Qué propiedades caracterizan a los vehículos con el frente débil?
4. Predicción y análisis de cadenas.- ¿Cuál será la relación de cambio del dolar mañana?
5. Análisis de desviación.- ¿Hay variaciones en la región o en la temporada por cambios en el clima?
6. Análisis de dependencias o asociaciones.- ¿Qué productos frecuentemente se compran conjuntamente?

El presente trabajo de tesis trata de resolver el problema del análisis de dependencias o asociaciones por medio de un modelo gráfico poniendo especial énfasis en datos discretos del tipo binario, siendo este tipo de datos los más sencillos y que se encuentran comunmente en la práctica.

Tomemos el siguiente ejemplo.

Supongamos que se tienen los datos referentes a un grupo de estudiantes de primaria de una determinada escuela y se toman los datos referentes a cuatro de las materias que cursaron durante el año escolar. Se sabe si pasaron o no una materia, entonces se tienen las cuatro variables siguientes.- X_0 =Pasaron la materia de Ciencias Sociales (si,no), X_1 =Pasaron la materia de Español (si,no), X_2 =Pasaron la materia de Ciencias Naturales (si,no) y X_3 =Pasaron la materia de Matematicas (si,no). Dada esta información, los modelos gráficos pueden contestar las siguientes preguntas.-

¿Hay alguna dependencia entre pasar la materia de Español con la materia de Matematicas?

¿Hay alguna relación entre Ciencias Sociales y Ciencias Naturales, dado que no se pasó Matemáticas ni Español?

¿Pasar la materia de Matematicas es independiente de pasar las demás?

Un conjunto de datos categóricos tiene un conjunto de variables o valores para cada categoría y es generalmente representado por medio de una tabla que muestra las observaciones obtenidas para cada combinación de valores de las variables, es decir, cuando cada elemento de un conjunto de muestras es clasificado simultáneamente de acuerdo a dos o más variables, a esta tabla se le denomina comunmente "tabla de contingencia".

Un ejemplo de una tabla de contingencia para el conjunto de variables del grupo de estudiantes representado por las variables X_0,..,X_3. se muestra en la figura (1.1).

n=200		X_2=0		X_2=1	
		X_3=0	X_3=1	X_3=0	X_3=1
X_0=0	X_1=0	0	1	2	15
	X_1=1	0	1	1	15
X_0=1	X_1=0	2	5	2	19
	X_1=1	10	18	16	93

Figura 1.1: Ejemplo de Tabla de Contingencia

Existen diferentes formas de visualizar los datos de una tabla de contingencia. En general podemos dividir los modelos de graficación en dos grupos:

Grupo 1:

Los que visualizan el conjunto de datos. Entre estos métodos se encuentran los dendogramas, los algoritmos de agrupamiento, los gráficos de dispersión (scatterplots), etc.

Grupo 2:

Los que visualizan las relaciones, niveles o medidas de interacción existentes entre las variables o atributos que representan los datos. Entre estos métodos se encuentran los modelos gráficos.

En la sección 1.2, explicamos primero las principales medidas de asociación que son la base de diferentes métodos de visualización del grupo 2. En la sección 1.3 se muestran algunos ejemplos de métodos del grupo 1 para visualizar datos.

Como el interés en este trabajo de tesis es en datos binarios, no se incluyen métodos como análisis de correspondencias.

1.2 Medidas de Asociación como base de visualización de relaciones

Las medidas de asociación se refieren comunmente a medidas que muestran la relación entre dos o más variables o atributos para un conjunto de datos categóricos representados por medio de una tabla de contingencia,

A continuación se verán las medidas de asociación o proporción más utilizadas en la práctica [7].

1.2.1 Proporción de paridad

Dada una tabla de probabilidades de 2×2 la paridad para una variable X_0, representada en la tabla por un renglón o columna, está definida como la proporción siguiente,

$$paridad\,(X_0) = \frac{p\,(X_0 = 1)}{p\,(X_0 = 0)}$$

Esta razón o proporción de paridad, básicamente muestra cuantas veces el evento $X_0 = 1$ es más probable que el evento $X_0 = 0$. Si el resultado de esta proporción es 1, muestra que X_0 es equiprobable. La proporción de paridad condicional de X_0 dado $X_1 = 1$ esta dada por la siguiente formula,

$$paridad\,(X_0 \mid X_1 = 1) = \frac{p\,(X_0 = 1 \mid X_1 = 1)}{p\,(X_0 = 0 \mid X_1 = 1)} \tag{1.1}$$

1.2.2 Proporción de producto cruzado

Una de las medidas de asociación más utilizada para variables o atributos de un conjunto de datos es la medida de proporción de producto cruzado, en ingles "odds ratio", y se define en su forma más sencilla para una tabla de probabilidades de 2×2 de la siguiente manera.

$$\alpha = \frac{p\,(X_0 = 1 \cap X_1 = 1)\,p\,(X_0 = 0 \cap X_1 = 0)}{p\,(X_0 = 0 \cap X_1 = 1)\,p\,(X_0 = 1 \cap X_1 = 0)} = \frac{p_{11}p_{00}}{p_{01}p_{10}} \tag{1.2}$$

donde p_{ij}, indica el valor para la variable i o j representada en la tabla de probabilidades, de la siguiente manera

	$X_1 = 0$	$X_1 = 1$
$X_0 = 0$	p_{00}	p_{01}
$X_0 = 1$	p_{10}	p_{11}

La medida α de asociación juega un papel clave en la construcción de modelos log-lineales. Algunas de las propiedades básicas de la medida de proporción de producto cruzado son:

1. α es invariante al intercambio de renglones y columnas. Si se realiza un intercambio de solo renglones o solo columnas entonces α cambiará de valor a $1/\alpha$.

2. α es invariante bajo ciertas transformaciones. Esto es, suponiendo que multiplicamos las probabilidades en el renglón 1 y el renglón 2 por escalares a y b respectivamente y la columna 1 y 2 por escalares c y d respectivamente. Considerando que a, b, c y d son mayores a 0 y renormalizando estos valores para que sumen de nueva cuenta 1 por renglones y columnas, veremos que las constantes de normalización se cancelan obteniendo:

$$\frac{(\cancel{a}\cancel{c}p_{00})(\cancel{b}\cancel{d}p_{11})}{(\cancel{a}\cancel{d}p_{01})(\cancel{b}\cancel{c}p_{10})} = \frac{p_{11}p_{00}}{p_{10}p_{01}} = \alpha$$

3. La medida α puede ser vista en función de la proporción de paridad, de la siguiente forma,

$$\alpha = \frac{paridad\,(X_0 \mid X_1 = 1)}{paridad\,(X_0 \mid X_1 = 0)} = \frac{\frac{p_{11}}{p_{01}}}{\frac{p_{10}}{p_{00}}}$$

El valor de α va desde 0 hasta ∞, pero si obtenemos el logaritmo natural de α entonces el valor resultante va desde $-\infty$ a $+\infty$. La razón de producto cruzado es simétrica en el sentido que $log\,(\alpha)$ y $log\,(1/\alpha)$ representan el mismo grado de asociación pero en direcciones opuestas,

$$\log\left(\frac{1}{\alpha}\right) = -\log(\alpha)$$

1.2.3 Funciones generales de asociación

Algunas medidas de asociación para tablas de 2×2 son funciones monótonas crecientes o decrecientes de α.

Sea $f(\alpha)$ una función positiva monótona creciente de α, tal que

$$f(1) = 1$$

Entonces una medida de asociación normalizada basada en $f(\alpha)$ cuyo máximo absoluto es 1 se expresa de la siguiente forma,

$$g(\alpha) = \frac{f(\alpha) - 1}{f(\alpha) + 1}$$

Yule propuso 2 diferentes $f(\alpha)$.
La medida de asociación Q, (1900),

$$Q = \frac{p_{00}p_{11} - p_{01}p_{10}}{p_{00}p_{11} + p_{01}p_{10}} = \frac{\frac{p_{00}p_{11}}{p_{01}p_{10}} - 1}{\frac{p_{00}p_{11}}{p_{01}p_{10}} + 1} = \frac{\alpha - 1}{\alpha + 1}$$

y la medida de agrupación Y, (1911),

$$Y = \frac{\sqrt{p_{00}p_{11}} - \sqrt{p_{01}p_{10}}}{\sqrt{p_{00}p_{11}} + \sqrt{p_{01}p_{10}}} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}$$

Ambos Q y Y tienen valores en el rango de $[-1, 1]$, tomando el valor de 0 cuando las variables renglón y columna son independientes.

1.2.4 Covarianza

La covarianza de X_0 y X_1 está definida por,

$$\sigma_{0,1} = E\left[(X_0 - \mu_0)(X_1 - \mu_1)\right] = \sum_{x_0}\sum_{x_1}(x_0 - \mu_0)(x_1 - \mu_1)\,p_{x_0,x_1}$$

La covarianza entre dos variables aleatorias es una medida de la naturaleza de la asociación entre las dos. El signo de la covarianza indica si la relación es positiva o negativa. Cuando X_0 y X_1 son independientes la covarianza es cero. Aunque esta medida es utilizada comunmente para datos cuantitativos, también se puede utilizar en datos discretos. Otra forma de representar la covarianza es.

$$\sigma_{0,1} = E(X_0X_1) - \mu_0\mu_1$$

1.2.5 Coeficiente de correlación

Para el caso de variables binarias la varianza es,

$$\begin{aligned} \sigma_0^2 &= \mu_0 - \mu_0^2 = \mu_0(1 - \mu_0) = p_{0+}p_{1+} \\ \sigma_1^2 &= \mu_1 - \mu_1^2 = \mu_1(1 - \mu_1) = p_{+0}p_{+1} \end{aligned}$$

donde μ_0 y μ_1 son las medias de las marginales de renglón y columna para una tabla de 2×2 respectivamente.

$$\mu_0 = p_{1+}, \qquad \mu_1 = p_{+1}$$

La covarianza entre las dos variables es entonces,

$$\sigma_{0,1} = p_{11} - \mu_0\mu_1 = p_{11} - p_{1+}p_{+1}$$

El coeficiente de correlación de producto de momentos se obtiene entonces dividiendo la covarianza por la raíz cuadrada del producto de las varianzas, de la siguiente manera,

$$\rho = \frac{p_{11} - p_{1+}p_{+1}}{\sqrt{p_{0+}p_{1+}p_{+0}p_{+1}}} = \frac{p_{00}p_{11} - p_{10}p_{01}}{\sqrt{p_{0+}p_{1+}p_{+0}p_{+1}}}$$

El coeficiente ρ es invariante al intercambio de ambos renglones o columnas, y cambia solo de signo si se intercambia ya sea el renglón o la columna. Si el renglón y columna son

independientes, entonces $\rho = 0$. Si $p_{01} = p_{10} = 0$, entonces $\rho = 1$, y si $p_{00} = p_{11} = 0$, entonces $\rho = -1$.

1.2.6 Reglas de Asociación

En 1993 Agrawal, Imielinski y Swami [1] introdujeron una clase de regularidades denominadas "reglas de asociación", las cuales son formas simples, pero útiles de representación de conocimiento.

Definición 1 *Una regla de asociación es una expresión de la forma $A \Rightarrow B$, donde A y B son eventos asociados con un grupo de variables aleatorias dadas.*

A es denominado el cuerpo de la regla y B la cabecera de la regla. El significado intuitivo de dicha regla es que una observación que contiene A tiende a contener B. El cuerpo de la regla expresa la condición que debe ser cumplida para que la cabecera de la regla sea verdadera.

Características de las reglas de asociación

Comunmente se asocia con una regla de asociación un soporte, una confianza y una elevación, conceptos que definimos a continuación.

El soporte de una regla es la ocurrencia relativa de las reglas de asociación detectadas dentro del conjunto completo de datos. El soporte se calcula en términos del número de datos.

$$\begin{aligned} Soporte\,(A) &= \frac{\text{número de datos conteniendo } A}{\text{número total de datos}} \\ Soporte\,(A \Rightarrow B) &= \frac{\text{número de datos conteniendo } A \text{ y } B}{\text{número total de datos}} \end{aligned}$$

La confianza de una regla de asociación es la fuerza o veracidad de dicha regla, y se calcula de la siguiente manera.

$$Confianza\,(A \Rightarrow B) = \frac{Soporte\,(A \Rightarrow B)}{Soporte\,(A)}$$

La elevación es una medida de la desviación de independencia entre 2 variables. La elevación se expresa de la siguiente forma,

$$Elevación\,(A \Rightarrow B) = \frac{Confianza\,(A \Rightarrow B)}{Soporte\,(B)}$$

Para medir que tanto contribuye A a la predicción de B, se puede utilizar la medida de diferencia de confianzas *doc* (differences of confidences)[35], para las reglas $A \Rightarrow B$ y $\rceil A \Rightarrow B$.

$$doc\,(A \Rightarrow B) = Confianza\,(A \Rightarrow B) - Confianza\,(\rceil A \Rightarrow B)$$

En términos de probabilidad de eventos las características de las reglas de asociación se pueden definir de la siguiente manera:

Soporte

$$\begin{aligned} Soporte\,(A) &= P\,(A) \\ Soporte\,(A \Rightarrow B) &= P\,(A \cap B) \end{aligned}$$

Confianza

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

Elevación

$$Elevacion\,(A \Rightarrow B) = \frac{P\,(A \cap B)}{P\,(A)\,P\,(B)}$$

Diferencia entre confianzas

$$\begin{aligned} doc\,(A \Rightarrow B) &= P\,(B \mid A) - P\,(B \mid \rceil A) \\ &= \frac{P\,(B \cap A)}{P\,(A)} - \frac{P\,(B \cap \rceil A)}{P\,(\rceil A)} \\ &= \frac{P\,(B \cap A)\,P\,(\rceil A) - P\,(B \cap \rceil A)\,P\,(A)}{P\,(A)\,P\,(\rceil A)} \end{aligned}$$

Como $P\,(B \cap \rceil A) = P\,(B) - P\,(B \cap A)$,

$$\begin{aligned} doc\,(A \Rightarrow B) &= \frac{P\,(B \cap A)\,P\,(\rceil A) - P\,(B)\,P\,(A) + P\,(B \cap A)\,P\,(A)}{P\,(A)\,P\,(\rceil A)} \\ &= \frac{P\,(B \cap A)\,(P\,(\rceil A) + P\,(A)) - P\,(B)\,P\,(A)}{P\,(A)\,P\,(\rceil A)} \\ &= \frac{P\,(A \cap B) - P\,(A)\,P\,(B)}{P\,(A)\,P\,(\rceil A)} \end{aligned}$$

Tipos de Reglas

El tipo de la regla es la validez de una regla basada en la influencia de la dependencia estadística entre el cuerpo de la regla y la cabecera. El tipo de la regla está determinado usando la prueba Chi-cuadrada para independencia estadística. Existen tres tipos de reglas: positivas, negativas y neutras.

Positivas.- A tiene una influencia positiva sobre las ocurrencias de B, si la elevación de la regla es mayor que 1.

Negativas.- A tiene una influencia negativa sobre las ocurrencias de B, si la elevación de la regla está entre 0 y 1.

Neutras.- A y B son independientes, es decir la presencia de A no nos dice nada acerca de la probabilidad de que esté presente B. La elevación de la regla es igual a 1.

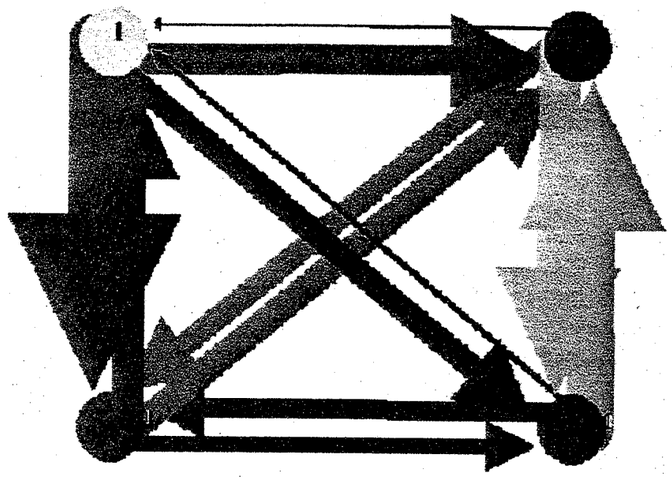


Figura 1.2: Ejemplo de Reglas de Asociación

1.2.7 Representación gráfica de Medidas de Asociación

Varias de las medidas de asociación conducen a alguna representación gráfica. En general el esquema genérico para construir una representación gráfica consiste en construir un grafo donde los nodos corresponden a las variables y una conexión refleja una interacción (asociación) particular. A continuación se muestra el tipo de representación utilizado comunmente para las reglas de asociación.

Las reglas de asociación se representan graficamente por medio de un grafo bi-dirigido, donde los nodos representan las variables pertenecientes al conjunto de variables multidimensionales S, y las flechas dirigidas representan las asociaciones entre las variables, es decir, la expresión de la forma $A \Rightarrow B$, se verá ejemplificada por medio de un flecha dirigida del nodo A al nodo B.

El ancho de las flechas dirigidas está relacionado con la confianza de la regla, entre más ancha sea la flecha, mayor será la confianza de la regla.

El color de la flecha indica el grado de elevación de la regla, puede también indicar en algunos casos el soporte de la regla. Entre más oscuro, mayor es la elevación de la regla.

En la figura (1.2) se muestra un ejemplo de reglas de asociación para el conjunto de datos de la figura (1.1) que se puede obtener en el programa Intelligent Miner de IBM. En este ejemplo se pueden apreciar dos reglas con bastante elevación (1⇒0) y (2⇒3).

En la figura (1.3) se muestra un tipo de representación alterno para las reglas de asociación, en este tipo de representación las reglas de asociación se muestran en una matriz, donde cada renglón corresponde al cuerpo de la regla, es decir, la parte izquierda, y las columnas corresponden a la cabera o parte derecha de la regla. Cada regla cumple con un mínimo soporte

y una mínima confianza establecida y se muestra en la matriz por medio de un cuadro. Las diferencias en intensidad de color del cuadro dependen de la confianza que cada regla tiene, por ejemplo reglas con un 99% de confianza tendrán un color más claro que las reglas que tengan un 100% de confianza. El tamaño de los cuadros esta dado por el soporte, de hecho el área es proporcional al cuadrado del soporte [35].

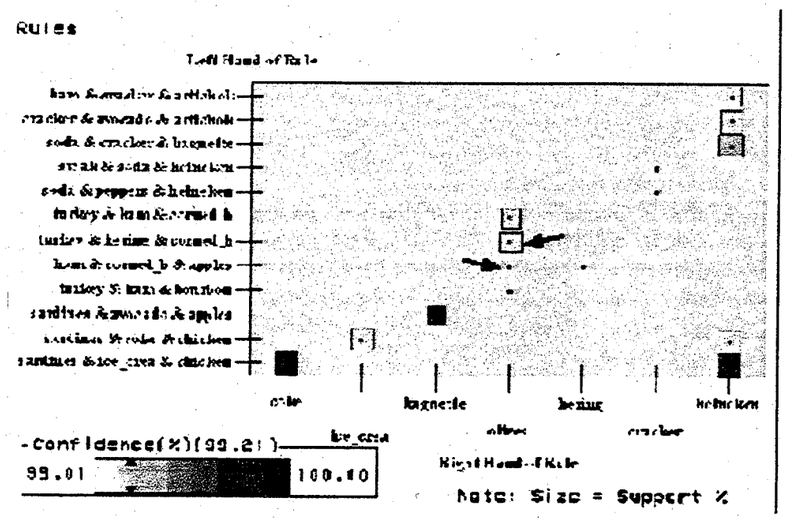


Figura 1.3: Ejemplo de Representación de Reglas de Asociación en SAS Enterprise Miner

1.3 Métodos para representar datos

1.3.1 Diagramas Parquet

Los diagramas Parquet fueron propuestos por Riedwyl y Schupbach [51] en 1983. Este tipo de diagramas Parquet se utiliza básicamente en el análisis de dos variables. Para mostrar un ejemplo de este tipo de diagramas y algunos otros gráficos en las siguientes secciones, se utilizarán los datos de la tabla 1. Esta tabla muestra los datos correspondientes a la relación entre el color del cabello y el color de ojos de 592 individuos [54], el objetivo principal al analizar este conjunto de datos es comprender la naturaleza de la asociación entre el color de cabello y el color de ojos.

	Color de cabello				
Color de ojos	Negro	Cafe	Rojo	Rubio	Total
Verde	5	29	14	16	64
Hazel	15	54	14	10	93
Azul	20	84	17	94	215
Cafe	68	119	26	7	220
Total	108	286	71	127	592

Tabla 1

Con cada celda se asocia un rectángulo cuyos anchos son proporcionales al total de la frecuencia en cada columna, n_{+j}, y cuyas alturas son proporcionales a las frecuencias totales en cada renglón, n_{i+}, entonces el área del rectángulo es proporcional a la frecuencia esperada, si hubiera independencia. Las frecuencias observadas se muestran por medio de un número de pequeños cuadros dentro de los rectángulos.

En la figura (1.4) se puede apreciar un diagrama parquet de los datos de la tabla 1. Este tipo de representación visual es muy útil, ya que las diferencias entre las frecuencias esperadas y las observadas aparecen como una densidad de sombreado en cada rectángulo. Se puede utilizar color para indicar cuando la desviación de independencia es negativa o positiva, o pueden utilizarse lineas punteadas para indicar una desviación negativa, y lineas continuas para indicar una desviación positiva, para gráficas sin color.

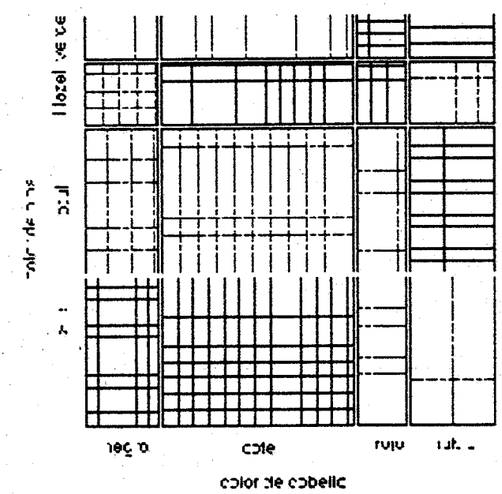


Figura 1.4: Diagrama Parquet

1.3.2 Gráficas de Mosaico

Las gráficas de mosaico (mosaic plots) [34], son una generalización recursiva de las gráficas de barras.

Para cada configuración de valores posibles de las variables, se asocia un rectángulo cuyo ancho es proporcional a la frecuencia marginal de las variables en cada columna. La altura de cada rectángulo está determinada por las probabilidades condicionales de las variables en el eje vertical dado el valor de las variables en el eje horizontal. Las celdas cuyo residuo es positivo tienen un contorno sólido, cuando el residuo es negativo el contorno es punteado. El color de los rectangulos esta determinado por el valor de los residuos bajo la hipótesis de independencia entre las variables. Un gráfico de mosaico sin color indicará independencia entre las variables. Pueden extenderse a más de dos variables [28].

En la figura (1.5) se muestra un gráfico de mosaico aumentado para los datos de la Tabla 1. En este gráfico los valores de los residuos se incluyen dentro de los rectangulos.

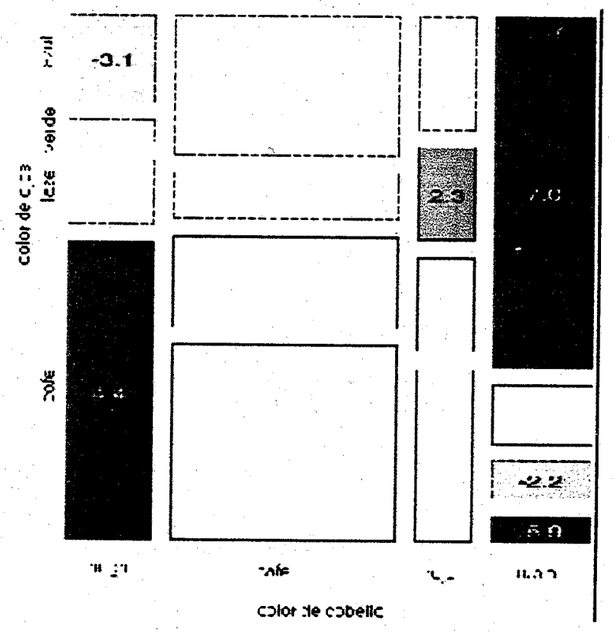


Figura 1.5:

1.3.3 Gráficos Fourfold

Este método gráfico está basado en la representación visual que mapea la frecuencia de una celda a un área. Está diseñado para desplegar tablas de 2x2 y de 2x2xk. La frecuencia n_{ij} de

cada celda de una tabla de fourfold se representa por medio de un cuarto de círculo, cuyo radio es proporcional a $\sqrt{n_{ij}}$, de tal manera que el área es proporcional al conteo en la celda [28].

El tipo de asociación entre las variables se muestra en el gráfico fourfold, si $\alpha \neq 1$, las celdas diagonales en una dirección difieren en tamaño con respecto a aquellas en la dirección opuesta. El gráfico fourfold, también utiliza color o sombreado para mostrar esta diferencia en dirección. Los anillos de confianza para el α observado permiten una prueba visual de la hipótesis $H_0 : \alpha = 1$. Los anillos en cuadrantes adyacentes se traslapan si y solo si los conteos observados son consistentes con la hipótesis nula. La figura (1.6) muestra el análisis realizado con un gráfico Fourfold sobre un conjunto de datos de admisiones realizadas en 1973 para estudios de posgrado en Berkeley para los seis departamentos más importantes. Se desea encontrar evidencia de preferencia en el sexo de los aplicantes con respecto al resultado [5].

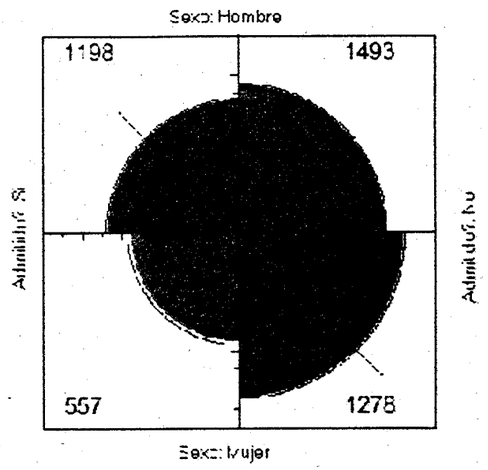


Figura 1.6: Gráfico Fourfold

La figura (1.6) muestra las frecuencias observadas numéricamente en las esquinas del gráfico. Entonces hubo 2691 aspirantes hombres, de los cuales 1193 (44.4%) fueron admitidos, comparando estos datos con las 1855 mujeres aspirantes de las cuales 557 (30%) fueron admitidas. La medida de asociación odds ratio es de 1.84 indicando que los hombres son admitidos en una proporción de casi el doble, es decir, que se admiten casi dos hombres por cada mujer. Como los cuadrantes sombreados no se alinean y los anillos de 99% de confianza alrededor de cada cuadrante no se traslapan, esto indica que la medida de asociación difiere significativamente de 1, si los cuatro cuadrantes estuvieran alineados, el odds ratio sería de 1. Finalmente el ancho de los anillos también da una representación visual de la precisión de los datos.

1.3.4 Agrupamiento

Las técnicas de agrpamiento (clustering) son comunmente usadas para encontrar dentro de un conjunto de datos los grupos o subconjuntos de datos que tienen características similares. Aunque están diseñados en primer lugar para datos continuos, se usan bastante para datos binarios.

El método de agrupamiento es un método iterativo definido de manera general como se muestra en el algoritmo siguiente:

Algoritmo 2 *Método de Agrupamiento*

1 Inicialización: Cada punto dado es un cluster

2 Repite lo siguiente hasta que la distancia sea mayor a un umbral establecido

i Busca dos clusters cercanos utilizando una formula para calcular la distancia

ii Combinar los clusters cercanos en un solo cluster

3 Terminar

El algoritmo termina cuando la distancia entre clusters rebasa un umbral predefinido y no se puede formar ningun nuevo cluster. También se pueden definir inicialmente el número de clusters determinado que se deseé encontrar.

Tipos de Agrupamiento

Existen dos tipos de métodos de agrupamiento principalmente: jerárquico y no jerárquico.

El método jerárquico consiste en formar grupos o clusters y representarlos por medio de un dendograma, este tipo de representación se muestra en la figura 1.7. Para este ejemplo se tomaron 50 muestras de los datos sobre alumnos y sus materias de la figura (1.1). Como se puede apreciar existen dos clusters iniciales, uno con 5 elementos, y el otro con 45 elementos.

Un ejemplo del tipo no jerárquico es el algoritmo de k-medias o k-means [24]. Este método consiste en dividir los datos en k grupos, tal que la variabilidad dentro de cada grupo sea mínima. El método k-means genera una solución aproximada al siguiente problema de optimización.

$$\text{minimizar } M(C) = \sum_{i=0}^{n-1} w_i D^2 (s_i, rep[s_i, C])$$

donde
$S = \{s_0, s_1, ..., s_{n-1}\}$ es un conjunto de n datos en el espacio real de dimension m, $\Re^m$.
El peso w_i refleja la relevancia de la observación s_i
$C = \{c_0, c_1, ..., c_{k-1}\}$ es el conjunto de k centros, o los puntos representativos de $\Re^m$.
$rep[s_i, C]$ es el punto más cercano en C a s_i, esto es,

$$D(s_i, rep[s_i, C]) = \min_{j \in \{0,...,k-1\}} D(s_i, c_j)$$

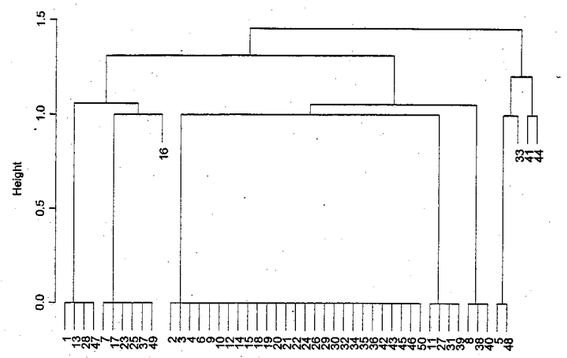


Figura 1.7: Ejemplo de un Dendograma

La medida de distancia D depende del conjunto de datos, para datos continuos es común utilizar la formula de distancia euclidiana, en datos discretos se utiliza comunmente la distancia de Hamming.

Distancia Máxima

$$D(x,y) = \max |x_i - y_i|$$

Distancia Mínima

$$D(x,y) = \min |x_i - y_i|$$

Distancia de Hamming

$$D(x,y) = \#\{i : x_i \neq y_i\}$$

Ejemplo:

$$\begin{array}{ccc} 00111 & \longleftrightarrow & 01011 \\ D & = & 2 \end{array}$$

1.4 Modelos Gráficos

Los modelos gráficos también denominados grafos de independencia condicional o redes de inferencia, se utilizan para la modelación de datos multivariados discretos de alta dimensión.

Un modelo gráfico muestra la estructura de interacción de independencia o dependencia condicional entre variables o atributos de un conjunto de datos. En los modelos gráficos los nodos representan atributos o variables, y cada arista denota una dependencia directa entre dos variables. Los modelos gráficos son usados para comprender mejor las relaciones existentes entre las variables y también son utilizados para facilitar el razonamiento en altas dimensiones, es por esto que los modelos gráficos son utilizados frecuentemente en sistemas expertos y sistemas de soporte de decisiones.

En la figura (1.8) se muestra un ejemplo de un modelo gráfico, en este caso para el conjunto de datos dado en la figura (1.1), con 4 variables. Entonces se puede apreciar visualmente que existen dos dependencias fuertes entre las variables X_0 y X_1 y entre las variables X_2 y X_3. Esta información nos dice que por ejemplo para el conjunto de datos de alumnos y las materias que aprobaron no hay independencias entre las materias de Ciencias Sociales y Español, y tampoco hay independencia entre las materias de Ciencias Naturales y Matemáticas. Por otro lado, si hay independencia entre por ejemplo Ciencias Sociales y Matemáticas.

Se puede ver que la figura (1.8) muestra alguna semejanza con la figura (1.2). En este caso las reglas de asociación con mayor confianza son las no independencias en el modelo gráfico. Sin embargo en un modelo gráfico se visualiza un modelo para todas las variables y no tanto un conjunto de patrones entre pares de variables.

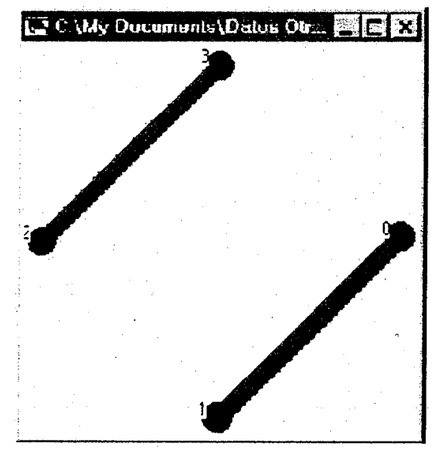


Figura 1.8: Ejemplo de Modelo Grafico

En este trabajo de tesis entonces nos enfocamos a los métodos de análisis de datos que unen la teoría de probabilidad y estadística con la teoría de grafos, los "Modelos Gráficos". En el capítulo 2 se vera la teoría base de los modelos gráficos y los tipos de modelos gráficos que existen, en el capítulo 3 se verán los métodos de estimación y selección de modelos, en el capítulo 4 se verán ejemplos de aplicación de estos modelos a diferentes conjuntos de datos, desde accesos a hojas Web, hasta conjuntos de datos proporcionados por investigadores en diversos artículos.

En el resto del trabajo de tesis, se plasman ejemplos de modelos gráficos y de modelos gráficos generalizados, que se han obtenido en un programa realizado en conjunto con este trabajo de tesis, denominado $\mathcal{MOGG}$ por "MOdelos Gráficos clásicos y Generalizados", desarrollado en el lenguaje de programación C++ Builder en la plataforma Windows. En el apéndice de este trabajo se muestran mayores detalles del programa dentro del manual de usuario.

1.4.1 Principales aportaciones

- En 2.4.1 se propone un algoritmo para encontrar independencias implicadas en modelos gráficos generalizados.
- En 2.4.3 se propone una visualización alterna para los modelos gráficos generalizados.
- En 3.3.1 se propone un método basado en el algoritmo de Newton Raphson para obtener los estimadores de máxima verosimilitud de un modelo gráfico generalizado.
- En 3.3.2 se propone la utilización de una solución inicial para el algoritmo de Newton Raphson.
- En 4.1.3 se propone la utilización de los modelos gráficos para el análisis de los accesos a un sitio Web.

Capítulo 2

Modelos Gráficos

2.1 Introducción

Los modelos gráficos son una unión entre la teoría de probabilidad y la teoría de grafos. Son una herramienta para modelar dos tipos de problemas comunes dentro de la matemática aplicada y la ingeniería: la incertidumbre y la alta dimensionalidad. Los modelos gráficos son modelos de probabilidad para observaciones aleatorias multivariadas, cuya estructura de independencia se caracteriza por medio de un grafo, que de manera visual ayuda a interpretar mejor y facilitar el entendimiento de las observaciones que representa el grafo.

A continuación, supongamos que las características de interés o atributos del conjunto de datos, son variables aleatorias discretas $X = (X_0, ..., X_n)$ donde $X_i \in \{0, 1\}$. Así la distribución subyacente esta definida por la probabilidad $p(X_0{=}x_0, X_1{=}x_1, ..., X_n{=}x_n)$. Se observa que en el caso $n = 0$, se obtiene la distribución Bernoulli, la cuál representa el tipo de datos categóricos discretos más simples.

2.1.1 Independencia

Definición 3 *Dos variables aleatorias X_0 y X_1 se dicen independientes, si su función de probabilidad conjunta se puede factorizar en el producto de sus probabilidades marginales de la siguiente forma,*

$$p(X_0 = x_0, X_1 = x_1) = p(X_0 = x_0)\, p(X_1 = x_1)$$

El concepto de independencia se puede ver como la ausencia de una relación existente entre las variables aleatorias. Para dos variables aleatorias X_i y X_j donde $i \neq j$, si X_i es independiente de X_j, denotado $X_i \perp X_j$, quiere decir que el evento que la variable X_i tome cierto valor es irrelevante para la probabilidad de que la variable X_j tome un cierto valor.

Para X_i, X_j y X_k, con i, j, k no iguales, la relación de independencia entre variables tiene las siguientes propiedades principales, entre otras.

1. No es reflexiva, ya que X_i no puede ser independiente de X_i
2. No es transitiva, ya que si $X_i \perp X_j$ y $X_j \perp X_k$, esto no implica que $X_i \perp X_k$.
3. Es simétrica, ya que $X_i \perp X_j$, implica que $X_j \perp X_i$.

2.1.2 Independencia condicional

Definición 4 *Sean X_0 y X_1 dos variables aleatorias discretas, la probabilidad condicional de que $X_0 = x_0$ ocurra dado que ocurrió $X_1 = x_1$ se escribe de la siguiente manera,*

$$p(X_0 = x_0 \mid X_1 = x_1) = \frac{p(X_0 = x_0, X_1 = x_1)}{p(X_1 = x_1)}$$

donde $p(X_0 = x_0 \mid X_1 = x_1)$ está definida solo si $p(X_1 = x_1) > 0$.

Definición 5 *Dadas las variables aleatorias discretas X_0, X_1 y X_2, se dice que X_0 y X_1 son condicionalmente independientes dado $X_2 = x_2$, si y solo si*

$$p(X_0 = x_0, X_1 = x_1 \mid X_2 = x_2) = p(X_0 = x_0 \mid X_2 = x_2)\, p(X_1 = x_1 \mid X_2 = x_2)$$

asumiendo que $p(X_2) > 0$,

Esto quiere decir que el evento que la variable X_0 tome cierto valor es irrelevante para la probabilidad de que la variable X_1 tome un cierto valor, dado que la variable X_2 toma un cierto valor, y se denota $X_0 \perp X_1 \mid X_2$.

La definición anterior puede ser ampliada para más variables, de tal manera que, $X_0 \perp X_1 \mid [X_2, X_3]$, significa que X_0 y X_1 son independientes dado cualquier valor en la partición del espacio muestral generado por las variables X_2 y X_3, de esta manera,

$$X_0 \perp X_1 \mid [X_2, X_3] \Leftrightarrow X_0 \perp X_1 \mid \begin{bmatrix} X_2 = 0, & X_3 = 0 \\ X_2 = 0, & X_3 = 1 \\ X_2 = 1, & X_3 = 0 \\ X_2 = 1, & X_3 = 1 \end{bmatrix}$$

Para i, j, k, l no iguales, la relación de independencia condicional entre variables aleatorias tiene las siguientes propiedades principales, entre otras [43] y [14]:

1. Es simétrica, ya que $X_i \perp X_j \mid X_k$, implica que $X_j \perp X_i \mid X_k$.
2. Se puede contraer, ya que si se tienen las siguientes dos independencias condicionales $X_i \perp X_j \mid X_k$ y $X_i \perp X_l \mid [X_j, X_k]$, entonces implica que $X_i \perp [X_j, X_l] \mid X_k$.
3. Se puede descomponer, si se tiene $X_i \perp [X_j, X_k] \mid X_l$, entonces se tienen las siguientes dos independencias condicionales $X_i \perp X_j \mid X_l$ y $X_i \perp X_k \mid X_l$.

2.2 Modelos log-lineales

Los modelos log-lineales son una parametrización particular de $p(X_0 = x_0, ..., X_n = x_n)$ para analizar tablas de contingencia. Su desarrollo ha hecho posible formular y ajustar patrones complejos de asociaciones entre los factores que clasifican una tabla multidimensional.

Los modelos log-lineales fueron introducidos por Birch en 1963 [6].

2.2.1 Expansión log-lineal

Tomando dos variables aleatorias binarias, se tiene la siguiente identidad,

$$\begin{aligned} p(X_0 = x_0, X_1 = x_1) &= p(0,0)^{(1-x_0)(1-x_1)}\, p(0,1)^{(1-x_0)x_1} \\ &\quad \cdot p(1,0)^{x_0(1-x_1)}\, p(1,1)^{x_0 x_1} \end{aligned}$$

válida para (x_0, x_1) en $\{0,1\}^2$. Para facilitar el trabajo tomamos los logaritmos de la identidad para p, para transformar la multiplicación de términos en una suma de términos. De esta manera se tiene el siguiente modelo log-lineal para dos variables aleatorias X_0 y X_1 [7] y [60],

$$\begin{aligned} \log p(X_0 = x_0, X_1 = x_1) &= \log p(0,0) + x_0 \log \frac{p(1,0)}{p(0,0)} + x_1 \log \frac{p(0,1)}{p(0,0)} \\ &\quad + x_0 x_1 \log \frac{p(1,1)\, p(0,0)}{p(0,1)\, p(1,0)} \end{aligned} \tag{2.1}$$

Haciendo una reparametrización a la ec.(2.1), se puede generalizar a una expansión log-lineal expresada en la ecuación (2.2),

$$\log p(X_0 = x_0, X_1 = x_1) = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_0 x_1 \tag{2.2}$$

El coeficiente β_0 es un término de normalización, β_1 y β_2 corresponden al logaritmo de la paridad de las variables aleatorias X_0 y X_1 respectivamente, definido en la ec. (1.1), finalmente β_3 define la medida de asociación entre las variables, es decir, es el logaritmo de la proporción de producto cruzado, definido en la ecuación (1.2).

En dos dimensiones en general, solo se pueden tener dos tipos de modelos distintos: cuando las dos variables involucradas son independientes, o cuando las dos están asociadas. Por ejemplo cuando existe independencia entre las variables X_0 y X_1, el cuarto término de la ec.(2.2) es cero, entonces se obtiene un modelo log-lineal como el que se muestra en la siguiente ecuación.

$$\log p(X_0 = x_0, X_1 = x_1) = \beta_0 + \beta_1 x_0 + \beta_2 x_1$$

Ahora en el caso para tres dimensiones, es decir, para tres variables aleatorias, un modelo log-lineal saturado (sin independencias) para las variables X_0, X_1, X_2 se representa mediante la ecuación siguiente,

$$\begin{aligned} \log p(X_0 = x_0, X_1 = x_1, X_2 = x_2) &= \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_0 x_1 \\ &\quad + \beta_5 x_0 x_2 + \beta_6 x_1 x_2 + \beta_7 x_0 x_1 x_2 \end{aligned}$$

y en general se tiene que para n dimensiones $X_0, ..., X_n$, el modelo log-lineal saturado se representa mediante la ecuación siguiente,

$$\log p(X_0 = x_0, ..., X_n = x_n) = \beta_0 + \beta_1 x_0 + ... + \beta_n x_n + ... + \beta_m x_0 x_1 ... x_n \tag{2.3}$$

donde $m = 2^n - 1$, es decir, se pueden tener hasta m términos dentro de un modelo log-lineal para n variables.

2.3 Modelos Gráficos Clásicos

Dada una distribución multivariada discreta $X = (X_1, ..., X_n)$, un modelo gráfico se puede definir de la siguiente manera.

Definición 6 *Un modelo gráfico representa una familia de distribuciones sobre $(X_0, ..., X_n)$ variables aleatorias multivariadas, que satisfacen las independencias condicionales de la forma*

$$X_i \perp X_j | (las\ otras\ variables). \tag{2.4}$$

contenidas en un grafo no dirigido $G(Z, L)$ donde Z es el conjunto de nodos asociados a las variables en el modelo, y L es el conjunto de aristas presentes entre los nodos. Cada independencia de la forma (2.4) corresponde a la ausencia de una arista en el grafo entre los nodos correpondientes [17].

Un ejemplo de un modelo gráfico para cuatro variables $X_0, ..., X_3$ se muestra en la figura (2.1)

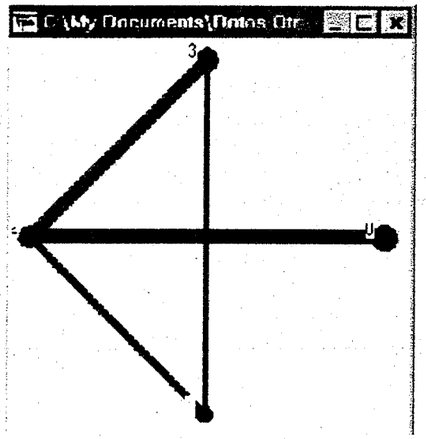


Figura 2.1: Ejemplo de Modelo Gráfico

El modelo log-lineal tomando la ec.(2.3) , es:

$$\begin{aligned} \log p\,(X_0 = x_0, .., X_3 = x_3) &= \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_0 x_2 \\ &\quad + \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_2 x_3 + \beta_9 x_1 x_2 x_3 \end{aligned}$$

Obsérvese que los modelos gráficos son una subclase de los modelos log-lineales.
Las independencias representadas en el modelo son:

$$\begin{aligned} (X_0 \perp X_1 | X_2 = x_2, X_3 = x_3) \qquad & \forall x_2, x_3 \\ (X_0 \perp X_3 | X_1 = x_1, X_2 = x_2) \qquad & \forall x_1, x_2 \end{aligned}$$

y se representan gráficamente mediante la ausencia de aristas entre las variables que son independientes.

2.3.1 Ventajas y desventajas de la utilización de modelos gráficos

Ventajas:

1. Ofrecen una interface visual entre el usuario y la probabilidad subyacente de los datos.
2. Ofrece una representación más compacta ayudando al usuario a que no se pierda en el gran número de parámetros que puede tener un modelo, esto hace que sean fáciles de interpretar.
3. Se basan en el concepto de independencia condicional, este concepto involucra la propiedad global de Markov, la cuál detalla un conjunto de reglas explícitas para interpretar el grafo de independencia.
4. Muchas de las propiedades en el modelo se pueden obtener fácilmente en términos de propiedades de la gráfica.
5. Pueden incorporar información apriori. El usuario puede especificar una estructura de independencia del dominio a ser modelado.
6. Existe una unificación entre tipos de variables, ya que pueden contener tanto variables continuas como discretas.

Desventajas:

1. Como reflejan en primer lugar las independencias, no permiten hacer un análisis "fino" de los tipos de dependencias. Más adelante, veremos una solución (parcial) para este problema.
2. Como en cualquier modelo multidimensional, un modelo gráfico aceptable no es único y no todos los diferentes modelos tienen algo en común.

2.3.2 Características de los grafos

Un grafo se denomina completo si contiene una arista entre cada par de nodos, es decir, todos los vertices estan unidos por una arista.

Un subconjunto de nodos S de un grafo G se denomina completo si existe una arista en G para cada par de nodos en S, es decir un subconjunto es completo si este induce un subgrafo completo.

Un subconjunto completo que es máximo es llamado clique, es decir un subconjunto completo que no está contenido en ningún otro subconjunto completo.

Cabe resaltar que en la literatura de Campos Aleatorios Markovianos, se consideran cliques a todos aquellos subconjuntos completos que no necesariamente son máximos. Pero en la literatura de modelos gráficos siempre que nos refiramos a cliques hablaremos de aquellos subconjuntos completos máximos.

Dados dos conjuntos de variables U y V se dicen estar separados de un conjunto de variables W con $U \cap V \subseteq W$, si todos los caminos entre U y V pasan por W [14].

2.3.3 Propiedades de Markov

Para un modelo gráfico dado, con un número finito de variables representadas por vertices, se tienen las siguientes propiedades:

- Propiedad de parejas de Markov: Si dos variables son no adyacentes, entonces son condicionalmente independientes dado las variables restantes.
- Propiedad local de Markov: Si una variable está condicionada solamente por las variables adyacentes, esta variable es independiente de todas las variables restantes. Esta propiedad define a los modelos gráficos no dirigidos, como Campos Aleatorios Markovianos.
- Propiedad global de Markov: Si U y V están separados por W, entonces $U \perp V \mid W$.

Las propiedades de Markov están relacionadas con la factorización de funciones de probabilidad, de tal manera que la probabilidad de un conjunto de variables aleatorias $p(X_0, ..., X_n)$ puede ser definida por medio de funciones más sencillas. Una forma de realizar la factorización es por medio de funciones potenciales.

Definición 7 *Se considera que se tiene una factorización de la función de probabilidad $p(X_0,..,X_n)$, si dada una serie de subconjuntos $C_0, ..., C_m$ del conjunto de variables $\{0, ..., n\}$, la función de probabilidad de X se puede escribir como un producto de $m+1$ funciones no negativas Ψ_i denominadas factores potenciales de la función de probabilidad,*

$$p(X_0, ..., X_n) = \prod_{i=0}^{m+1} \Psi_i(X_{c_i}) \tag{2.5}$$

donde $i = 0, ..., m+1$ y $X_{c_i} = \{X_j : j \in C_i\}$ [14].

Los conjuntos $C_0, ..., C_m$ no necesariamente son conjuntos disjuntos y las funciones Ψ_i no son funciones únicas.

Considerando el teorema de Hammersley y Clifford [33].

Teorema *Una distribución de probabilidad $p(.) > 0$ satisface la propiedad de parejas de Markov con respecto a un grafo no dirigido G, si y solo si este factoriza a $p(.)$ de acuerdo a los cliques de G.*

Se puede probar que para un modelo gráfico dado, cumplir la propiedad de parejas de Markov implica que las tres propiedades de Markov antes descritas son equivalentes [43].

Los cliques son utilizados dentro de los modelos gráficos para definir los conjuntos de variables que son utilizados para las funciones Ψ_i en la ecuación (2.5), para simplificar la obtención de la función de probabilidad $p(X_0, ..., X_n)$.

A continuación describimos un método propuesto Bron y Kerbosch en 1971 [10] para obtenerlos.

2.3.4 Obtención de los cliques en un grafo

La figura (2.2) muestra un grafo con 4 nodos (0,1,2,3). Los cliques para este grafo son los siguientes: (0,1,3) y (1,2,3).

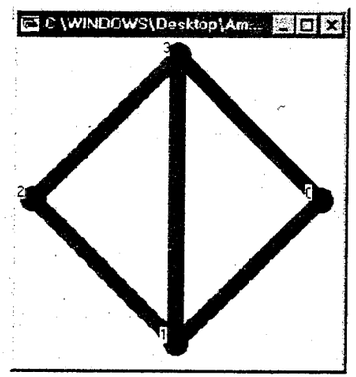


Figura 2.2: Grafo con 2 cliques

En la figura (2.3) se muestra un grafo con 6 nodos. en este caso hay tres cliques: (1,3,4,5), (1,2,5) y (0,5).

Un algoritmo eficiente para encontrar todos los cliques de un grafo no dirigido es el propuesto por Bron y Kerbosch en 1971 [10]. Este algoritmo obtiene los cliques de un grafo de manera recursiva, donde se van encontrando los nodos que pueden formar un subconjunto de nodos completo hasta que se encuentra un subconjunto que es máximo. El algoritmo cuenta principalmente con los siguientes tres conjuntos de datos:

1. El conjunto "Subgrafos".- Este conjunto es extendido cada vez que se encuentra un nodo candidato a formar un clique.
2. El conjunto "Candidatos".- Este conjunto contiene todos los nodos que pueden servir de extensión a un clique en el conjunto "Subgrafos".
3. El conjunto "No_Candidatos".-Este conjunto contiene todos los nodos que han servido en pasos anteriores como extensión del conjunto "Subgrafos".

Estos son los conjunto de datos que participan en el proceso, de esta manera el algoritmo propuesto Bron y Kerbosch [10] se puede expresar de la siguiente manera.

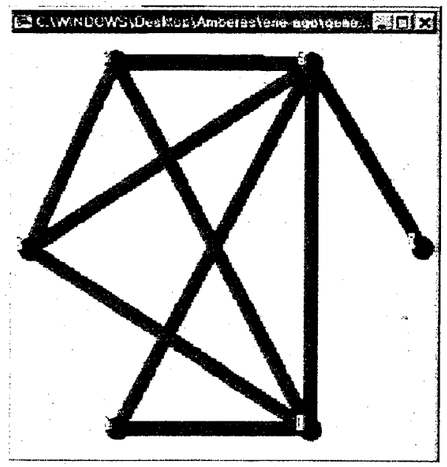


Figura 2.3: Grafo con 3 cliques

Algoritmo 8 *Para un grafo dado, hacer lo siguiente.*

1. *Almacenar todos los nodos todos los nodos del grafo en el conjunto "Candidatos".*
2. *Hacer conjuntos "Subgrafos" y "No_ Candidatos" vacíos.*
3. *Llama a función Extiende(Subgrafos, Candidatos, No_ Candidatos).*

Función Extiende(Subgrafos, Candidatos, No_ Candidatos)

1. if("*Candidatos*" $\neq \phi$)

   (a) *Seleccionar el nodo del conjunto "Candidatos" que tenga más aristas adyacentes.*

   (b) *Añadir nodo seleccionado al conjunto "Subgrafos".*

   (c) *Eliminar el nodo seleccionado del conjunto "Candidatos".*

   (d) *Asignar conjunto "Candidatos" a nuevo conjunto "Candidatos-Temp" y asignar conjunto "No_ Candidatos" a nuevo conjunto "No_ Candidatos-Temp".*

   (e) *Eliminar de los conjuntos "Candidatos-Temp" y "No_ Candidatos-Temp" los nodos que no sean adyacentes al nodo candidato seleccionado.*

   (f) if(("*Candidatos-Temp*" y "*No_ Candidatos-Temp*") $\neq \phi$ y "*No_ Candidatos–Temp*" *no contiene un nodo que sea adyacente a todos los nodos en "Candidatos-Temp"*)

      i. *Extiende(Subgrafos, Candidatos-Temp, No_ Candidatos-Temp)*

   (g) else if(("*Candidatos-Temp*" y "*No_ Candidatos-Temp*")== ϕ)

      i. *Almacenar conjunto "Subgrafos" en lista de cliques.*

   (h) *Eliminar del conjunto "Subgrafos" y "Candidatos" el nodo candidato seleccionado.*

   (i) *Colocar el nodo candidato seleccionado en el conjunto "No_ Candidatos".*

El algoritmo termina cuando se han encontrado todos los posibles cliques en el grafo, esto sucede cuando el conjunto "Candidatos" inicial es vacío. Este algoritmo genera cliques en un orden no predecible, pero tiende a generar los cliques más grandes primero y generar secuencialmente cliques con las mayores intersecciones en común. La optimización utilizando "salto de ramificaciones", se utiliza cuando se encuentra un nodo en el conjunto "No_candidatos" que es adyacente a todos los nodos en "Candidatos", esto mejora notablemente la rapidez del algoritmo.

Ejemplo

Tomando el grafo de la figura (2.2), los siguientes pasos muestran el procedimiento que sigue el algoritmo para la obtención de los cliques.

	Subgrafos	Candidatos	No_Candidatos
Inicio $\Rightarrow$	ϕ	(0,1,2,3)	ϕ
Seleccionar nodo 1 $\Rightarrow$	(1)	(0,2,3)	ϕ
Seleccionar nodo 3 $\Rightarrow$	(1,3)	(0,2)	ϕ
Seleccionar nodo 0 $\Rightarrow$	(1,3,0)	ϕ	ϕ
	(1,3)	(2)	(0)
Seleccionar nodo 2 $\Rightarrow$	(1,3,2)	ϕ	ϕ
	(1,3)	ϕ	(0,2)
	(1)	ϕ	(0,2,3)
"Candidatos" inicial == $\phi \Rightarrow$	ϕ	ϕ	(0,1,2,3)

2.3.5 Representación Numérica de Grafos

Un grafo puede ser representado numéricamente utilizando matrices.

Matriz de adyacencia

Sea $G = (Z, L)$ un grafo de n nodos y sea $A\,(a_{ij})$ una matriz $n \times n$, donde

$$a_{ij} = \begin{cases} 1, & \text{si } L_{ij} \in L \\ 0, & \text{en caso contrario} \end{cases} \quad ; \forall i \neq j$$

La matriz A se denomina matriz de adyacencia del grafo G. Cuando $a_{ij} = 0$, entonces no existe ninguna arista del nodo Z_i al nodo Z_j. Cuando $a_{ij} = 1$ indica que el nodo Z_i está conectado al nodo Z_j, o que los nodos son adyacentes.

Características de la matriz de adyacencia

La matriz de adyacencia de un grafo no dirigido es simétrica.

Dado que $L_{ii} \notin L$ para todos los valores de i, los elementos diagonales de A son nulos.

Matriz de alcanzabilidad

La matriz de alcanzabilidad, $T = (t_{ij})$, de un grafo G se define como

$$t_{ij} = \begin{cases} 1, & \text{si existe algún camino del nodo } Z_i \text{ al nodo } Z_j \\ 0, & \text{en caso contrario} \end{cases}$$

La matriz de alcanzabilidad está claramente relacionada con las potencias de la matriz de adyacencia.

Dado un grafo con n nodos, si existe un camino del nodo Z_i al nodo Z_j, entonces también existe un camino de longitud menor que n de Z_i a Z_j. Por tanto, la matriz de alcanzabilidad puede ser obtenida a partir de un número finito de potencias de la matriz de adyacencia, $A, A^2, A^3, ..., A^{n-1}$. El número de potencias necesario es $n-1$, entonces se tiene

$$t_{ij} = \begin{cases} 0, & \text{si } a_{ij}^k = 0, \forall k < n \\ 1, & \text{en caso contrario} \end{cases}$$

2.4 Modelos Gráficos Generalizados

En los modelos gráficos clásicos las independencias condicionales se tienen que asumir para todos los valores de las variables restantes. En la practica esta restricción es a veces muy fuerte y no permite obtener o descubrir más información sobre la interacción que hay entre los datos. En ocasiones se quisiera tener independencias condicionales para solo algunos valores de las variables, por ejemplo tener que X_0 es independiente de X_1 dado que $X_2 = 0$, es decir, cuando X_2 tiene un valor de 1 no hay independencia entre X_0 y X_1. Los modelos separables (split models) [36], resuelven este problema, separando los datos según el valor de la variable X_2, de esta manera generalizan los dos conjuntos separados obteniendo un modelo para cuando $X_2 = 0$ y otro modelo para $X_2 = 1$. Este tipo de modelos tienen la siguiente desventaja: una variable que ha sido utilizada para la separación de modelos, en este caso X_2, no puede ser utilizada para establecer independencias con respecto a otras variables dentro de ese modelo. Un modelo gráfico generalizado no tiene esa desventaja, ya que evita esas restricciones dentro del conjunto de independencias que se establecen.

Tomando la definición de Modelos Gráficos Generalizados [38].

Definición 9 *Un modelo gráfico generalizado se define como una familia de distribuciones caracterizadas por un conjunto de independencias de la forma*

$$X_i \perp X_j \mid \{X_k = x_k, k \notin \{i, j\}\}$$

Un modelo gráfico generalizado es un modelo gráfico clásico con la característica adicional que si dos variables son independientes solamente para ciertos valores de las variables restantes, entonces se dibuja una conexión entre los nodos y se agrega una etiqueta a la arista indicando los casos en los cuales no son independientes.

En la figura (2.4) se muestra un ejemplo de un modelo gráfico generalizado para 3 variables. En este ejemplo solo se tiene una independencia.

$X_1 \perp X_2 \mid X_0 = 0$

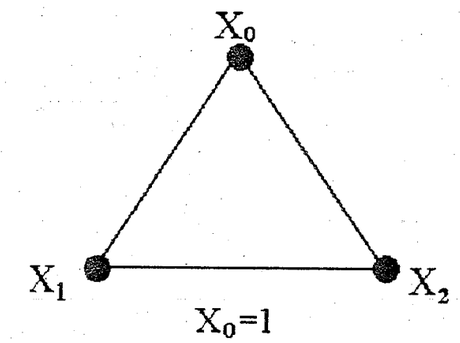


Figura 2.4: Ejemplo de Modelo Gráfico Generalizado

A diferencia de los modelos gráficos clásicos, donde cualquier grafo especifica una familia no vacía de distribuciones, en los modelos gráficos generalizados, se debe tener especial cuidado en las independencias especificadas, porque existen relaciones entre las independencias, que pueden llevar a que un modelo gráfico generalizado esté mal especificado, y por lo tanto no sea válido. A este problema se le llama "problema de consistencia".

2.4.1 Consistencia de un Modelo Gráfico Generalizado

Un modelo gráfico generalizado, se dice consistente cuando no se han encontrado independencias implicadas que contradigan alguna no independencia en el modelo. Esto está definido por la siguiente propiedad [38].

Propiedad 1:

Suponiendo que X es una variable aleatoria multivariada binaria, si

$$\begin{aligned} X_0 \perp X_1 &\mid X_2 = x_2, X_3 = x_3, ..., X_n = x_n \\ X_0 \perp X_1 &\mid X_2 = 1 - x_2, X_3 = x_3, ..., X_n = x_n \\ X_0 \perp X_2 &\mid X_1 = x_1, X_3 = x_3, ..., X_n = x_n \end{aligned}$$

entonces

$$X_0 \perp X_2 \mid X_1 = 1 - x_1, X_3 = x_3, ..., X_n = x_n$$

La prueba para esta propiedad se encuentra en [38].

Tomando como ejemplo el modelo gráfico de la figura (2.5) se tiene que:

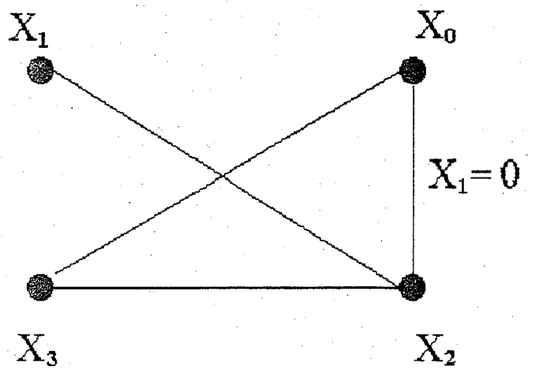


Figura 2.5: Inconsistencia en un Modelo Grafico Generalizado

Hay 2 independencias definidas para todos los valores restantes de las variables, expresadas de manera condensada de la siguiente forma.

$$\begin{aligned} X_0 \perp X_1 &\mid X_2 = x_2, X_3 = x_3 \\ X_1 \perp X_3 &\mid X_0 = x_0, X_2 = x_2 \end{aligned} \tag{2.6}$$

La tercera independencia está definida para algunos valores de las variables restantes,

$$X_0 \perp X_2 \mid X_1 = 1, X_3 = x_3 \tag{2.7}$$

Las independencias descritas en (2.6) y (2.7), por la propiedad 1, determinan la siguiente independencia implicada,

$$X_0 \perp X_2 \mid X_1 = 0, X_3 = x_3 \tag{2.8}$$

Como se puede observar, las independencias implicadas encontradas contradicen el modelo, ya que en este se define una no independencia entre X_0 y X_2, cuando $X_1 = 0$. Por lo tanto este modelo gráfico generalizado no es consistente.

Para encontrar las independencias implicadas de un modelo gráfico generalizado, es conveniente utilizar hipercubos para representar las interacciones entre las n variables binarias.

Definimos para $(X_0, ..., X_n)$ un hipercubo de dimensión $(n+1)$,con longitud de arista 1, donde cada vertice es de la forma $(x_1, ..., x_n)$ y $x_i \in \{0, 1\}$. Cada vertice corresponde a una cadena binaria de longitud n y dos vertices están unidos por una arista cuando sus cadenas binarias difieren solamente en un bit. En la figura (2.6) se muestra un hipercubo de dimensión 3 para (X_0, X_1, X_2).

De esta manera cada independencia y no independencia en el modelo gráfico es asociada con una cara del hipercubo y se dice que cada cara del hipercubo que define una independencia

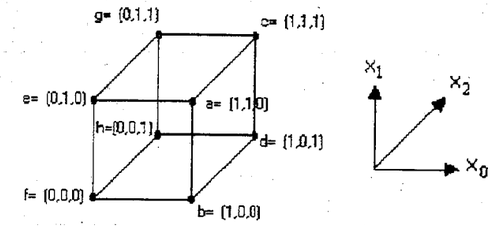


Figura 2.6: Hipercubo de 3 dimensiones

es una cara marcada [56]. De esta forma la independencia,

$$\begin{aligned} X_i \perp X_j &\mid X_1 = x_1, ..., X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \\ ..., X_{j-1} &= x_{j-1}, X_{j+1} = x_{j+1}, ..., X_n = x_n \end{aligned}$$

se asocia con la cara definida por los vertices,

$$\begin{aligned} &(x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_{j-1}, 0, x_{j+1}, ..., x_n) \\ &(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_{j-1}, 1, x_{j+1}, ..., x_n) \\ &(x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_{j-1}, 1, x_{j+1}, ..., x_n) \\ &\text{y} \\ &(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_{j-1}, 0, x_{j+1}, ..., x_n) \end{aligned}$$

A continuación se define el concepto de camino de implicación

Definición 10 *Dada una serie de caras marcadas en un hipercubo, un camino de implicación es aquel camino dentro del hipercubo que comenzando con una arista de una cara no marcada permite realizar una caminata a través de aristas paralelas pertenecientes a caras marcadas adyacentes. Se dice que hay una implicación cuando a partir de una arista de una cara no marcada se puede llegar a la arista paralela a la arista inicial de la misma cara no marcada a través de un camino de implicación.*

La cara no marcada que tiene la arista inicial y final en el camino de implicación corresponde a una independencia implicada, es decir una independencia que no estaba definida en el modelo gráfico generalizado, pero que el conjunto de independencias definidas requieren de esta independencia para definir una distribución. En [56] se muestra que cada independencia implicada es caracterizada por un camino de implicación.

Una cara marcada puede representarse numéricamente como un vector de izquierda a derecha (variable 0, variable 1, ..., variable n), donde las dos variables independientes colocan un * y las restantes el valor que tengan. De esta manera la independencia,

$$X_0 \perp X_1 \mid X_2 = 0$$

se puede representar con el vector (*,*,0). La representación vectorial de las aristas del hipercubo se expresa colocando un * solo en una de las variables independientes y la otra variable independiente toma algún valor, de esta manera las aristas para el vector (*,*,0) pueden representarse como: [0,*,0], [*,1,0], [1,*,0] y [*,0,0]. En la figura (2.7) se muestra con mayor claridad este tipo de representación y su relación con el hipercubo de tres dimensiones.

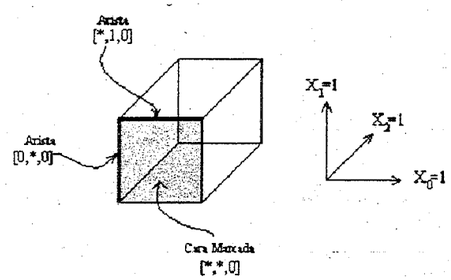


Figura 2.7: Ejemplo de representación numérica de una cara marcada y dos de sus aristas

Ahora proponemos un algoritmo para encontrar independencias implicadas en un modelo gráfico generalizado.

Algoritmo para encontrar independencias implicadas

El algoritmo propuesto en esta tésis para encontrar independencias implicadas utiliza como parte central una función recursiva que permite explorar en un hipercubo cada una de sus aristas y las aristas paralelas a ellas. Para realizar esto primeramente se convierten todas las independencias y dependencias encontradas en el modelo gráfico generalizados a representaciones numéricas en términos de caras marcadas (independencias) y caras no marcadas (no independencias). Estas caras y sus respectivas aristas son almacenadas en una lista de caras marcadas y en una lista de caras no marcadas respectivamente. Una cara no marcada que se encontró implicada se almacena en una lista de caras implicadas. A continuación se muestra como ejemplo la estructura de la lista de caras marcadas.

Cara 1		Cara 2		Cara N
Arista 1		Arista 1		Arista 1
Arista 2	$\mapsto$	Arista 2	$\mapsto \ldots \mapsto$	Arista 2
Arista 3		Arista 3		Arista 3
Arista 4		Arista 4		Arista 4

Lista de caras marcadas

A continuación se muestra el algoritmo propuesto en este trabajo de tesis para encontrar independencias implicadas

Algoritmo 11 *Para un modelo gráfico generalizado con independencias, hacer lo siguiente.*

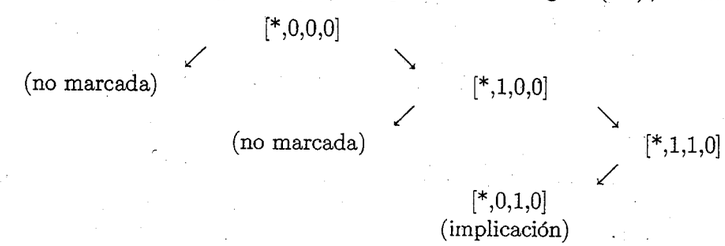
1. Para cada *par de variables X_i, X_j que en al menos una asignación de valores para las demás variables son independientes y en al menos una son dependientes, hacer lo siguiente.*
   (a) *Lista_no_marcadas*$\leftarrow$*Caras no marcadas y sus aristas.*
2. *Lista_marcadas*$\leftarrow$*Caras marcadas y sus aristas.*
3. *num_implicaciones* $\leftarrow 1$
4. Mientras *num_implicaciones* > 0, *hacer lo siguiente.*
   (a) *num_implicaciones*$\leftarrow 0$
   (b) Mientras *no se llegue al final de la Lista_no_marcadas.*
      i. *Toma arista k de Lista_no_marcadas.*
      ii. *Llama a función Analiza(k, k).*
5. *if (Lista_implicadas es no vacía), el modelo no es válido.*

Función Analiza(i,k)

1. if ($i \neq k$ *y k e i definen una cara no marcada, hacer lo siguiente.*)
   (a) *Guardar cara no marcada en Lista_marcadas y en Lista_implicadas.*
   (b) *Remover cara no marcada de la Lista_no_marcadas.*
   (c) *num_implicaciones* $\leftarrow$ *num_implicaciones* $+ 1$
2. else
   (a) Para cada *arista j de Lista_marcadas, tal que la arista j es paralela a la arista i y la arista j y la arista i definen una cara marcada, hacer lo siguiente.*
      i. *Analiza(j, k)*

Utilizando este algoritmo para el ejemplo de la figura (2.5), la figura (2.8) muestra con unas flechas el camino de implicación, sobre las aristas marcadas en color azul, que sigue el algoritmo. Como se puede apreciar, las caras en color gris corresponden a las independencias implicadas mostradas en la ecuación (2.8).

A continuación se muestra el recorrido que el algoritmo realiza sobre las aristas para encontrar la primera independencia implicada. Tomando como cara no marcada inicial (*,0,*,0), correspondiente a la cara en gris del primer hipercubo de la figura (2.8), se tiene:



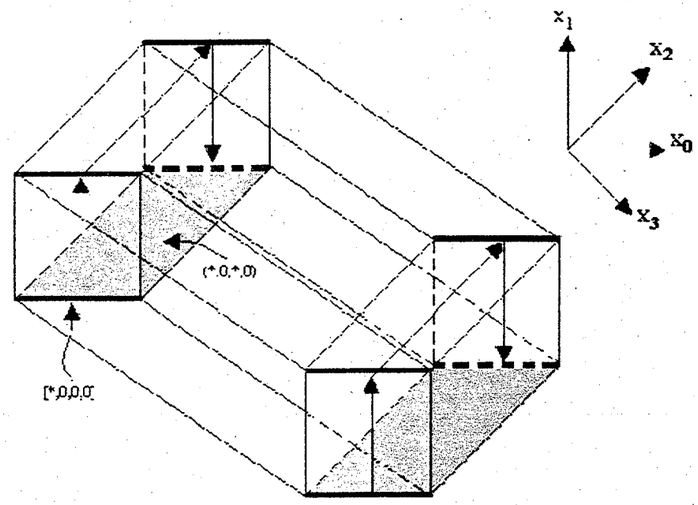


Figura 2.8: Ejemplo dos caminos de implicación

2.4.2 Relación entre Oddsratios

El problema de inconsistencia surge porque los oddsratios están relacionados entre si. Tomando como ejemplo el caso de tres variables como se muestra en la figura (2.6).

Tomando la ecuación (1.2) y asignando una probabilidad $p(.)$ a cada combinación de valores de las variables representadas por vertices en el hipercubo. El oddsratio θ_1, relacionado con la cara del hipercubo formada por los vertices (a,b,c,d), se define de la siguiente manera,

$$\theta_1 = pcr\,(X_1, X_2 \mid X_0 = 1) = \frac{p\,(1,1,1)\,p\,(1,0,0)}{p\,(1,0,1)\,p\,(1,1,0)} \tag{2.9}$$

Y de igual manera las caras θ_2,θ_3,θ_4, para las caras con vertices *(a,b,e,f)*, *(e,f,g,h)* y *(c,d,g,h)*, respectivamente. Se puede mostrar que se puede determinar el valor θ_1, si se conocen θ_2,θ_3,θ_4, con la siguiente ecuación,

$$\theta_1 = \theta_2^{-1}\theta_3\theta_4 \tag{2.10}$$

Prueba:
Sustituyendo los valores de $\theta_2, \theta_3, \theta_4$ en la ecuación (2.10) , se tiene,

$$\theta_1 = \frac{\cancel{p}(0,1,0)\boxed{p(1,0,0)}\cancel{p}(0,1,1)\cancel{p}(0,0,0)\boxed{p(1,1,1)}\cancel{p}(0,0,1)}{\boxed{p(1,1,0)}\cancel{p}(0,0,0)\cancel{p}(0,0,1)\cancel{p}(0,1,0)\cancel{p}(0,1,1)\boxed{p(1,0,1)}}$$

Esto se puede generalizar para n variables. En consecuencia, si se tienen algunas independencias sabemos que ciertos θ's serán igual a 1. Por la ecuación (2.10), si θ_2,θ_3,θ_4 son igual a 1, se obtiene que $\theta_1 = 1$, en otras palabras, se obtiene una independencia implicada.

2.4.3 Visualización alterna de Modelos Gráficos Generalizados

En lugar de la representación para modelos gráficos generalizados con etiquetas en las aristas, se propone en este trabajo de tesis la utilización de una visualización alterna, donde las aristas se dividen en segmentos rectangulares con diferentes tamaños y colores. Un ejemplo de la representación propuesta para el ejemplo de la figura (2.4) , se muestra en la figura (2.9) . Este ejemplo muestra la representación alterna de modelos gráficos generalizados y en él se pueden encontrar las siguientes extensiones propuestas.

1. Dividir las aristas en segmentos rectangulares. Cada segmento presente representa una no independencia entre las variables adyacentes a la arista para una asignación de valores de las demás variables. La asignación de valores está ordenada de un nodo menor a un nodo mayor de manera ascendente, por ejemplo en la figura (2.9) del nodo 0 al nodo 1. En este caso el segmento verde y gris representan una interacción cuando $X_2 = 0$, y $X_2 = 1$, respectivamente.

2. La ausencia de un segmento en una arista, indica independencia entre las variables adyacentes a la arista dado el valor de las variables restantes correspondientes.

3. Colocar colores de relleno a los segmentos. Si un segmento tiene color rojo, esto indica una medida de asociación de correlación o covarianza (dependiendo de la opción que se haya elegido) negativa entre las variables, si el segmento es de color verde, esto indica una relación positiva. Si un segmento tiene color gris, como se verá en el capitulo siguiente, se puede aceptar la hipótesis de independencia correspondiente al nivel de significancia proporcionado por el usuario.

4. El ancho de cada segmento es proporcional a la magnitud de la asociación entre las variables correspondientes.

2.5 Historia de los Modelos Gráficos

La utilización de modelos gráficos para el análisis estadístico de datos ha ido aumentando paulativamente desde hace 20 años. En 1980 Darroch, Lauritzen y Speed [17] introdujeron los modelos gráficos para variables discretas, juntaron los conceptos e ideas de teoría de grafos, independencia condicional y modelos log-lineales. Mostraron como un subconjunto de modelos log-lineales, los modelos gráficos, pueden ser fácilmente interpretados aplicando las propiedades de Markov de sus grafos de independencia asociados. Estos modelos utilizan un grafo para representar un modelo estadístico.

En 1984 Kiiveri, Speed y Carlin [41] introdujeron los modelos gráficos recursivos, los cuales se representan por un grafo dirigido acíclico en el cuál los nodos representan variables y las

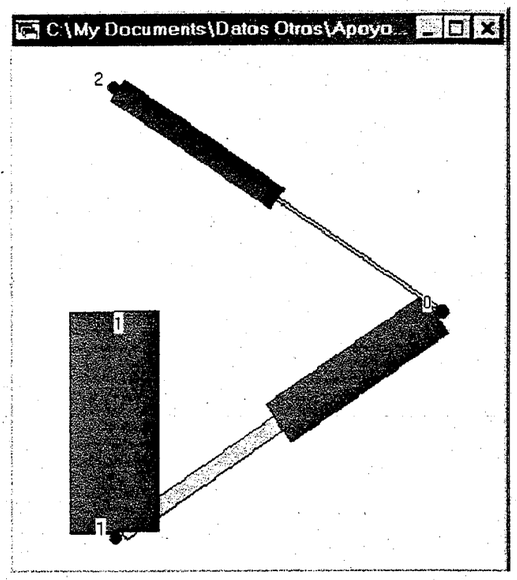


Figura 2.9: Representación Visual de un Modelo Gráfico Generalizado

aristas dirigidas representan asociaciones. Las independencias condicionales son determinadas por las aristas y su orientación.

En 1989 Lauritzen y Wermuth [45] introdujeron los modelos gráficos para variables mixtas, donde algunas variables son discretas y otras continuas. Las variables discretas y las variables continuas se representan con diferentes tipos de nodos. Estos modelos se utilizan para modelos gráficos para tablas de contingencia y modelos de selección de covarianza.

También en 1989 Frydenberg y Lauritzen [29] y en 1990 Wermuth y Lauritzen [59], introdujeron los modelos recursivos a bloques o modelos gráficos de cadena, dichos modelos incorporan asociaciones simétricas y directas entre variables. Estos tipos de modelos abarcan los modelos gráficos no dirigidos y los recursivos, como casos especiales y se representan gráficamente por medio de grafos de cadena que contienen aristas dirigidas y no dirigidas.

En 1990 Edwards [20], [21], introdujo modelos de interacción jerárquica, donde se incluyen los modelos gráficos mixtos como un caso especial de modelos.

En 1995 Hojsgaard y Thiesson [37] tratan en su artículo, la forma de realizar inferencias en modelos recursivos a bloques, así como también presentan una aplicación para el diagnóstico de enfermedades de la arteria coronaria. También introducen el programa BIFROST (Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques), el cuál facilita la inferencia automática en modelos recursivos a bloques con base en modelos gráficos no dirigidos, así como la transformación de un modelo recursivo a bloques en un modelo recursivo, el cuál puede ser exportado al shell de sistemas expertos HUGIN [2].

En 1998 Hojsgaard [36] en su tesis doctoral introduce los modelos gráficos separables o split models, que básicamente tratan de resolver el siguiente problema: "¿Como incorporar en los modelos gráficos que la estructura cualitativa de asociación entre algunas variables difiere dependiendo de los valores de las variables?". También muestra aplicaciones variadas de estos modelos utilizando el programa BIFROST.

Capítulo 3

Estimación y Selección de Modelos

3.1 Introducción

En este capítulo se muestran los métodos y algoritmos de selección y estimación utilizados para obtener modelos gráficos válidos, que ayudan a encontrar de una manera más eficaz las independencias más relevantes en los modelos gráficos clásicos y los modelos gráficos generalizados.

Para mostrar los métodos utilizados para encontrar modelos gráficos clásicos y generalizados válidos, se utilizará un conjunto de datos correspondiente a la circunstancia en la cual ocurrieron accidentes entre jugadores de futbol americano [13]. El conjunto de datos en forma de tabla de contingencia se muestran en la figura (3.1). La variable X_0 indica si el accidente ocurrió en defensa ($X_0 = 0$) o en ataque ($X_0 = 1$), X_1 indica si el accidente ocurrió aventando el balón ($X_1 = 0$) o no ($X_1 = 1$) y X_2 indica si el accidente ocurrió en una tacleada ($X_2 = 0$) o en una bloqueada ($X_2 = 1$).

n=725		X_2=0	X_2=1
X_0=0	X_1=0	125	129
	X_1=1	85	31
X_0=1	X_1=0	216	61
	X_1=1	62	16

Figura 3.1: Tabla de Contingencia para el conjunto de datos de los accidentes de los jugadores de futbol

A continuación usaremos la siguiente notación.

n_x denota el número de observaciones en la celda x. En el caso de una tabla 2×2 , n_{ij} denota la frecuencia en la celda i, j de la tabla de contingencia. Un signo "+" en un subíndice indica sumatoria sobre todo valor de esa variable. Por ejemplo,

$$n_{i+} = \sum_j n_{ij}, \qquad n_{+j} = \sum_i n_{ij}$$

n_{++}, denota el número total de observaciones en la tabla,

$$n_{++} = \sum_{i,j} n_{ij}$$

3.2 Construcción de Modelos Gráficos Clásicos

El proceso de construcción de un modelo gráfico consiste básicamente en los siguientes cuatro pasos:

1. Formulación de un modelo gráfico particular.
2. Obtener los estimadores de máxima verosimilitud correspondientes.
3. Realizar una prueba de hipótesis de estos estimadores.
4. Obtener el valor de P del modelo.

Si el valor de P es mayor a un umbral de significancia establecido, típicamente de 5% o 10%, el modelo gráfico no es rechazado.

3.2.1 Estimadores de máxima verosimilitud

Como se vió en secciones anteriores, la distribución de las variables aleatorias discretas X en los modelos gráficos pueden ser parametrizadas por un modelo log-lineal. Dado que este modelo está determinado por valores de parámetros desconocidos, se requiere de una medida de proporción que nos indique la divergencia existente entre este modelo log-lineal y los datos observados. La función de verosimilitud proporciona esta medida y expresa que tan probable (plausible) es observar un conjunto de datos para una cierta asignación de valores para los parámetros del modelo

Suponiendo que las observaciones provienen de una muestra, la verosimilitud es igual a,

$$\prod_x p(N_x = n_x) \qquad ó \qquad \prod_x p(x)^{n_x}$$

Tomando el logaritmo se obtiene.

$$l(p; n) = \sum_x n_x \log p(x) \tag{3.1}$$

Si parametrizamos $p(x)$ en términos de parámetros β obtenemos,

$$l(\beta; n) = \sum_x n_x \log p_\beta(x) \tag{3.2}$$

El objetivo es minimizar la divergencia entre el modelo log-lineal y lo observado, esto es equivalente a maximizar la función de log-verosimilitud,

$$\max_\beta l(\beta; n) \tag{3.3}$$

Para realizar esto se iguala con cero la derivada de la función $l\ (\beta; n)$ en la ecuación (3.2) con respecto a cada parámetro β y se resuelve el sistema de ecuaciones resultante.

A continuación veremos como se resuelve este problema en el caso de una tabla de 2×2 y posteriormente se verá un algoritmo iterativo usado frecuentemente para resolver este problema.

Estimación en dos dimensiones

El caso de dos dimensiones es el caso más sencillo de estimacion de parámetros, de esta manera para una tabla de contingencia de 2×2, bajo el modelo,

$$\log \widehat{p}(x_0, x_1) = \beta_0 + \beta_1 x_0 + \beta_2 x_1$$

los estimadores de máxima verosimilitud son:

$$\widehat{p}(X_0 = i, X_1 = j) = \frac{(n_{i+})\,(n_{+j})}{(n_{++})\,(n_{++})}$$

Ver en el apendice sección $(A.1)$, para mayor detalle en la obtención de las formulas cerradas de los estimadores de máxima verosimilitud.

De igual manera se pueden obtener formulas directas de los estimadores $\widehat{p}$ para otros modelos. Cuando esto no es posible es necesario utilizar un algoritmo iterativo para resolver (3.3). Especificamente para modelos clásicos se tiene el algoritmo de Ajuste Iterativo Proporcional (Iterative Proportional Fitting, IPF) , este algoritmo tiene sus inicios en el año de 1940 en un artículo publicado por Deming y Stephan [19]. Existen otros métodos para obtener los estimadores a través de procesos iterativos, como el método propuesto por Bartlett en 1935 [3]. Este método fue tal vez el primero en describir como obtener los estimadores de máxima verosimilitud para un modelo que no poseé formulas cerradas para los estimadores. Otros autores han sugerido el uso de técnicas basadas en el método Newton-Raphson, como se verá en 3.3.1.

Estimación utilizando el algoritmo de Ajuste Iterativo Proporcional.

Los pasos para obtener los estimadores de máxima verosimilitud utilizando el método de Ajuste Iterativo Proporcional ó IPF se describen a continuación, posteriormente se explica cada paso con más detalle.

1. Obtener las frecuencias marginales para todos los cliques en el modelo gráfico propuesto.
2. Obtener las probabilidades iniciales para el IPF
3. Realizar el algoritmo IPF

Obtención de las frecuencias para todos los cliques Este proceso básicamente realiza un ciclo sobre todos los cliques del modelo gráfico y en cada clique recorre la tabla de contingencia acumulando la suma de todas las frecuencias para cada valor de las variables en los cliques.

$$n_{x_c} = \sum_{y:y_c = x_c} n_y$$

Una vez terminado el ciclo, las frecuencias marginales en los cliques n_{x_c}, se almacenan en un vector para que sean utilizadas despues por el proceso de IPF.

Obtención de las probabilidades iniciales El proceso para obtener las probabilidades iniciales para el IPF, consiste en obtener las probabilidades marginales de la siguiente forma,

$$\widehat{p}^{(0)}\,(X_0 = x_0, .., X_m = x_m) = \prod_{j=0}^{m} \frac{n_{x_j}}{n_{++..+}}$$

Para un ejemplo donde se tienen tres variables X_0, X_1 y X_2 para cuando todas valen 0, suponiendo independencia la formula sería,

$$\widehat{p}^{(0)}\,(X_0 = 0, X_1 = 0, X_2 = 0) = \frac{n_{0++} n_{+0+} n_{++0}}{n^3_{+++}}$$

Algoritmo 12 *Algorito IPF propuesto por Deming y Stephan [19]*

1. $k \leftarrow 0$
2. Repite
   (a) $k \leftarrow k + 1$
   (b) Repetir para *todos los cliques c.*
      i. Para *todos los x*
      $$\widehat{p}^{(k)}\,(x) = \widehat{p}^{(k-1)}\,(x)\,\frac{n_{x_c}/n_{++..+}}{\widehat{p}^{(k-1)}\,(x_c)}$$
      Hasta *que* $\left|\widehat{p}^{(k)}\,(x) - \widehat{p}^{(k-1)}\,(x)\right| < \varepsilon$

De esta manera realizando los pasos descritos anteriormente se pueden obtener los estimadores $\widehat{p}\,(x)$ para cualquier modelo gráfico clásico.

El método IPF tiene las siguientes propiedades [7]:

1. Siempre converge, si $p > 0$.
2. Utiliza una regla de paro (ε), que permite asegurar un grado de precisión en la obtención de los estimadores.
3. El algoritmo no requiere de casos especiales de procesamiento cuando se tienen celdas sin observaciones.
4. Cuando es usado en modelos para los cuáles existen formulas cerradas el algoritmo obtiene los estimadores en una iteración.

3.2.2 Pruebas de Hipótesis sobre Modelos

Una vez que se han obtenido los estimadores $\widehat{p}(x)$ para un modelo dado M, el siguiente paso consiste en verificar que el modelo sea aceptable. Esto se logra realizando una prueba de hipótesis sobre este modelo. Para eso, se mide la distancia entre las observaciones y las frecuencias esperadas bajo un modelo dado. Las dos medidas de distancias más usadas son la prueba estadística Pearson χ^2 y la prueba estadística de proporción de log-verosimilitud G^2 [50] .

La prueba estadística Pearson χ^2 se calcula de la siguiente manera,

$$\chi^2 = \sum_x \frac{(n_{++..}\widehat{p}(x) - n_{++..}\widehat{p}_0(x))^2}{n_{++..}\widehat{p}_0(x)}$$

donde $\widehat{p}$ y $\widehat{p}_0$ son los estimadores de máxima verosimilitud de $\widehat{p}$ bajo el modelo M y M_0 : M es el modelo que se desea probar y M_0 es el modelo saturado, es decir, un modelo completo sin independencias.

La prueba estadística G de log-verosimilitud se calcula de la siguiente manera,

$$G^2 = 2\sum_x n_{++..} \log \frac{\widehat{p}(x)}{\widehat{p}_0(x)}$$

Asumiendo que $p \in P$, ambas pruebas tienen una distribución asintótica $(n_{++..+} \to \infty)$ χ^2 con grados de libertad igual a,

$$df = \dim(M_0) - \dim(M)$$

Las dimensiones $dim(M)$ de los modelos, se pueden obtener siguiendo las siguientes reglas

- Para un modelo completo M_0,

$$\dim(M_0) = 2^m - 1$$

donde m es el número de nodos en el grafo.

- Para un modelo no completo M. La suma de las dimensiones de M_c menos la dimensión de las intersecciones en los cliques, dan los grados de libertad del modelo [44]:

$$\dim(M) = \left(\sum \dim(M_c)\right) - \left(\sum \dim(M_{c\in\cap})\right)$$

En la sumatoria del primer término, la suma esta dada sobre todos los cliques del grafo. En la segunda sumatoria, la suma se da sobre todos los conjunto completos que pertencen a la intersección de 2 o más cliques.

La dimensión de un clique c en M está dada por:

$$\dim(M_c) = 2^q - 1$$

donde q es el número de nodos en el clique,

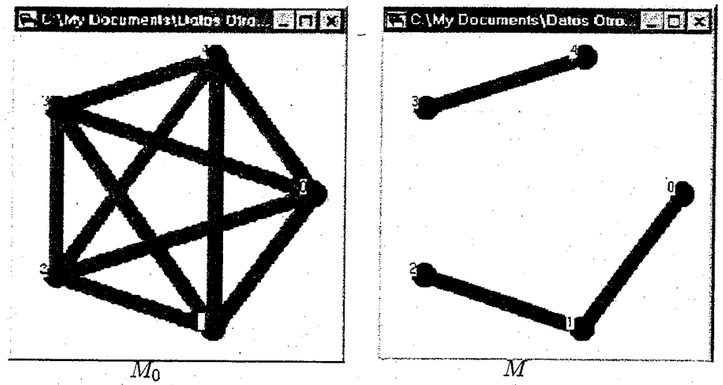
La dimensión de las intersecciones en los cliques está dada por:

$$\dim(M_{c\in\cap}) = 2^r - 1$$

donde r es el número de nodos que forman parte de la intersección entre dos o más cliques de C.

Ejemplo

Para los siguientes grafos M_0 y M:



M_0 M

$$\begin{aligned}
dim(M_0) &= 2^5 - 1 = 31 \\
dim(M) &= \dim(Clique: 0-1) + \dim(Clique: 1-2) + \dim(Clique: 3-4) \\
&\quad - \dim(Clique: 0-1 \cap Clique: 1-2) \\
dim(M) &= [3 * (2^2 - 1)] - 1 = 8
\end{aligned}$$

Así obtenemos,

$$df = 31 - 8 = 23$$

3.2.3 Obtención del Valor P de un modelo

El valor de P corresponde a la probabilidad de obtener una distancia igual o mayor a la observada entre los datos observados y el modelo M, si el modelo gráfico M es válido. De esta manera el valor de P expresa si se puede contribuir al azar (por tener una muestra pequeña) el hecho que la distancia no es cero.

Dados los estimadores de máxima verosimilitud y el valor de la prueba Pearson χ^2 o log-verosimilitud, el valor de P de un modelo gráfico dado es igual a $\left(1 - \Pr\left[\chi^2_{df} < x\right]\right)$.

A continuación se explica la implementación del calculo de $\Pr\left[\chi^2_{df} < x\right]$.

3.2.4 Aproximación del valor de $\Pr\left[\chi^2_{df} < x\right]$

La distribución chi-cuadrada es un caso especial de la distribución general Gamma y tiene un parámetro denominado "grados de libertad" o por sus siglas en inglés "degrees of freedom" (df), que determinan la forma de la distribución. Con menos grados de libertad la distribución χ^2 se tiende hacia su extremo izquierdo. Conforme los grados de libertad aumentan, la distribución tiene menos sesgo y se acerca a la simetría. Si la distribución χ^2 tiene un valor de grados de libertad grande, esta se aproxima a una distribución normal. En la figura(3.2) se muestran dos distribuciones χ^2 para diferentes grados de libertad. La media de una χ^2 esta dada por sus grados de libertad. La moda es $df - 2$ y la mediana es aproximadamente $df - 7$ [15], [32].

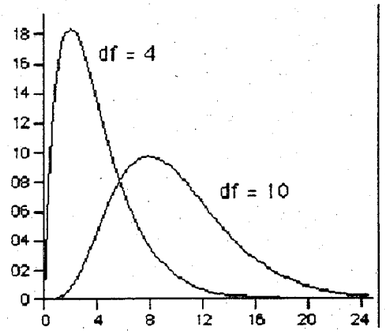


Figura 3.2: Ejemplos de distribuciones Chi-Cuadrada

Como se mencionó anteriormente si los grados de libertad (df) tienden a infinito, la distribución χ^2_{df} tiende a una distribución normal. La aproximación de la distribución χ^2_{df} está basada en este concepto utilizando la aproximación a la distribución normal como aproximación a la distribución χ^2_{df}. Entre las mejores aproximaciones se encuentra las siguientes.

Aproximación de Fisher [26],

$$\Pr\left[\chi^2_{df} < x\right] \doteqdot \Phi\left(\sqrt{2x} - \sqrt{2\,(df) - 1}\right)$$

Aproximación de Wilson-Hilferty [62],

$$\Pr\left[\chi^2_{df} < x\right] \doteqdot \Phi\left(\left\{\left(\frac{x}{df}\right)^{\frac{1}{3}} - 1 + \frac{2}{9}\,(df)^{-1}\right\}\sqrt{\frac{9\,(df)}{2}}\right) \tag{3.4}$$

Esta aproximación tiene más precisión.

$\Phi(\cdot)$ denota la función de distribución cumulativa normal:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-\frac{1}{2}x^2}\,dx$$

La aproximación a la distribución normal utilizada en $\mathcal{MOGG}$ es la propuesta por Zelen y

Severo [64], este tipo de aproximaciones emplean expresiones polinomiales, y son aproximaciones con alta precisión,

$$\Phi(x) \doteqdot 1 - \frac{1}{2}\left(1 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x_4\right)^{-4}$$

con $a_1 = 0.196854$; $a_2 = 0.115194$; $a_3 = 0.000344$; $a_4 = 0.019527$, el error en $\Phi(x)$, para $x \geq 0$, es menor a 2.5×10^{-4}.

3.3 Construcción de Modelos Gráficos Generalizados

La estimación en este tipo de modelos gráficos es generalmente una tarea más complicada que en el caso de modelos gráficos clásicos, ya que en este caso existen más restricciones en los parámetros. Como no existe una extensión del IPF para este tipo de modelos, proponemos un método basado en el algoritmo de Newton-Raphson para resolver (3.3). Para eso consideramos una distribución positiva p como un vector $v(\log(p))$ definido de la siguiente forma,

$$v(\log(p)) = \begin{pmatrix} \log\left(P\left(X_0 = 0, X_1 = 0, ..., X_n = 0\right)\right) \\ \log\left(P\left(X_0 = 1, X_1 = 0, ..., X_n = 0\right)\right) \\ ... \\ \log\left(P\left(X_0 = 1, X_1 = 1, ..., X_n = 1\right)\right) \end{pmatrix}^T$$

Para una independencia de la forma

$$X_0 \perp X_1 \mid X_2 = i_2, X_3 = i_3, ..., X_n = i_n \tag{3.5}$$

definimos un vector a con 1 en las posiciones en las cuales,

$$P\left(X_0 = 1, X_1 = 1, X_2 = i_2 ..., X_n = i_n\right)$$

y

$$P\left(X_0 = 0, X_1 = 0, X_2 = i_2 ..., X_n = i_n\right)$$

aparecen en $v(\log(p))$, con -1 en las posiciones en las cuales,

$$P\left(X_0 = 0, X_1 = 1, X_2 = i_2 ..., X_n = i_n\right)$$

y

$$P\left(X_0 = 1, X_1 = 0, X_2 = i_2 ..., X_n = i_n\right)$$

aparecen en $v(\log(p))$ y 0 en las demás posiciones. De esta manera $\langle v(\log(p)), a\rangle$ corresponde al log-oddsratio. Bajo (3.5) el log-oddsratio es igual a cero o equivalente a:

$$v(\log(p)) \perp a$$

Es fácil extender lo anterior para el caso general $X_i \perp X_j \mid \{X_k = x_k, k \notin \{i, j\}\}$.

Ejemplo Para la independencia $X_0 \perp X_2 \mid X_1 = 0$, el vector de independencia a es:

$$a = (-1, 1, 0, 0, 1, -1, 0, 0)$$

De igual manera para la independencia $X_0 \perp X_2 \mid X_1 = 1, X_3 = 0$, el vector de independencia a es:

$$a = (0, 0, 0, 0, -1, 0, 1, 0, 0, 0, 0, 0, 1, 0, -1, 0).$$

3.3.1 Estimación utilizando el método de Newton-Raphson

Sabemos que la log-verosimilitud se expresa como sigue,

$$l(\beta; n) = \sum_i n_i \log p_\beta(i) \tag{3.6}$$

Sabemos que $v(\log(p)) \in Vect(a_1, a_2, ..)^\perp$. Así podemos escribir $v(\log(p))$ como una combinación lineal de un conjunto de vectores v_j.

$$v(\log(p)) = \sum_j \beta_j v_j \tag{3.7}$$

O equivalente a construir una matriz V con elementos v_j.

$$v(\log(p)) = V\beta \tag{3.8}$$

Por definición de $v(\log(p))$, (3.8) es equivalente a,

$$\widehat{p_i} = e^{(V\beta)_i} \qquad , \forall i$$

Tenemos entonces que la log-verosimilitud puede ser expresada como,

$$l(\beta; n) = \sum_i n_i (V\beta)_i \tag{3.9}$$

Para encontrar los estimadores de máxima verosimilitud se maximiza la función (3.9) en βs. Dado que $\sum p = 1$, tenemos que maximizar (3.9) bajo la restricción.

$$\sum e^{(V\beta)_i} = 1 \tag{3.10}$$

Podemos convertir este problema de optimización con restricciónes a otro sin restricciones de la siguiente manera:

Sacamos de la sumatoria en (3.10) el termino $e^{(V\beta)_i}$ cuando $i = 0$,

$$e^{(V\beta)_0} \left(\sum_{i \geq 1} e^{(V\beta)_i} \right) = 1$$

Despejando $(V\beta)_0$ obtenemos,

$$(V\beta)_0 = -\log \sum_{i \geq 1} e^{(V\beta)_i} \tag{3.11}$$

De esta manera la ecuación de log-verosimilitud (3.9) se puede descomponer en dos terminos de la siguiente manera,

$$l(\beta; n) = \sum_{i \geq 1} n_i (V\beta)_i + n_0 (V\beta)_0$$

Sustituyendo (3.11) en la ecuación anterior obtenemos una función de log-verosimilitud donde el parámetro β_0 no entra dentro de la maximización de la función.

$$l(\beta; n) = \sum_{i \geq 1} n_i (V\beta)_i - n_0 \log \sum_{i \geq 1} e^{(V\beta)_i}$$

De esta manera los pasos para obtener los estimadores de máxima verosimilitud utilizando el método propuesto basado en el algoritmo de Newton-Raphson son los siguientes:

1. Obtener el conjunto de vectores a ortogonal a $v(\log(p))$.
2. Obtener una base de vectores ortogonales V, para el espacio ortogonal al conjunto de vectores a.
3. Maximizar (3.9) utilizando el método Newton-Raphson.
4. Realizar una prueba Pearson χ^2 para los estimadores $\widehat{p}(x)$ obtenidos.
5. Obtener el valor de P.

3.3.2 Estimación utilizando el método IPF & Newton-Raphson

Para acelerar la convergencia del algoritmo de Newton Raphson se propone una variante del método anterior. El método propuesto consiste en elegir de una manera particular los valores β de inicialización del método Newton Raphson. Para realizar esto buscamos el modelo gráfico clásico con el mínimo número de aristas, tal que cada independencia está contenida en el modelo gráfico generalizado que se quiere estimar, y calculamos los estimadores de máxima verosimilitud $\widehat{p}^*$ para este modelo. Utilizando mínimos cuadrados buscamos un β^* tal que se minimiza,

$$\|\log \widehat{p}^* - V\beta^*\|^2$$

finalmente este β^* se utiliza para inicializar el método Newton Raphson.

Ejemplo Para el modelo gráfico generalizado de la figura (3.3), se deben encontrar primero los estimadores $\widehat{p}^*$ del modelo gráfico clásico de la figura (3.4) y luego obtener los β^* que mejoren la convergencia del método Newton-Raphson.

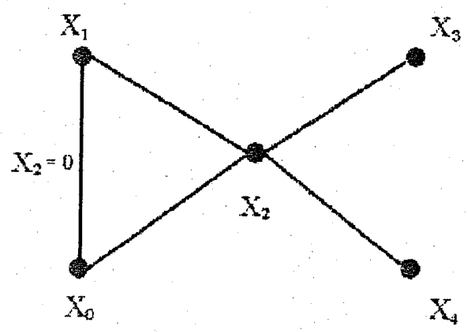


Figura 3.3: Modelo gráfico generalizado con β^* como inicialización para el Newton-Raphson.

3.4 Comparación entre los métodos de estimación

A continuación se muestra una tabla donde se realiza una comparación entre los dos métodos propuestos en este trabajo de tesis para la estimación de modelos gráficos generalizados. Las pruebas se realizaron en una PC Pentium III a 800MHz con 64MB en memoria RAM . La tabla muestra el tiempo en segundos (s) que les toma a los métodos Newton Raphson y Newton Raphson inicializado con IPF en converger y el número de iteraciones para cada método respectivamente (iter). Las pruebas se realizaron sobre diferentes modelos y de diferentes dimensiones.

Modelo con 6 variables:

Métodos	1 independencia	3 independencias	5 independencias
NR	.270s, 11 iter.	.270s, 11 iter.	.220s, 11 iter.
IPF & NR	.270s, 2 y 11 iter.	.270s, 1 y 11iter.	.220s, 2 y 11iter

Modelo con 7 variables:

Métodos	1 independencia	3 independencias	5 independencias
NR	3s, 18 iter.	14s, 77 iter.	4s, 24 iter.
IPF & NR	9s, 2 y 48 iter.	8s, 2 y 48 iter.	7s, 2 y 40 iter.

Modelo con 8 variables:

Métodos	1 independencia	3 independencias	5 independencias
NR	45s, 18 iter.	35s, 17 iter.	54s, 28 iter.
IPF & NR	63s, 2 y 23 iter.	31s, 2 y 16iter.	45s, 2 y 22 iter.

Se observa que en ocasiones hay una mejora en el tiempo de procesamiento tomado por el Newton-Raphson inicializado con IPF sobre el Newton-Raphson sin inicialización, pero no se encontró algún patrón que determinara cuando se debe utilizar la inicialización con IPF y cuando no.

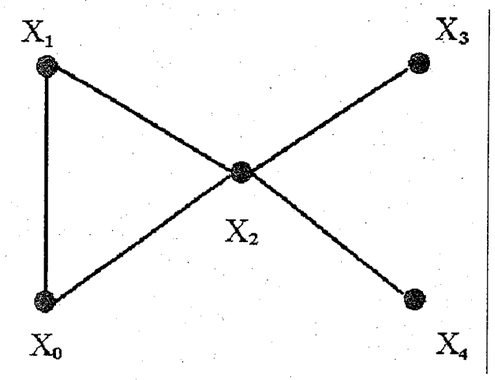


Figura 3.4: Modelo gráfico clásico para la estimación de los β^*

3.5 Selección de Modelos Gráficos

3.5.1 Selección en Modelos Gráficos Clásicos

Una estrategia para encontrar el mejor modelo gráfico, podría ser generar todos los modelos posibles y calcular sus valores P respectivos, y dentro de los modelos que son aceptados bajo un cierto nivel de significancia elegir el que parezca más adecuado y que tenga mayor simplicidad de parámetros. Esta técnica es muy ineficiente y además el número de posibles modelos crece exponencialmente de acuerdo al número de variables.

Por ejemplo para una tabla de dimensión 3, se pueden tener los 8 posibles modelos de independencia siguientes.

$$
\begin{aligned}
\log P(X_0, X_1, X_2) &= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_0 X_1 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_5 X_0 X_2 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_6 X_1 X_2 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_0 X_1 + \beta_5 X_0 X_2 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_0 X_1 + \beta_6 X_1 X_2 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_5 X_0 X_2 + \beta_6 X_1 X_2 \\
&= \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_2 + \beta_5 X_0 X_2 + \beta_6 X_0 X_2 \\
&\quad + \beta_7 X_1 X_2 + \beta_8 X_0 X_1 X_2
\end{aligned}
$$

Para una tabla de dimensión 4, hay 64 posibles modelos y para una tabla de dimensión 10, el número de modelos es de 35,184,372,088,832, en general para n dimensiones se pueden tener $2^{\binom{n}{2}}$ diferentes modelos.

A continuación se muestra un método de selección de modelos gráficos clásicos.

Método de Etiquetado

Para cada arista (i, j) del modelo dado, se calcula el valor de P del modelo incluyendo la independencia condicional entre la variable i y j y se dibuja la arista con un ancho inversamente proporcional a este valor. En base a esto el usuario decide cuales aristas quitar para formular un nuevo modelo. Este proceso se repite hasta no encontrar un modelo más sencillo y que sea aceptable bajo un nivel de significancia especificado por el usuario.

Ejemplo 1 El etiquetado clásico correspondiente para el conjunto de datos de la figura (3.1) se muestra en la figura (3.5)

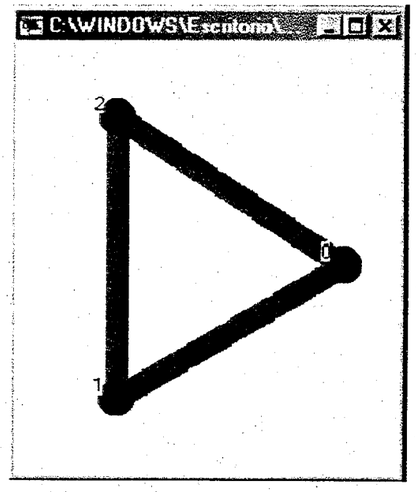


Figura 3.5: Label del modelo gráfico clásico para los datos de los jugadores de futbol americano

Los valores P para cada arista de la figura (3.5) son los siguientes,
Valores P
Arista: 0-1 = 0.00023
Arista: 0-2 = 4.76837E-7
Arista: 1-2 = 0.00022
Como se puede ver ninguna arista es factible para salir del modelo, ya que todas tienen valor P menor al umbral de significancia propuesto de 0.05.

Ejemplo 2 Aplicando el método de etiquetado al conjunto de datos de la tabla de contingencia mostrada en la figura (1.1), correspondiente a los datos de alumnos y las materias que pasaron, obtenemos el modelo gráfico de la figura (3.6).

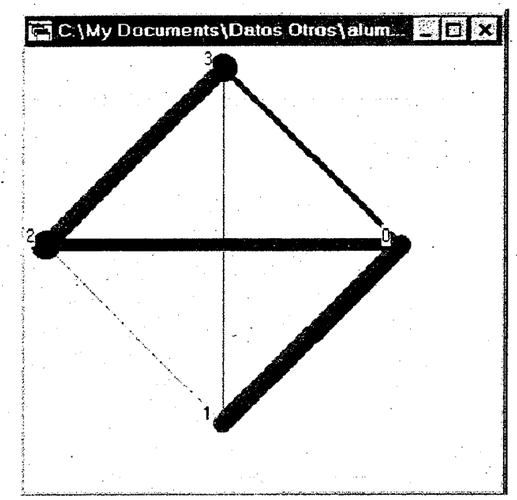


Figura 3.6: Label del modelo gráfico clásico para los datos de los alumnos y las materias

Los valores P para cada arista de la figura (3.6) son los siguientes,
Valores P
Arista: 0-1 = 0.00074
Arista: 0-2 = 0.30733
Arista: 0-3 = 0.77155
Arista: 1-2 = 0.97150
Arista: 1-3 = 0.93364
Arista: 2-3 = 0.08461
En este modelo existen 4 aristas que son candidatas para salir del modelo. Si quitamos siempre la arista que tenga el valor de P más grande y repetimos este proceso de selección, obtenemos el método llamado Backwise.

3.5.2 Selección en Modelos Gráficos Generalizados

En los modelos gráficos generalizados implementamos de la misma manera el método de etiquetado. Si el valor de P es mayor al nivel de significancia especificado por el usuario el color del segmento es de color gris. Una estrategia propuesta es utilizar el método Backwise para obtener el modelo gráfico clásico más simple y despues aplicar el método de etiquetado generalizado para seleccionar ahora los segmentos que pueden salir del modelo.

Ejemplo del método de etiquetado

En la figura (3.7) se puede observar el resultado obtenido al aplicar el método de etiquetado al modelo saturado del conjunto de datos de los jugadores de futbol americano de la tabla

de contingencia de la figura (3.1). En estos modelos se utilizo como medida de asociación la correlación entre variables.

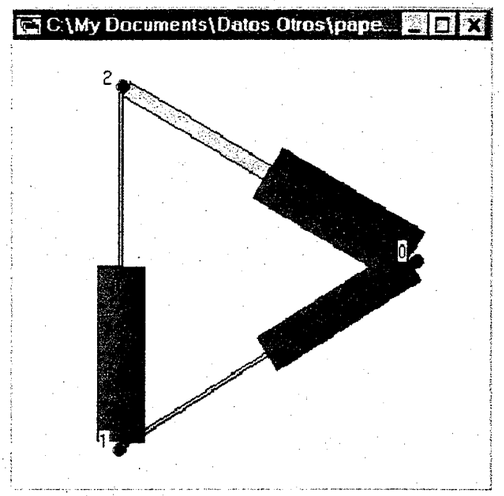


Figura 3.7: Label generalizado del modelo gráfico generalizado para los datos de los jugadores de futbol americano

Los valores P obtenidos en $\mathcal{MOGG}$ para cada cuadro de las aristas son los siguientes
Valores P
0-1 (0): 9.19699E-5
0-1 (1): 0.76731
0-2 (0): 4.17232E-7
0-2 (1): 0.3235
1-2 (0): 9.23871E-5
1-2 (1): 0.78788

Ejemplo de selección de un modelo

En la figura (3.8) se muestra el modelo gráfico generalizado con etiquetas para el conjunto de datos de los accidentes en futbol americano. En [38] se muestra que este modelo gráfico generalizado es aceptable a un nivel de significancia de 0.1.

En la figura (3.9) se muestra el mismo modelo gráfico generalizado obtenido en $\mathcal{MOGG}$, el número de iteraciones newton 8, el valor de la prueba Pearson χ^2 es de 3.74555 y el valor p de 0.28946. El valor P rebasa el umbral de significancia de 0.1, por lo tanto el modelo es válido o aceptado con 10% de significancia.

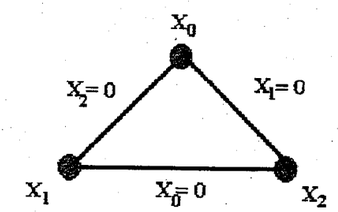


Figura 3.8: Modelo Grafico Generalizado para los datos de accidentes de futbol americano

3.6 Historia de la estimación y busqueda de modelos.

La historia de los métodos de selección o busqueda de modelos se remonta a finales de los años setenta con el trabajo de Goodman en 1971 [31], donde se propone un procedimiento similar al Stepwise (paso a paso) de regresión multiple, que permite agregar y eliminar términos de un modelo. Bishop, Frienberg y Holland en 1975 [7], dan una explicación teórica y práctica de métodos de análisis multivariado discreto, métodos para estimación de máxima verosimilitud para tablas de contingencia completas e incompletas y métodos para selección de modelos. Algunas variantes del método de Goodman fueron sugeridas por Brown [11] y Wermuth [58] en 1976 y Benedetti y Brown [4] en 1978, ellos utilizan diferentes medidas de asociación entre variables para decidir cuál término debe entrar o salir del modelo. Whittaker y Aitkin [61] en ese mismo año proponen la utilización de un procedimiento de pruebas simultaneas para establecer un conjunto de modelos jerárquicos. Edwards y Kreiner en 1983 [22] sugirieron la utilización de solo un conjunto de modelos jerárquicos que permiten representar por medio de un grafo la estructura de asociación entre las variables. El procedimiento de selección de modelos compara modelos obtenidos al agregar y eliminar aristas del grafo.

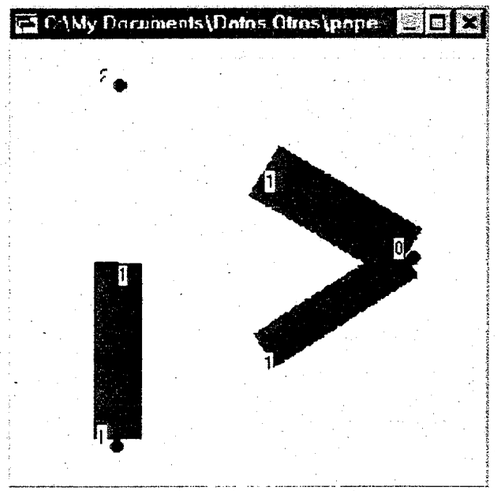


Figura 3.9: Modelo Grafico Generalizado para los datos de accidentes en futbol

Capítulo 4

Aplicaciones

4.1 Modelación de accesos a páginas WWW

4.1.1 Introducción

El rápido crecimiento de la red Internet y el aumento gradual de usuarios que accesan a este servicio dan lugar a dos necesidades o retos en este tipo de comunicación:

1. Mejorar el servicio.- Como hacer el acceso más rapido, robusto y eficiente a hojas Web, personalizar el sitio Web, agilizar las ventas o tramites usando este servicio, etc.
2. Obtener información sobre el usuario.- Como usar la información de los accesos realizados por los usuarios para comprender el comportamiento de los usuarios. ¿Qué podemos aprender de los usuarios y sus intereses?

Muchos métodos en la actualidad se enfocan al primero. Nosotros nos enfocaremos a proponer una solución para el segundo.

Para obtener algún tipo de información, que nos de una idea de la interacción entre el usuario y el Web, las páginas por las que ha navegado, las compras que ha realizado, el tiempo que ha permanecido en ciertas hojas, si son usuarios frecuentes, etc. La mayoría de los métodos que existen en la actualidad encuentran patrones, grupos o cadenas de navegación que permiten contestar las preguntas 1-5 descritas al inicio capitulo 1. De esta manera se propone la utilización de los modelos gráficos para contestar la pregunta número 6 y de esta manera contribuir a un buen análisis de los accesos realizados a páginas Web.

Existen diferentes formas de obtener información sobre los accesos realizados por los usuarios. Una fuente importante de información y que comunmente se utiliza es la proporcionada por los archivos log que se almacenan en el servidor de hojas Web. Estos archivos registran cada día los accesos realizados a determinadas hojas Web dentro del sitio, que máquina realizó el acceso, el día, hora y protocolo utilizado. Un ejemplo de un registro típico almacenado en un archivo log es como el que se muestra a continuación.

148.233.95.58–[01/Jan/2001:10:31:23-0600]
"GET/info_general/computacion.html HTTP/1.1"200 8679
"http://www.cimat.mx/estudios/programas/maestria/computo.html"

El registro corresponde a una visita a una hoja Web del servidor fractal en el CIMAT. La primera serie de 4 numeros corresponde a la dirección de la máquina desde donde se esta realizando la requisición. Entre corchetes se encuentra la fecha y la hora en la que se realiza el proceso. El número siguiente que en este caso es 0600 corresponde al protocolo de requisición. Entre comillas se encuentra la página requerida por el usuario. en seguida se muestra el protocolo de transición, que en este caso es el protocolo http/1.1. El código 200 indica al servidor el estado del proceso. El número 8679 indica el tamaño de la página requerida.

Un archivo log extendido además de registrar la información detallada en el párrafo anterior, registra desde que hoja se realizó el acceso. Este tipo de registros sirven para poder inferir la navegación de los usuarios y así obtener las cadenas de hojas accesadas. Otra forma de obtener información es la que es proporcionada en línea por los usuarios, las transacciones que estos realizan dentro de una hoja Web y los registros en línea que algunos sitios Web solicitan como un ID y un password.

El análisis de accesos a hojas de internet en un servidor Web es entonces de gran importancia en la actualidad, ya que un buen análisis proporcionará tanto al usuario, como al administrador de la hoja Web una idea de lo que sucede realmente en el sitio y la forma en que los usuarios interactúan en él.

A continuación se hace una breve descripción de las herramientas más comunes y las técnicas que utilizan.

4.1.2 Métodos utilizados frecuentemente

Existen bastantes métodos, que obtienen datos estadísticos a partir de los archivos log, algunos métodos proporcionan estadísticas de los accesos, es decir, quienes los realizaron , desde que maquinas, tiempos transcurridos, días con mayor número o menor número de accesos. Los siguientes programas utilizan este tipo de métodos: Webalizer, Access Watch y Http-Analize, todos estos utilizan Java scripts, Perl scripts o shell scripts. Muchas herramientas de este tipo se utilizan para el análisis de tráfico en Internet y producen reportes periódicos que ayudan a mejorar el sistema, la seguridad y facilitar las modificaciones al sitio, así como también para dar soporte a toma de decisiones. Una desventaja de estas herramientas es que no encuentran patrones muy sofisticados. Existen algunos trabajos de investigación interesantes, como el realizado por Padmanabhan y Mogul [48], los cuales describen una técnica para realizar el mantenimiento de un sitio Web por medio de su servidor Web con información estadística mostrando la inter-dependencia entre hojas Web en la forma de un grafo dirigido: el más probable estadísticamente es el que debe ser enviado a la cola de envíos del cliente. Eick, Nelson y Schmidt [23] describen un sistema que utiliza el sistema de visualización SeeSoft. Este sistema es una herramienta visual para analizar los archivos log y ayuda a visualizar rápidamente los problemas generales que pudiera tener el sitio Web.

Existen otros métodos que son utilizados comúnmente y que caen dentro del área de minería de datos del Web, como son: las reglas de asociación, los algoritmos de agrupamiento y los patrones de secuencias.

Minería de datos en el Web

En la minería de datos en el Web comúnmente se utiliza el archivo log, pero también se puede realizar minería de datos sobre el contenido y la estructura del sitio Web.

La mineria de Datos en el Web no solo ayuda a contestar las preguntas: ¿Qué fue lo que el visitante realizó en el sitio?, ¿Cuál fue su comportamiento?, también sirve para contestar preguntas del tipo ¿Cuales son los gustos de los visitantes?, ¿Cómo reaccionarán los visitantes, si realizo estos cambios en el sitio?, ¿Qué tipo de visitante es el que actualmente esta accesando el sitio?, etc.

Para realizar procesos de minería de datos en el Web se requiere principalmente de tres etapas: Preprocesamiento, Descubrimiento de Patrones y Análisis de Patrones como se muestra en la figura (4.1).

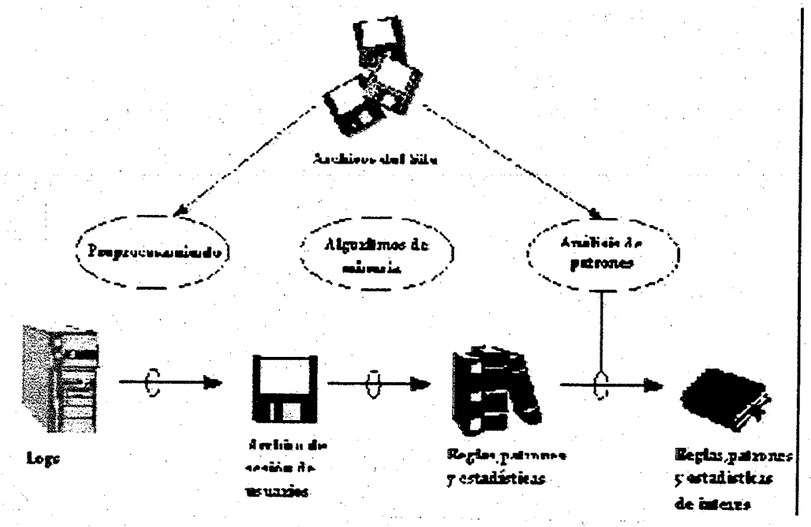


Figura 4.1: Proceso de Minería de Datos en el Web

Preprocesamiento Esta etapa consiste en convertir la información grabada en los archivos logs, el contenido de las páginas Web y la información correspondiente a la estructura del Web sitio. Generalmente el procesar el archivo log es el proceso más complicado debido a que no es evidente la manera de asociar los accesos de ciertas maquinas con un usuario en particular debido a que se pueden tener los siguientes tipos de accesos:

- Una sola dirección IP y muchas sesiones de servidor.- En ocasiones los servicios de Internet tienen un servidor proxy. Los usuario accesan el Web a través de este servidor proxy, el cuál tomará en ocasiones una misma dirección para realizar diferentes solicitudes de páginas, simulando entonces que un solo usuario realiza muchas solicitudes de páginas. Esto genera problemas para reconocer diferentes usuarios con un mismo IP.
- Muchas direcciones IP y una sola sesión en el servidor.- existen también tipos de servidores que cambian constantemente la dirección IP que asignan a una sesión, debido al tráfico

y para mejorar en general los tiempos de accesos, esto se traduce también en que no se pueda determinar que accesos realizó el mismo usuario.

- Utilización del "Cache".- La mayoría de los navegadores utilizan el "cache" o un espacio en disco duro donde se almacenan las hojas web visitadas ultimamente. Esto impide que se puedan obtener las cadenas de navegación con facilidad, ya que no sé puede saber con certeza si el usuario regreso a una página anteriormente accesada.

Estos problemas complican la obtención de secuencias de navegación restringiendo los análisis a considerar solamente las visitas al Web sitio con un determinado lapso de tiempo, las hojas Web que han sido accesadas, la estructura del Web sitio y el contenido de este.

Algoritmos de minería En esta etapa, se utilizan todo tipo de algoritmos para encontrar patrones interesantes a partir de los datos preprocesados. Los algoritmos que más son utilizados son las reglas de asociación, algoritmos de agrupamiento y patrones de secuencias. A continuación se muestran algunos métodos y sistemas que utilizan varias técnicas para descubrir conocimiento a partir de datos del Web. En seguida se muestran las técnicas que solo utilizan Reglas de Asociación. Después se muestran las que solo utilizan agrupamiento y finalmente en se muestran las que utilizan Patrones de Secuencias.

Karuna, Joshi, Yesha y Krishnapuram [40], describen la creación de una warehouse usando archivos logs y una *OLAP* (On-Line Analytical Processing) relacional. A esta warehouse se le introducen los datos de los registros en los archivos log, así como los resultados obtenidos al realizar la minería de los datos de este archivo log. La etapa de minería de datos de los archivos log utiliza dos técnicas de minería de datos: agrupamiento y generación de reglas de asociación. Para descubrir reglas de asociación, se utilizó el software *SGI´s Mineset*, que implementa una variación del algoritmo apriori. Para la técnica de agrupamiento se usó el algoritmo *Fuzzy C-medoids*.

Silani [53] muestra la implementación de procesos de e-intelligence en un Web sitio utilizando los procesos típicos de planeación del servidor Web, conducta de navegación, perfil de los visitantes y predicción de compras. Utilizan técnicas de minería de datos como reglas de asociación, agrupamiento y redes neuronales para clasificar y obtener patrones de los datos del archivo log y del análisis de las secuencias de navegación.

Wu, Yu y Ballman [63] muestran la herramienta *SpeedTracer* desarrollada para entender las conductas de navegación de los usuarios utilizando técnicas de minería de datos en los archivos log. *SpeedTracer* primero identifica las sesiones realizadas por los usuarios haciendo una reconstrucción de las rutas que realizaron. Después se aplican los algoritmos de minería para descubrir las rutas más comunes y los grupos de páginas que se visitaron conjuntamente con mayor frecuencia.

Reglas de Asociación Las reglas de asociación se utilizan comúnmente para relacionar las páginas de un Web sitio. Las reglas de asociación pueden revelar las correlaciones entre páginas y ayudar a los diseñadores del Web sitio a reestructurarlo. También pueden servir como una heurística para reducir la latencia percibida por el usuario cuando carga una página, ya que se puede enviar anticipadamente la hoja Web que más probablemente vaya a accesar el usuario, mientras este revisa el contenido de la página actual dentro de su visita.

Algunos trabajos de investigación realizados en esta línea son los propuestos por Cooley, Tan y Srivastava [16], los cuales desean encontrar el "grado de interés" en el contexto del uso de la minería en el Web. Ellos proponen la utilización de un conjunto de proposiciones iniciales sobre el uso del Web sitio. Utilizando un modelo cuantitativo propuesto basado en soporte lógico, determinan el grado de interés de cada patrón. Una vez establecido el modelo y los grados de interés utilizan la estructura y los accesos realizados a las hojas Web, para obtener la evidencia a favor o en contra de una proposición. El modelo cuantitativo esta integrado en el sistema *WebSift* (Web sitio Information Filter), el cual comprende varias herramientas para el preprocesamiento de datos, como: descubrimiento de patrones y análisis de patrones. Consideran dos tipos medida del grado de interés.- objetiva y subjetiva. La objetiva mide la relación de reglas basado en los datos usados en el proceso de minería. Los umbrales en las medidas objetivas como confianza, soporte y chi-cuadrada reducen el número de reglas generadas, pero frecuentemente caen fuera del objetivo de solo reportar reglas que son de interés potencial al analista. Las medidas subjetivas del grado de interés definen dos criterios para evaluar reglas y patrones. Una regla es inesperada si sorprende al analista de datos, y accionable si el analista puede tomar ventaja al actuar con ella. La implementación actual incluye el filtro de información, una herramienta de grafos para reglas de asociación y su visualización, y un sistema de busquedas a través de SQL. El sistema *WebSift* ha sido implementado usando una base relacional SQL y Java.

Lan, Bressan y Ooi [42] describen una estrategia para reducir el tráfico en el Web determinando que hoja Web es más probable a ser accesada por un usuario, para que sea mandada antes de que esta sea solicitada. Esto lo obtienen realizando un análisis de los patrones encontrados en los archivos log y utilizan las reglas de asociación para realizar una efectiva selección de que hojas Web colocar en la cola de envíos. Una vez que una regla de la forma "Hoja1⇒Hoja2" ha sido identificada y seleccionada, el servidor Web decide enviar la Hoja2, si la Hoja1 ha sido solicitada. La medida de viabilidad de dichas reglas esta dada por la confianza de la regla. De esta manera el usuario encontrará la hoja Web disponible en su cache en cuanto la solicite, reduciendo asi la latencia y tráfico en el Web. Los resultados en los experimentos realizados con esta estrategia muestran que el uso de la arquitectura con proxy mejora los resultados y que el uso de cache es una mejora. Sin embargo la implementación de esta estrategia no se puede tener actualmente con los navegadores comerciales existentes y servidores proxy dada la definición e implementación de HTTP.

Agrupamiento Las técnicas de agrupamiento dentro de la minería de datos en el Web se utilizan para agrupar el conjunto de visitas que se han realizado a un Web sitio en grupos que muestren características similares. Existen dos tipos de clusters interesantes a descubrir: clusters de usuarios y clusters de páginas. Los clusters de usuarios establecen grupos de usuarios que exhiben patrones de navegación similares. Este tipo de conocimiento es útil para inferir el comportamiento de usuarios futuros, ya sea para personalizar el Web sitio o para dividir el mercado alcanzable para ese usuario. Los clusters de páginas establecen grupos de páginas con contenido relacionado, esta información puede ser útil para facilitar los procesos de busqueda realizados por las maquinas de busqueda en Internet. De igual manera estos tipos de clusters pueden utilizarse conjuntamente para mostrar hipervículos relacionados a un determinado usuario de acuerdo a las busquedas que este ha realizado y la navegación que ha realizado dentro del Web sitio.

Algunos métodos que utilizan la técnica de agrupamiento para datos de Internet son los propuestos por Nasraoui, Frigui, Joshi y Krishnapuram [47], quienes describen un método para realizar la personalización del Web utilizando los access logs. En este artículo ellos definen una nueva medida de distancia para realizar el agrupamiento entre sesiones en el Web, capturando también la organización del Web sitio. Proponen además una extensión del algoritmo de agrupamiento de aglomeración competitiva para conjuntos de datos relacionales. Esta extensión da lugar a un nuevo algoritmo denominado *CARD* (Competitive Agglomeration for Relational Data), el cual puede trabajar con medidas de distancia o similitud complejas y subjetivas.

Perkowitz y Etzioni [49] definen un sitio Web adaptable como aquellos sitios que automaticamente mejoran su organización y presentación de acuerdo a lo que van aprendiendo de los patrones de acceso de los visitantes. En su artículo muestran el problema de la sinstesís de indexación de páginas, que consiste en la creación automatica de páginas de navegación que contengan un conjunto de ligas a páginas sobre un tema en particular. Proponen la utilización de un algoritmo denominado *PageGather* que básicamente sirve para descubrir ligas candidatas que formarán la base para los nuevos índices de páginas, todo esto a partir del archivo log de un servidor. PageGather utiliza técnicas de estadística para descubrir ligas o hipervínculos candidatos, pero estos candidatos no corresponden a conceptos intuitivos. Por eso en este artículo también se introduce el nuevo concepto de *conceptual cluster mining*, que consiste en buscar un número pequeño de clusters que correspondan a conceptos establecidos. Se presenta también en este artículo el algoritmo *SCML*, que combina el algoritmo estadístico de agrupamiento con el algoritmo de aprendizaje de conceptos.

Fu, Sandhu y Shih [30] describen un método para organizan los registros de accesos en sesiones que representan episodios de interacción entre usuarios Web y el servidor Web. Generalizan las sesiones de acuerdo a la jerarquía de las páginas usando inducción orientada a atributos para reducir la dimension de los datos. Las sesiones generalizadas son clasificadas usando un metodo de agrupamiento jerárquico.

Estivill-Castro y Yang [25] muestran la utilización de algoritmos de clustering para agrupar visitantes en el Web. Muestran que estos algoritmos son robustos y que convergen rápidamente. En su artículo muestran la utilización de una medida de similitud de uso. Siendo $U_{u_i}[j]$ la entrada $j - ésima$ del vector binario U_{u_i}, si la página p_j ha sido accesada por el usuario u_i,se coloca un 1 y cero en otro caso. La medida de similitud del uso esta definida como el coseno del ángulo entre los vectores,

$$Uso(u_i, u_j) = \frac{U_{u_i}^T U_{u_j}}{\|U_{u_i}\| \, \|U_{u_j}\|}$$

Patrones de Secuencias Esta técnica descubre patrones en las secuencias de hojas accesadas. Dada una serie de visitas se obtienen las secuencias más representativas, que representan los patrones de navegación comunes dentro del sitio Web. Con este tipo de información los administradores del Web pueden predecir los patrones de visita futuros para colocar propaganda dirigida a un cierto tipo de clientes. Los patrones de secuencias regularmente se representan usando grafos dirigidos cíclicos y tienen la ventaja de que se pueden interpretar facilmente, pero como se mencióno anteriormente tienen la desventaja de que son difíciles de obtener.

Algunos métodos que utilizan esta técnica son los propuestos por, Buchner, Baumgarten, Anand. Mulvenna y Hughes [12], quienes proponen un mecanismo de minería de secuencias

para el descubrimiento de patrones de navegación denominado *MiDAS* (Mining Internet Data for Associative Sequences). Este sistema ayuda a establecer la topología del Web sitio, en la construcción de jerarquías de conceptos sobre los datos que son minados y en la formulación de plantillas de queries que guían el proceso de minería realizado por *MiDAS*. El principal componente del algoritmo *MiDAS* es un árbol de patrones, que consiste en un grafo directo acíclico, donde cada nodo representan una hoja que ha sido alcanzada con éxito y los arcos representan la relación entre dos nodos. Existen dos diferentes tipos de ligas para describir las relaciones entre dos nodos, los arcos de secuencia que conectan a dos nodos que van a través de sub-secuencias (multiples visitas en el Web sitio), y arcos de tupla, los cuales conectan dos nodos que estan en la misma sub-secuencia (misma visita).

Spiliopoulou, Pohle y Faulstich [55] estudian la contribución de cada hoja Web en términos de eficiencia y contacto final del sitio. Los patrones de navegación de clientes y no clientes son descubiertos utilizando la herramienta de minería *WUM* (Web Utilization Miner). Los patrones de clientes seleccionados son mapeados a patrones de no clientes, para identificar diferencias y detectar páginas que son responsables de la diferente conducta entre clientes y no clientes, dichas páginas finalmente tienen que ser rediseñadas. El archivo log es filtrado para quitar los accesos de los visitantes que accesaron las páginas un corto periodo de tiempo. Posteriormente se parte el log preprocesado en un log con los compradores y un log con los no compradores. El descubrimiento de patrones se aplica en cada log por separado y de acuerdo a las mismas restricciones estadísticas y estructurales. Después se comparan los patrones de navegación descubiertos en los dos archivos log.

Los lenguajes de minería de datos de *MiDAS* y *WUM* son similares en el sentido de que usan plantillas para describir muchas características deseables de los patrones de navegación. Pero difiernen en la noción de patron de navegación. En *MiDAS* un patrón de navegación es una secuencia de eventos que satisfacen las restricciones dadas por el experto. El concepto de patrón de navegación en *WUM* incluye la secuencia de eventos que satisfacen las restricciones del experto y las rutas que conectan estos eventos. El resultado no es una secuencia sino un arbol compuesto de estas rutas. De esta manera el diseñador puede distinguir entre las rutas que son populares y las que son raras.

Borges y Levene [9] muestran un algoritmo de minería para el descubrimiento de patrones de navegación basado en gramáticas probabilísticas, donde los caminos o rutas son modelados con valores de probabilidad. Este método toma los datos del archivo log, divide los registros en sesiones y obtiene secuencias de navegación de estos. Finalmente se modelan los registros de navegación del usuario como una serie de gramáticas probabilísticas, donde las cadenas generadas con alta probabilidad corresponden a los trayectos preferidos por el usuario. En la figura (4.2) se muestra un ejemplo del tipo de análisis y gráfico que se puede obtener con esta metodología. El ejemplo muestra seis sesiones de usuario con un total de 24 páginas requeridas, donde el estado A1 fue visitado cuatro veces, dos de las cuales fue el primer estado en la sesion.

Schechter, Krishnan y Smith [52] describen un algoritmo para la creación eficiente de perfiles. Esto permite describir la conducta al requerir las hojas Web, almacenando todas las rutas en un árbol . Muestran que se puede predecir la conducta en la petición usando ruta de perfiles con alta probabilidad. De esta manera justifican la generación dinámica de contenido antes de que el cliente la requiera. Para seleccionar la ruta a ser usada para predicción, utilizan un algortimo de pasos hacia atrás, tomando el ultimo URL accesado como el más probable a predecir. Este algoritmo toma sus datos del archivo log y realiza un preprocesamiento de estos utilizando la

ID	Trail
1	$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4$
2	$A_1 \rightarrow A_5 \rightarrow A_3 \rightarrow A_4 \rightarrow A_1$
3	$A_5 \rightarrow A_2 \rightarrow A_4 \rightarrow A_6$
4	$A_5 \rightarrow A_2 \rightarrow A_3$
5	$A_5 \rightarrow A_2 \rightarrow A_3 \rightarrow A_6$
6	$A_4 \rightarrow A_1 \rightarrow A_5 \rightarrow A_3$

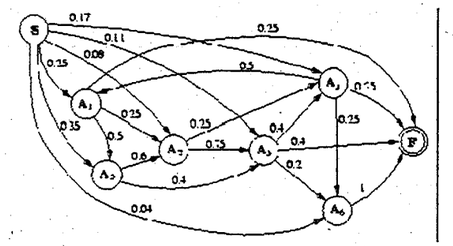


Figura 4.2: Patrones de navegación, Borges et al.

herramienta de Microsoft sitio Server Usage Analyst.

Hay algunos otros métodos que no utilizan el archivo log como parte importante en la obtención de información de los accesos en Internet.

Entre estos métodos se encuentran los propuestos por Joachims, Freitag y Mitchell [39], quienes proponen la utilización de una herramienta interactiva denominada WebWatcher. Esta herramienta toma información sobre las busquedas que el usuario realiza en el Web. Con esta información aplica técnicas de aprendizaje supervisado para encontrar los patrones de navegación relevantes.

Murray [46] muestra un método para inferir atributos demográficos a partir de la navegación de un usuario utilizando los términos que introduce el usuario en algun buscador y las hojas Web accesadas por él en Internet. Una vez que se tiene esta información se utiliza una técnica de recuperación de información llamada *Latent Semantic Analysis* (LSA) para construir un vector en el cuál se puedan representar los datos de uso del Web con cada usuario. Este vector LSA permite inferir atributos demográficos introduciendo este vector a una red neural de tres capas. La red ha sido entrenada usando el método de gradiente conjugado escalado.

Análisis de patrones Esta es la etapa final del proceso de minería de datos, y generalmente consiste en realizar un análisis detallado de los resultados obtenidos por los diferentes algoritmos de minería. En ocasiones se utilizan varios algoritmos con los mismos datos, lo que hace más complicado el proceso de análisis. El proceso de análisis debe generar también los pasos a seguir para mejorar el Web sitio, caracterizar los usuarios, los accesos, y mejorar el servicio que se proporciona al usuario. De igual manera al usuario debe darle una idea clara de como guiar su navegación hacia lo que realmente desea obtener o ver en ese Web sitio.

4.1.3 Utilización de $\mathcal{MOGG}$ para el análisis de accesos en Internet

Cada vez que un usuario accesa a una serie de páginas en un servidor se dice que ha realizado una "visita". Esta visita forma una cadena de páginas del tipo $p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow p_1$.

Definiendo los grupos de páginas que se accesan como:

$$\{G_0, G_1, ..., G_n\}$$

Se tiene entonces que cada visita se puede ver como una serie de variables binarias $(X_0, X_1, .., X_n)$, donde $X_i = 1$ si y solo sí se visitó al menos una página del grupo G_i.

Para mostrar la utilización de los modelos gráficos dentro de la modelación de usuarios de Internet, se utilizó el programa $\mathcal{MOGG}$ en conjunto con el programa Amberes (desarrollado en el CIMAT) y el "Intelligent Miner for Data" de IBM. El programa "Amberes" es un software realizado en Perl que genera datos discretos binarios de una serie de accesos de Internet almacenados en archivo log. Amberes genera también datos estadísticos sobre los tiempos de accesos, maquinas que más accesaron el Web sitio, muestra el número de cadenas o secuencias, etc.

El objetivo al utilizar estos datos dentro de $\mathcal{MOGG}$ es primordialmente encontrar relaciones y dependencias en los accesos o requisiciones de páginas que se encuentran en el servidor del CIMAT, para esto solo nos enfocamos solo a los accesos realizados por maquinas fuera de la propia institución. Se desea encontrar que páginas estan fuertemente relacionadas. Los archivos log fueron tomados del servidor "fractal" del CIMAT dentro de diferentes periodos de tiempo, obteniendo diferentes resultados. A continuación se muestran tres ejemplos representativos.

Ejemplo 1

El conjunto de datos que se analiza en este ejemplo corresponde a los grupos de páginas de las maestrías que se imparten en el CIMAT. Dado que los usuarios entraron a alguna hoja en http://www.cimat.mx/estudios/programas/, se desea saber las relaciones existentes entre los accesos a las hojas de las diferentes maestrías. Los grupos de páginas y variables que se modelan son:

G_0=/estudios/programas/maestria/estadistica.html
G_1=/estudios/programas/maestria/computo.html
G_2=/estudios/programas/maestria/aplicadas.html
G_3=/estudios/programas/maestria/basicas.html

Los modelos gráficos se obtuvieron tomando 5% de umbral de significancia para el valor de P

Periodo del 3 de Enero del 2000 al 25 de Agosto del 2000. Antes de analizar los datos se dará primero una descripción general.

Panorama General La figura siguiente muestra el histograma de la longitud de las cadenas.

Longidudes	frecuencia	
0-1	73283	70.21%
2-3	14829	14.20%
4-5	6680	6.40%
6-7	3522	3.37%
8-9	1908	1.82%
10-11	1191	1.14%
12-13	737	0.70%
14-15	505	0.48%
16-17	378	0.36%
18-19	261	0.25%
>20	1081	1.3%
No mostrados=	0	
Total=	104375	

Histograma de longitud de cadenas para el ejemplo 1, periodo Enero-Agosto del 2000

La figura siguiente muestra el histograma del rango de tiempos de las cadenas, dado en segundos.

Rango	frecuencia	
0-59	76313	73.11%
60-119	8120	7.77%
120-179	4941	4.73%
180-239	3569	3.41%
240-299	2895	2.77%
300-359	1872	1.79%
360-419	1323	1.26%
420-479	1055	1.1%
480-539	809	0.77%
540-599	593	0.56%
>600	2885	2.76%
No mostrados=	0	
Total=	104375	

Histograma para rangos de tiempo de las cadenas para el ejemplo 1, periodo Enero-Agosto del 2000

La figura siguiente muestra el histograma de las páginas destino más accesadas.

Páginas	frecuencia	
/	21343	17.81%
/error_Not_Modified	14596	12.18%
/error_Not_Found	13338	11.13%
/info_general	12796	10.67%
/error_Moved_Permanently	9316	7.77%
/info_general/basicas.html	6878	5.73%
/biblioteca	6472	5.40%
/talleres	5897	4.92%
/farmat	5201	4.34%
/info_general/computacion	5162	4.30%
/estudios	4625	3.86%
/personal/investigadores	4073	3.40%
/~gil/rincon	3937	3.28%
/info_general/estadisticas.html	3568	2.97%
/estudios/cursos/00/posgrado	2625	2.19%
No mostrados=	0	
Total=	119831	

Histograma de las páginas destino más accesadas para el ejemplo 1, periodo Enero-Agosto del 2000

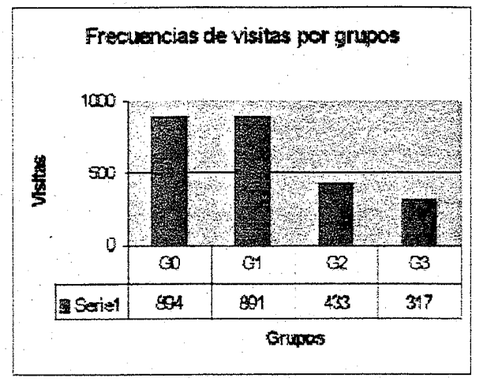
La figura siguiente muestra el histograma de las páginas desde las cuales se realizó la requisición de alguna página del CIMAT.

Análisis para grupos.

Máquinas	frecuencia	
www.cimat.mx	186256	66.58%
(desconocido)	40877	14.70%
www.altavista.com	20325	7.30%
ask.yahoo.com	5352	1.92%
search.msn.com.mx	4880	1.75%
cimat.mx	3355	1.20%
www.terra.es	3322	1.19%
search.espanol.yahoo.com	3185	1.14%
mx.search.yahoo.com	2516	0.90%
google.yahoo.com	2152	0.77%
www.google.com	1917	0.68%
espanol.yahoo.com	1310	0.47%
mx.yahoo.com	878	0.31%
www.yahoo.com	876	0.31%
search.msn.com	841	0.30%
No mostrados=	0	
Total=	279856	

Histograma de las páginas desde las cuales se realizó la requisición para el ejemplo 1, periodo Enero-Agosto del 2000

La figura siguiente muestra la frecuencia de visitas por grupos de páginas.



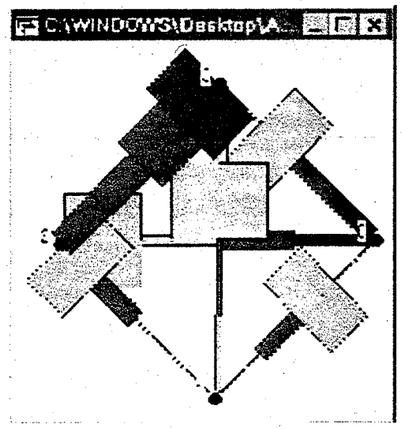
Histograma de frecuencias de visitas por grupos para el ejemplo 1, periodo Enero-Agosto del 2000

La tabla de contingencia para este periodo se muestra en la figura siguiente.

n=2439		$X_2=0$		$X_2=1$	
		$X_3=0$	$X_3=1$	$X_3=0$	$X_3=1$
$X_0=0$	$X_1=0$	1748	60	106	53
	$X_1=1$	157	3	8	3
$X_0=1$	$X_1=0$	261	2	5	13
	$X_1=1$	17	0	2	1

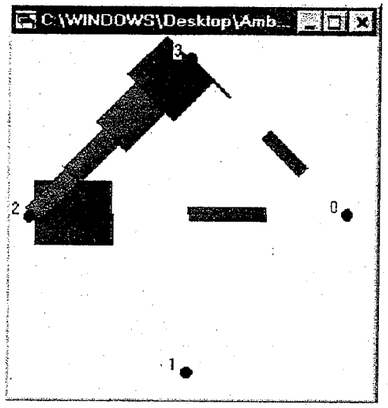
Tabla de Contingencia para el ejemplo 1, periodo Enero-Agosto del 2000

El modelo gráfico generalizado saturado se muestra en la figura siguiente.



Modelo gráfico generalizado saturado para el ejemplo 1, periodo Enero-Agosto del 2000

En la figura siguiente se muestra el modelo gráfico generalizado simplificado.



Modelo gráfico generalizado simplificado para el ejemplo 1, periodo Enero-Agosto del 2000.

El modelo gráfico tiene un valor de P de 0.0776.

En la figura siguiente se muestran los residuos normalizados obtenidos para el modelo gráfico generalizado simplificado de la figura anterior.

		$X_2=0$		$X_2=1$	
		$X_3=0$	$X_3=1$	$X_3=0$	$X_3=1$
$X_0=0$	$X_1=0$	-0.48254	1.01267	0.8067	0.19283
	$X_1=1$	0.55135	-0.69668	-0.11354	-0.66156
$X_0=1$	$X_1=0$	1.10449	-1.95168	-2.31545	0.02681
	$X_1=1$	-0.81535	-0.78528	0.79524	-0.09202

Residuos normalizados para el ejemplo 1, periodo Enero-Agosto del 2000

Dado el modelo gráfico generalizado simplificado se pueden apreciar las siguientes relaciones entre las páginas de las maestrías.-

- Computación es independiente de las demás hojas de maestría.

$$X_1 \perp [X_0, X_1, X_2]$$

- Estadística y Aplicadas son independientes cuando Básicas no está presente.

$$X_0 \perp X_2 \mid X_1 = x_1, X_3 = 0$$

- Estadística y Básicas son independientes cuando Aplicadas no está presente.

$$X_0 \perp X_3 \mid X_1 = x_1, X_2 = 0$$

- Aplicadas y Básicas son condicionalmente dependientes.

$$X_2 \not\perp X_3 \mid X_0 = x_0, X_1 = x_1$$

- Dependencia negativa entre Aplicadas y Básicas cuando Estadística y Computación están presentes y positiva en las demás.

$$\begin{aligned} \textit{Negativa } X_2 &\not\perp X_3 \mid X_0 = 1, X_1 = 1 \\ \textit{Positiva } X_2 &\not\perp X_3 \mid X_0 = 0, X_1 = x_1 \ y \\ X_2 &\not\perp X_3 \mid X_0 = 1, X_1 = 0 \end{aligned}$$

- Dependencia negativa entre Estadística y Aplicadas cuando Computación y Básicas están presentes y positiva cuando Computación no está presente y Básicas si.

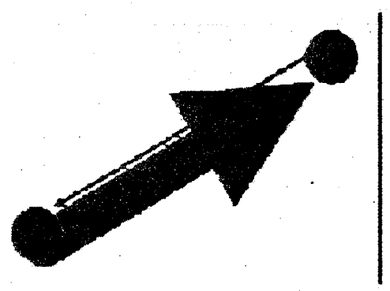
$$\begin{aligned} \textit{Negativa } X_0 &\not\perp X_2 \mid X_1 = 1, X_3 = 1 \\ \textit{Positiva } X_0 &\not\perp X_2 \mid X_1 = 0, X_3 = 1 \end{aligned}$$

- Dependencia negativa entre Estadística y Básicas cuando Computación y Aplicadas están presentes y positiva cuando Computación no está presente y Aplicadas si.

$$\begin{aligned} \textit{Negativa } X_0 &\not\perp X_3 \mid X_1 = 1, X_2 = 1 \\ \textit{Positiva } X_0 &\not\perp X_3 \mid X_1 = 0, X_2 = 1 \end{aligned}$$

Las reglas de asociación encontradas se muestran en la figura siguiente.

Soporte (%)	Confianza(%)	Elevación	Cuerpo		Cabecera
10.1302	51.85	1.88	3	→	2
10.1302	36.6500	1.88	2	→	3



Reglas de Asociación para el ejemplo 1, periodo Enero-Agosto del 2000.

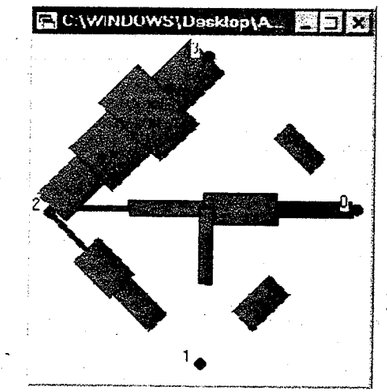
Existen dos reglas: Básicas $\Longrightarrow$ Aplicadas y Aplicadas $\Longrightarrow$ Básicas.

Periodo del 25 de Septiembre de 1999 al 31 de Diciembre del 2000. La tabla de contingencia para este periodo se muestra en la figura siguiente.

n=3365		X_2=0		X_2=1	
		X_3=0	X_3=1	X_3=0	X_3=1
X_0=0	X_1=0	2388	73	125	64
	X_1=1	217	4	14	19
X_0=1	X_1=0	394	2	7	16
	X_1=1	29	1	6	6

Tabla de contingencia para el ejemplo 1, periodo Septiembre de 1999 a Diciembre del 2000

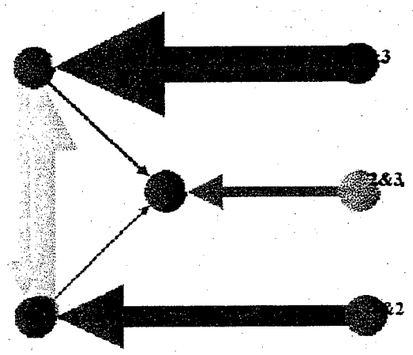
En la figura siguiente se muestra el modelo gráfico generalizado obtenido.



Modelo gráfico generalizado para el ejemplo 1, periodo Septiembre de 1999 a Diciembre del 2000

El modelo gráfico tiene un valor de P de 0.0867.
Las reglas de asociación encontradas se muestran en la figura siguiente.

Soporte (%)	Confianza(%)	Elevación	Cuerpo		Cabecera
9.0278	100.0	3.89	1&3	→	2
9.0278	76.47	3.80	1&2	→	3
15.2778	75.86	2.95	3	→	2
15.2778	59.46	2.95	2	→	3
9.0278	59.09	1.22	2&3	→	1
11.8056	45.95	0.95	2	→	1
9.0278	44.83	0.92	3	→	1



Reglas de Asociación para el ejemplo 1, periodo Septiembre de 1999 a Diciembre del 2000

Ejemplo 2

En este ejemplo se analizan los grupos de páginas con información sobre los investigadores del CIMAT. Dado que los usuarios entraron a alguna hoja en http://www.cimat.mx/personal/, se desea saber las relaciones existentes entre los accesos a las hojas de las diferentes areas de investigación. Los grupos de páginas y variables que se modelan son:

G_0=/personal/investigadores/estadistica.html
G_1=/personal/investigadores/compu.html
G_2=/personal/investigadores/basicas.html
G_3=/personal/investigadores/visitantes.html

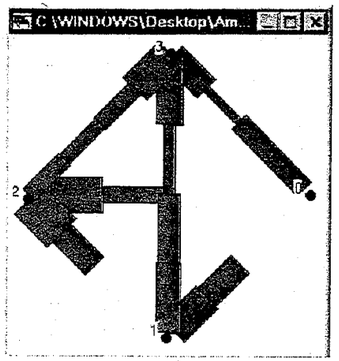
Los modelos gráficos se obtuvieron tomando 5% de umbral de significancia para el valor de P

Periodo del 25 de Septiembre de 1999 al 3 de Enero del 2000. La tabla de contingencia para este periodo se muestra en la figura siguiente.

n=1386		$X_2=0$		$X_2=1$	
		$X_3=0$	$X_3=1$	$X_3=0$	$X_3=1$
$X_1=0$	$X_0=0$	738	27	244	27
	$X_0=1$	79	15	24	12
$X_1=1$	$X_0=0$	112	16	20	9
	$X_0=1$	13	4	12	34

Tabla de Contingencia para el ejemplo 2, periodo Septiembre de 1999 a Enero del 2000

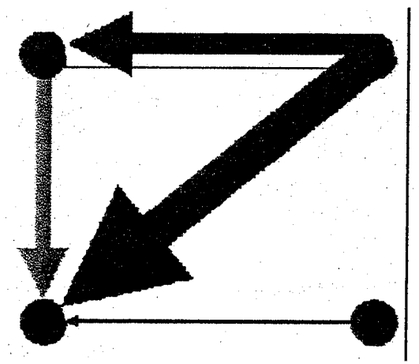
En la figura siguiente se muestra el modelo gráfico generalizado obtenido.



Modelo gráfico generalizado para el ejemplo 2, periodo Septiembre de 1999 a Enero del 2000

El modelo gráfico tiene un valor de P de 0.1905.
Las reglas de asociación encontradas se muestran en la figura siguiente.

Soporte (%)	Confianza(%)	Elevación	Cuerpo		Cabecera
10.0309	45.14	1.52	3	→	1
10.0309	33.68	1.52	1	→	3
12.6543	56.94	0.97	3	→	2
12.6543	42.49	0.72	1	→	2
11.5741	34.09	0.58	0	→	2



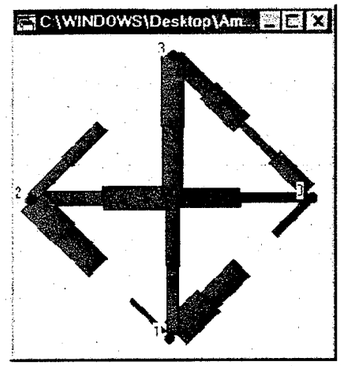
Reglas de asociación para el ejemplo 2, periodo Septiembre de 1999 a Enero del 2000

Periodo del 4 de Enero del 2000 al 25 de Agosto del 2000. La tabla de contingencia se muestra en la figura siguiente.

n=4030		$X_2=0$		$X_2=1$	
		$X_3=0$	$X_3=1$	$X_3=0$	$X_3=1$
$X_0=0$	$X_1=0$	1938	83	870	113
	$X_1=1$	231	32	69	33
$X_0=1$	$X_1=0$	325	66	64	18
	$X_1=1$	19	24	46	99

Tabla de contingencia para el ejemplo 2, periodo Enero-Agosto del 2000

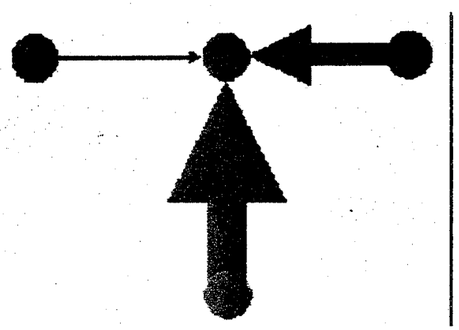
En la figura siguiente se muestra el modelo gráfico generalizado obtenido.



Modelo gráfico generalizado para el ejemplo 2, periodo Enero a Agosto del 2000

El modelo gráfico tiene un valor de P de 0.2726.
Las reglas de asociación encontradas se muestran en la figura siguiente.

Soporte (%)	Confianza(%)	Elevación	Cuerpo		Cabecera
12.5717	56.20	0.9	3	→	2
11.8069	44.67	0.71	1	→	2
10.8509	34.34	0.55	0	→	2



Reglas de asociación para el ejemplo 2, periodo Enero a Agosto del 2000

Periodo del 25 de Septiembre del 2000 al 31 de Diciembre del 2000. La tabla de contingencia para este periodo se muestra en la figura siguiente.

n=1606		$X_2=0$		$X_2=1$	
		$X_3=0$	$X_3=1$	$X_3=0$	$X_3=1$
$X_0=0$	$X_1=0$	830	32	390	46
	$X_1=1$	88	14	42	28
$X_0=1$	$X_1=0$	72	9	17	3
	$X_1=1$	6	5	6	18

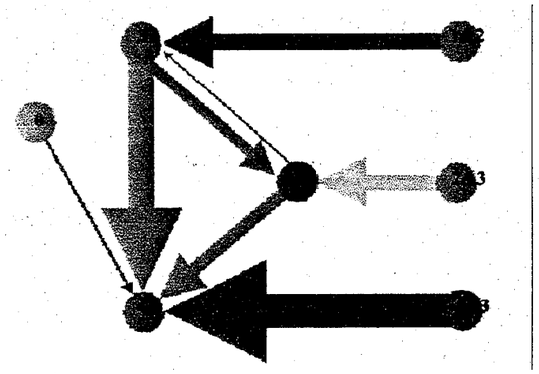
Tabla de contingencia para el ejemplo 2, periodo Septiembre-Diciembre del 2000

En la figura siguiente se muestra el modelo gráfico generalizado obtenido.

Modelo gráfico generalizado para el ejemplo 2, periodo Septiembre a Diciembre del 2000

El modelo gráfico tiene un valor de P de 0.06028.
Las reglas de asociación encontradas se muestran en la figura siguiente

Soporte (%)	Confianza(%)	Elevación	Cuerpo		Cabecera
5.9278	48.94	2.45	2&1	→	3
5.9278	48.42	1.82	3&2	→	1
8.3763	41.94	1.57	3	→	1
8.3763	31.40	1.57	1	→	3
5.9278	70.77	1.0	3&1	→	2
12.2423	61.29	0.86	3	→	2
12.1134	45.41	0.64	1	→	2
5.6701	32.35	0.46	0	→	2



Reglas de asociación para el ejemplo 2, periodo Septiembre-Diciembre del 2000

Ejemplo 3 En este ejemplo se analizan los grupos de páginas con ligas en la página de inicio del CIMAT. Dado que los usuarios entraron a alguna hoja en http://www.cimat.mx/info_general/, se desea saber las relaciones existentes entre los accesos a las hojas de las otras diferentes ligas dentro del mismo servidor. Los grupos de páginas y variables que se modelan son:

G_0=/talleres/
G_1=/rincon/
G_2=/estudios/
G_3=/famat/
G_4=/biblioteca/
G_5=/personal/investigadores/

Los modelos gráficos se obtuvieron tomando 5% de umbral de significancia para el valor de P

Periodo del 25 de Septiembre de 1999 al 3 de Enero del 2000. La tabla de contingencia para este periodo se muestra en la figura siguiente.

n=6025			$X_3=0$				$X_3=1$			
			$X_4=0$		$X_4=1$		$X_4=0$		$X_4=1$	
			$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$
$X_0=0$	$X_1=0$	$X_2=0$	3622	351	359	58	118	44	32	10
		$X_2=1$	667	97	73	9	51	13	14	4
	$X_1=1$	$X_2=0$	102	26	25	5	11	3	2	1
		$X_2=1$	46	10	9	2	7	5	1	1
$X_0=1$	$X_1=0$	$X_2=0$	122	21	10	2	4	2	1	1
		$X_2=1$	37	6	3	9	3	1	1	0
	$X_1=1$	$X_2=0$	7	2	3	0	1	1	0	0
		$X_2=1$	9	3	0	2	1	0	0	1

Tabla de contingencia para el ejemplo 3, periodo Septiembre de 1999 a Enero del 2000

En la figura siguiente se muestra el modelo gráfico generalizado obtenido.

Modelo gráfco generalizado para el ejemplo 3, periodo Septiembre de 1999 a Enero del 2000

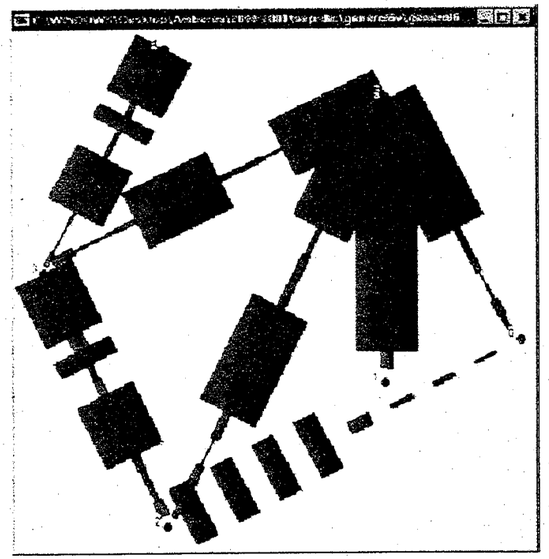
El modelo gráfico tiene un valor de P de 0.0626.
No se encontraron reglas de asociación con el mínimo de 5% de confianza y 5% de soporte.

Periodo del 25 de Septiembre del 2000 al 31 de Diciembre del 2000. La tabla de contingencia para este periodo se muestra en la figura siguiente.

n=8429			$X_3=0$				$X_3=1$			
			$X_4=0$		$X_4=1$		$X_4=0$		$X_4=1$	
			$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$
$X_0=0$	$X_1=0$	$X_2=0$	5751	485	558	57	275	73	50	9
		$X_2=1$	439	54	40	16	54	31	11	8
	$X_1=1$	$X_2=0$	5	3	0	1	0	0	0	0
		$X_2=1$	0	0	0	0	0	0	0	0
$X_0=1$	$X_1=0$	$X_2=0$	260	48	27	10	23	9	5	1
		$X_2=1$	28	16	4	5	8	7	0	6
	$X_1=1$	$X_2=0$	0	0	0	0	0	0	0	0
		$X_2=1$	0	0	0	0	0	0	0	0

Tabla de contingencia para el ejemplo 3, periodo Septiembre a Diciembre del 2000

En la figura siguiente se muestra el modelo gráfico generalizado obtenido.



Modelo gráfico generalizado del ejemplo 3, periodo Septiembre a Diciembre del 2000

El modelo gráfico tiene un valor de P de 0.1265.
No se encontraron reglas de asociación con el mínimo de 5% de confianza y 5% de soporte.

4.2 Otras Aplicaciones

A continuación se muestra un conjunto de datos categóricos obtenido de otra fuente de información. Comparamos el resultado obtenido en $\mathcal{MOGG}$ con lo que se obtiene usando un modelo log-lineal Los modelos gráficos generalizados obtenidos tratan de mostrar sus ventajas sobre los modelos gráficos clásicos. Para estos modelos gráficos se especificó un umbral de significancia de 10%.

4.2.1 Ejemplo 1

Este ejemplo muestra un conjunto de datos tomado comúnmente en la literatura de análisis de tablas de contingencia. El conjunto de datos de Mujeres y las Matematicas [27], mostrado en la tabla de contingencia de la figura siguiente, es el resultado de un estudio realizado por Lacampagne en 1979 que consistió en realizar una encuesta a 1,190 estudiantes de cuatro escuelas urbanas y cuatro suburbanas de nivel medio superior, en este estudio se deseaba conocer la actitud de los estudiantes con respecto a las matemáticas. El estudio fue enfocado a la respuesta que los estudiantes daban a la pregunta ¿Creés que necesitarás las matemáticas en tu futuro?, las variables en la tabla de contingencia y el modelo son las siguientes.

X_0.– Tipo de escuela (0=Suburbana, 1=Urbana)
X_1.– Sexo (0=Mujer, 1=Hombre)
X_2.– Asistieron a conferencias (0=Si, 1=No)

X_3.– Planes futuros (0=Universidad, 1=Trabajar)
X_4.– Preferencia de cursos (0=Matemáticas, 1=Artes liberales)
X_5.– ¿Creés que necesitarás las matemáticas en tu futuro? (0=De acuerdo, 1=En desacuerdo)

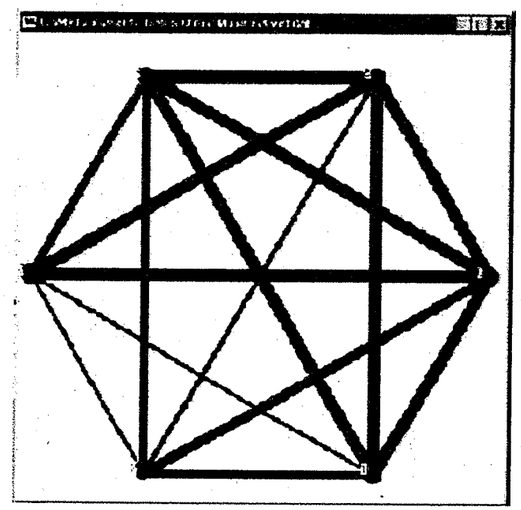
La tabla de contingencia se muestra en la figura siguiente.

n=1190			X_7=0				X_7=1			
			X_1=0		X_1=1		X_1=0		X_1=1	
			X_2=0	X_2=1	X_2=0	X_2=1	X_2=0	X_2=1	X_2=0	X_2=1
X_3=0	X_4=0	X_5=0	37	27	51	48	51	55	109	86
		X_5=1	16	11	10	19	24	28	21	25
	X_4=1	X_5=0	16	15	7	6	32	34	30	31
		X_5=1	12	24	13	7	55	39	26	19
X_3=1	X_4=0	X_5=0	10	8	12	15	2	1	9	5
		X_5=1	9	4	8	9	8	9	4	5
	X_4=1	X_5=0	7	10	7	3	5	2	1	3
		X_5=1	8	4	6	4	10	9	3	6

Tabla de contingencia para el conjunto de datos de las mujeres y las matemáticas

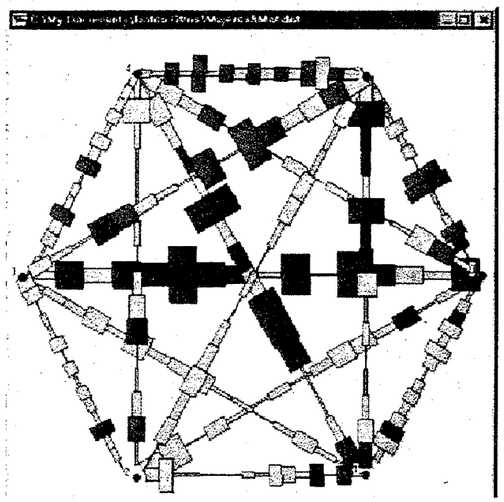
Análisis con $\mathcal{MOGG}$

La figura siguiente muestra el modelo gráfico clásico saturado.



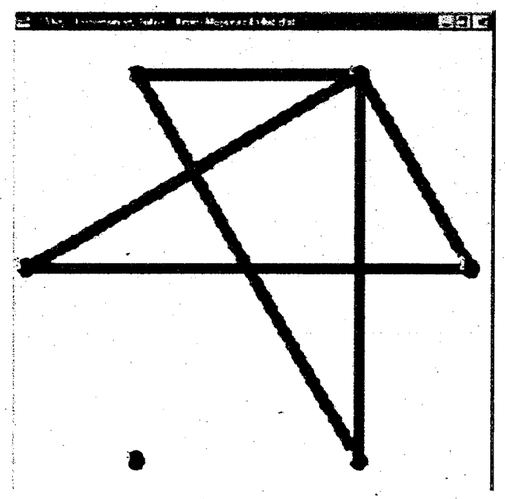
Modelo gráfico clásico saturado para el conjunto de datos de las mujeres y las matemáticas

La figura siguiente muestra el modelo gráfico generalizado saturado.



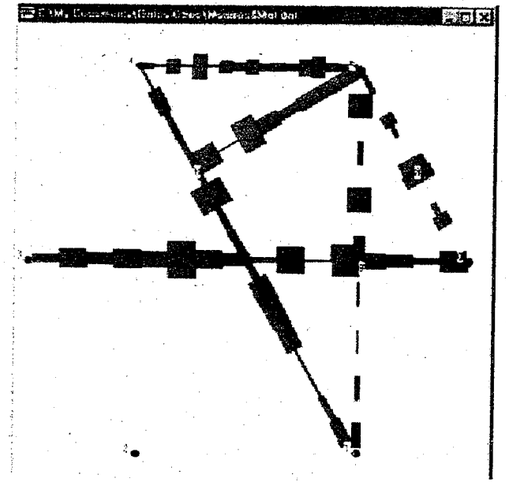
Modelo gráfico generalizado saturado para el conjunto de datos de las mujeres y las matemáticas

La figura siguiente muestra el modelo gráfico clásico simplificado.



Modelo gráfico clásico para el conjunto de datos de las mujeres y las matemáticas

El modelo gráfico tiene un valor de P de 0.1835.
La figura siguiente muestra el modelo gráfico generalizado simplificado.



Modelo gráfico generalizado para el conjunto de datos de las mujeres y las matemáticas

El modelo gráfico tiene un valor de P de 0.1228.
En la figura siguiente se muestran los residuos normalizados obtenidos para el modelo gráfico generalizado obtenido en la figura anterior.

			$X_2=0$				$X_2=1$			
			$X_4=0$		$X_4=1$		$X_4=0$		$X_4=1$	
			$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$	$X_5=0$	$X_5=1$
$X_0=0$	$X_1=0$	$X_3=0$	1.808	1.135	-0.624	-1.801	-0.212	1.141	-0.099	1.068
		$X_3=1$	0.321	-0.625	-0.541	1.271	-0.594	-0.51	1.272	-1.[illegible]
	$X_1=1$	$X_3=0$	0.394	-0.912	-1.304	0.21[illegible]	-1.563	1.238	1.228	0.567
		$X_3=1$	0.522	1.911	-1.591	-1.279	-0.553	1.93	-0.538	-0.187
$X_0=1$	$X_1=0$	$X_3=0$	-0.839	-1.00[illegible]	-1.118	2.028	-0.717	0.485	1.5[illegible]3	0.197
		$X_3=1$	0.287	0.063	-0.327	0.06	-1.333	1.12	-0.242	0.113
	$X_1=1$	$X_3=0$	0.837	-1.28	1.273	0.091	0.916	-0.393	-0.454	-1.14
		$X_3=1$	0.735	0.[illegible]25	1.995	0.039	0.454	0.313	1.29	0.31

Residuos normalizados para el ejemplo 1, datos de las mujeres y las matemáticas

Se puede ver las siguientes relaciones, entre otras.-

- El tipo de escuela y la respuesta a la pregunta son solamente independientes cuando los planes futuros son ir a la universidad.

$$X_0 \perp X_5 \mid X_1 = x_1, X_2 = x_2, X_3 = 0, X_4 = x_4$$

pero,

$$X_0 \not\perp X_5 \mid X_1 = x_1, X_2 = x_2, X_3 = 1, X_4 = x_4$$

con correlación del mismo tipo.

- El sexo y la respuesta a la pregunta son independientes cuando prefieren los cursos de artes liberales.

$$X_1 \perp X_5 \mid X_0 = x_0, X_2 = x_2, X_3 = x_3, X_4 = 1$$

pero,

$$X_1 \not\perp X_5 \mid X_0 = x_0, X_2 = x_2, X_3 = x_3, X_4 = 0$$

- Los planes futuros y la respuesta a la pregunta de la encuesta son independientes cuando el tipo de escuela es suburbana.

$$X_3 \perp X_5 \mid X_0 = 0, X_1 = x_1, X_2 = x_2, X_4 = x_4$$

pero,

$$X_3 \not\perp X_5 \mid X_0 = 1, X_1 = x_1, X_2 = x_2, X_4 = x_4$$

con correlación del mismo tipo.

- Tipo de escuela y los planes futuros son condicionalmente dependientes.

$$X_0 \not\perp X_3 \mid [X_1, X_2, X_4, X_5]$$

- El sexo y la preferencia de cursos son condicionalmente dependientes.

$$X_1 \not\perp X_4 \mid [X_0, X_2, X_3, X_5]$$

- La preferencia de cursos y la contestación a la pregunta de la encuesta son condicionalmente dependientes.

$$X_4 \not\perp X_5 \mid [X_0, X_1, X_2, X_3]$$

Análisis con modelo log-lineal

Utilizando el paquete R se utilizó la función *glm* para ajustar un modelo log-lineal al conjunto de datos de las mujeres y las matemáticas, obteniendo los siguientes resultados:

Coeficientes de los parámetros β, error estandarizado, valor z y $P(>|z|)$

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.61092    0.16440  21.964  < 2e-16 ***
X0u                0.32091    0.21595   1.486 0.137274
X1m                0.32091    0.21595   1.486 0.137274
X2n               -0.31508    0.25311  -1.245 0.213189
X3j               -1.30833    0.35641  -3.671 0.000242 ***
X4l               -0.83833    0.29921  -2.802 0.005082 **
X5d               -0.83833    0.29921  -2.802 0.005082 **
X0u:X1m            0.43861    0.27462   1.597 0.110230
X0u:X2n            0.39059    0.31915   1.224 0.221006
X0u:X3j           -1.93035    0.80414  -2.401 0.016372 *
X0u:X4l            0.37224    0.37468   0.993 0.320473
X0u:X5d            0.08456    0.38833   0.218 0.827628
X1m:X2n            0.25446    0.32327   0.787 0.431208
X1m:X3j           -0.13859    0.47955  -0.289 0.772587
X1m:X4l           -1.14759    0.50199  -2.286 0.022249 *
X1m:X5d           -0.79091    0.45731  -1.729 0.083724 .
X2n:X3j            0.09194    0.53765   0.171 0.864224
X2n:X4l            0.25054    0.43958   0.570 0.568706
X2n:X5d           -0.05961    0.46634  -0.128 0.898283
X3j:X4l            0.48165    0.57653   0.835 0.403470
X3j:X5d            0.73297    0.54830   1.337 0.181291
X4l:X5d            0.55065    0.48514   1.135 0.256363
X0u:X1m:X2n       -0.56696    0.40385  -1.404 0.160352
X0u:X1m:X3j        0.88314    0.93266   0.947 0.343688
X0u:X1m:X4l        0.32353    0.58767   0.551 0.581962
X0u:X1m:X5d       -0.10214    0.57202  -0.179 0.858279
X0u:X2n:X3j       -0.86059    1.35161  -0.637 0.524309
X0u:X2n:X4l       -0.26543    0.54008  -0.491 0.623102
X0u:X2n:X5d        0.13826    0.57675   0.240 0.810552
X0u:X3j:X4l        0.90073    1.04079   0.865 0.386805
X0u:X3j:X5d        1.40710    0.99343   1.416 0.156659
X0u:X4l:X5d        0.74472    0.58827   1.266 0.205533
X1m:X2n:X3j        0.19183    0.69246   0.277 0.781760
X1m:X2n:X4l       -0.34407    0.73702  -0.467 0.640615
X1m:X2n:X5d        0.76209    0.64074   1.189 0.234284
X1m:X3j:X4l        0.96526    0.84914   1.137 0.255641
X1m:X3j:X5d        0.49081    0.79283   0.619 0.535879
X1m:X4l:X5d        1.69763    0.75812   2.239 0.025139 *
X2n:X3j:X4l        0.32928    0.81307   0.405 0.685494
X2n:X3j:X5d       -0.52817    0.89643  -0.589 0.555728
X2n:X4l:X5d        0.81730    0.68676   1.190 0.234016
```

```
X3j:X4l:X5d              -0.31176    0.84518  -0.369 0.712230
X0u:X1m:X2n:X3j           0.22604    1.53271   0.147 0.882756
X0u:X1m:X2n:X4l           0.62874    0.85325   0.737 0.461198
X0u:X1m:X2n:X5d          -0.42938    0.79633  -0.539 0.589749
X0u:X1m:X3j:X4l          -3.25472    1.62034  -2.009 0.044573 *
X0u:X1m:X3j:X5d          -1.79498    1.31634  -1.364 0.172690
X0u:X1m:X4l:X5d          -1.48928    0.90224  -1.651 0.098813 .
X0u:X2n:X3j:X4l          -0.53754    1.72033  -0.312 0.754690
X0u:X2n:X3j:X5d           1.26046    1.62937   0.774 0.439175
X0u:X2n:X4l:X5d          -1.30034    0.83144  -1.564 0.117828
X0u:X3j:X4l:X5d          -1.67676    1.32289  -1.267 0.204977
X1m:X2n:X3j:X4l          -1.30619    1.27955  -1.021 0.307339
X1m:X2n:X3j:X5d          -0.27966    1.17591  -0.238 0.812014
X1m:X2n:X4l:X5d          -1.98467    1.09271  -1.816 0.069328 .
X1m:X3j:X4l:X5d          -1.68521    1.25362  -1.344 0.178861
X2n:X3j:X4l:X5d          -1.27933    1.29445  -0.988 0.322996
X0u:X1m:X2n:X3j:X4l       2.93106    2.38044   1.231 0.218208
X0u:X1m:X2n:X3j:X5d      -0.05305    2.02573  -0.026 0.979109
X0u:X1m:X2n:X4l:X5d       1.70990    1.29568   1.320 0.186936
X0u:X1m:X3j:X4l:X5d       4.07954    2.10483   1.938 0.052601 .
X0u:X2n:X3j:X4l:X5d       1.76237    2.13131   0.827 0.408295
X1m:X2n:X3j:X4l:X5d       2.99389    1.91754   1.561 0.118448
X0u:X1m:X2n:X3j:X4l:X5d  -3.93553    3.06649  -1.283 0.199353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ajustando un modelo en particular al conjunto de datos:

```
glm(counts ~X0+X1+X3+X4+X5+ X1*X4+ X1*X5 + X4*X5+ X0*X3*X5)
```

se obtiene:

```
   Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1862  -0.6750  -0.2010   0.7637   1.7289

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.31628    0.09042  36.676  < 2e-16 ***
X0u          0.72640    0.08464   8.582  < 2e-16 ***
X1m          0.52342    0.08437   6.204 5.51e-10 ***
X3j         -1.05605    0.13676  -7.722 1.15e-14 ***
X4l         -0.52040    0.10401  -5.003 5.63e-07 ***
X5d         -0.76395    0.14686  -5.202 1.97e-07 ***
```

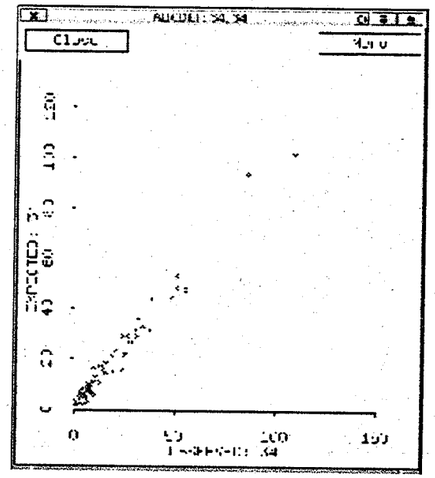


```
X0u:X3j      -1.67081    0.23745  -7.036 1.97e-12 ***
X0u:X5d       0.02316    0.14248   0.163   0.8709
X1m:X4l      -0.75024    0.12606  -5.951 2.66e-09 ***
X1m:X5d      -0.51011    0.12586  -4.053 5.05e-05 ***
X3j:X5d       0.28880    0.21631   1.335   0.1818
X4l:X5d       0.98329    0.12698   7.744 9.67e-15 ***
X0u:X3j:X5d   0.95899    0.32741   2.929   0.0034 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Null deviance: 1056.233  on 63  degrees of freedom
Residual deviance:   61.638  on 51  degrees of freedom
AIC: 362.83

P=0.15
```

En la figura siguiente se muestra la gráfica obtenida de los residuos.



Gráfica de los residuos para el modelo log-lineal del conjunto de datos de las mujeres y las matemáticas

Se observa que este modelo es un caso especial del modelo gráfico clásico simplificado que obtuvimos antes. El modelo gráfico generalizado simplificado que nosotros obtuvimos contiene todas las independencias encontradas en el modelo log-lineal. Además encontramos varias más, porque no coinciden con igualar ciertos parámetros en el modelo log-lineal a cero.

Capítulo 5

Conclusiones

Las principales contribuciones en esta tésis fueron:

- En 2.4.1 se propuso un algoritmo para encontrar independencias implicadas en modelos gráficos generalizados. Este algoritmo es eficiente y efectivo
- En 2.4.3 se propone una visualización alterna para los modelos gráficos generalizados.
- En 3.3.1 se propone un método para obtener los estimadores de máxima verosimiliutd de un modelo gráfico generalizado basado en el algorito Newton Raphson.
- En 3.3.2 se propone una mejora al método anterior proporcionando una solución inicial al algoritmo de Newton Raphson.
- En 4.1.3 se propone la utilización de los modelos gráficos generalizados para el análisis de los accesos a un sitio Web.

De este trabajo de tesis se puede concluir que:

- El algoritmo propuesto para encontrar independencias implicadas es un algoritmo novedoso, eficiente y efectivo.
- El nuevo tipo de visualización para los modelos gráficos generalizados muestra más información al utilizar el ancho del segmento como medida visual de la magnitud de la asociación entre variables y utilizar color verde o rojo para expresar el tipo de asociación o gris para indicar una posible independencia condicional. Todo esto hace que los modelos gráficos generalizados proporcionen mayor información y sean más atractivos a la vista.
- Los dos métodos que se proponen para obtener los estimadores de máxima verosimilitud son efectivos y su comparación no proporcionó una preferencia de uno sobre otro.
- La estimación en modelo gráficos generalizados utilizada en [38], se realiza utilizando formulas cerradas. En este trabajo de tesis se propuso un método de estimación iterativo que se puede aplicar a todo tipo de modelos.

- En [8], se propuso el uso de modelos gráficos para minería de datos en el Web. En este trabajo de tesis se muestra que el uso de modelos gráficos generalizados, da una descripcion más fina que los modelos gráficos clásicos.
- Se comparó los resultados con reglas de asociación. en general no se encontraron incongruencias. Sin embargo, un modelo gráfico permite predecir y cuantificar su calidad.

Dentro del trabajo futuro se puede considerar lo siguiente:

- Extensión de los algoritmos de estimación y selección de modelos para datos discretos.
- Implementar los algoritmos dentro de un ambiente Multi-plataforma para poder incluir la herramienta de análisis de accesos a Internet a un equipo servidor de hojas Web.
- Realizar un análisis más detallado de los accesos al servidor de CIMAT, utilizando modelos gráficos.
- Realizar más experimentos sobre la utilización del Newton Raphson con el IPF. Con el fin de encontrar algún patrón que indique cuando aplicar cada método.

Apéndice A

Estimación

A.1 Obtención de formulas cerradas para los estimadores de máxima verosimilitud.

Las formulas cerradas para los estimadores de máxima verosimilitud se pueden obtener de varias formas, aqui veremos dos de ellas.

Tomando la función de log-verosimilitud definida en la sección (3.2.1).

$$l(p;n) = \sum_{x} n_x \log p(x) \tag{A.1}$$

Tomando el caso de dimensión 2 para el modelo,

$$\log p(x_i, x_j) = \beta_0 + \beta_1 x_i + \beta_2 x_j$$

la primer forma de obtener las formulas de para obtener los estimadores es la siguiente: Desarrollando la ecuación $(A.1)$,

$$\begin{aligned} l(p;n) &= \sum_{i,j} n_{i,j} \log p_{i,j} \\ &= \sum_{i,j} n_{i,j} (\beta_0 + \beta_1 x_i + \beta_2 x_j) \end{aligned}$$

Se quiere maximizar la función $l(p;n)$, la cual puede escribirse,

$$\max_{\beta_0,\beta_1,\beta_2} n_{00}(\beta_0) + n_{01}(\beta_0 + \beta_2) + n_{10}(\beta_0 + \beta_1) + n_{11}(\beta_0 + \beta_1 + \beta_2) \tag{A.2}$$

Se debe cumplir que,

$$\sum_{i,j} e^{\beta_0 + \beta_1 x_i + \beta_2 x_j} = 1$$

Sacando el termino e^{β_0} de la sumatoria tenemos que,

$$e^{\beta_0}\left(1+e^{\beta_1}+e^{\beta_2}+e^{\beta_1+\beta_2}\right)=1 \tag{A.3}$$

Despejando β_0,

$$\beta_0=-\log\left(1+e^{\beta_1}+e^{\beta_2}+e^{\beta_1+\beta_2}\right)$$

Juntando terminos β_0 en $(A.2)$y sustituyendo β_0 por el resultado anterior,

$$\max_{\beta_1,\beta_2} -n_{++}\log\left(1+e^{\beta_1}+e^{\beta_2}+e^{\beta_1+\beta_2}\right)+n_{01}\beta_2+n_{10}\beta_1+n_{11}\left(\beta_1+\beta_2\right) \tag{A.4}$$

Igualando la derivada parcial con respecto a β_1 de la función de máximización $(\grave{A}.4)$, tenemos:

$$-n_{++}\frac{\left(e^{\beta_1}+e^{\beta_1+\beta_2}\right)}{1+e^{\beta_1}+e^{\beta_2}+e^{\beta_1+\beta_2}}+n_{10}+n_{11}=0$$

Multiplicando ambos lados de la ecuación anterior por -$1/n_{++}$,

$$\frac{\left(e^{\beta_1}+e^{\beta_1+\beta_2}\right)}{1+e^{\beta_1}+e^{\beta_2}+e^{\beta_1+\beta_2}}=\frac{n_{1+}}{n_{++}}$$

Sustituyendo e^{β_0} de la ecuación $(A.3)$ en la ecuación anterior,

$$\begin{aligned} e^{\beta_0}\left(e^{\beta_1}+e^{\beta_1+\beta_2}\right) &= \frac{n_{1+}}{n_{++}} \\ e^{\beta_0+\beta_1}+e^{\beta_0+\beta_1+\beta_2} &= \frac{n_{1+}}{n_{++}} \end{aligned} \tag{A.5}$$

Por definición,

$$\begin{aligned} p_{10} &= e^{\beta_0+\beta_1} \\ p_{11} &= e^{\beta_0+\beta_1+\beta_2} \end{aligned}$$

entonces sustituyendo p_{10} y p_{11} en $(A.5)$,

$$p_{10}+p_{11}=\frac{n_{1+}}{n_{++}}$$

tenemos entonces que el estimador $\widehat{p}_{1+}$ puede obtenerse por la formula cerrada:

$$\widehat{p}_{1+}=\frac{n_{1+}}{n_{++}}$$

De la misma manera podemos obtener que el estimador $\widehat{p}_{+1}$ ahora para la derivada parcial con respecto a β_1 de la función de máximización $(A.4)$:

$$\widehat{p}_{+1}=\frac{n_{+1}}{n_{++}}$$

Los estimadores para $\widehat{p}_{0+}$ y $\widehat{p}_{+0}$ se pueden obtener de la relación:

$$\begin{aligned} p_{0+} &= 1-p_{1+} \\ p_{+0} &= 1-p_{+1} \end{aligned} \tag{A.6}$$

De esta manera la formula para obtener los estimadores de máxima verosimilitud para el modelo se puede escribir como:

$$\widehat{p}(X_0=i,X_1=j)=\frac{n_{i+}n_{+j}}{n_{++}n_{++}}$$

Otra forma de obtener los estimadores es considerando la función de verosimilitud igual a,

$$\begin{aligned} vero\,(p;n) &= \prod p_{ij}^{n_{ij}} \\ &= \prod (p_{i+}p_{+j})^{n_{ij}} \end{aligned}$$

Desarrollando para todos los i,j :

$$\begin{aligned} vero\,(p;n) &= (p_{0+}p_{+0})^{n_{00}}\,(p_{1+}p_{+0})^{n_{10}}\,(p_{0+}p_{+1})^{n_{01}}\,(p_{1+}p_{1j})^{n_{11}} \\ &= p_{0+}^{n_{0+}}\,p_{+0}^{n_{+0}}\,p_{1+}^{n_{1+}}\,p_{+1}^{n_{+1}} \end{aligned}$$

Sustituyendo ecuación $(A.6)$ en la ecuación anterior, se tiene:

$$vero\,(p;n)=p_{0+}^{n_{0+}}\,(1-p_{0+})^{n_{1+}}\,p_{+0}^{n_{+0}}\,(1-p_{+0})^{n_{+1}}$$

Maximizando la función de verosimilitud en p_{0+}, implica,

$$\max_{p_{0+}} n_{0+}\log p_{0+}+n_{1+}\log\left(1-p_{0+}\right)$$

Esto es igual a:

$$\frac{n_{0+}}{p_{0+}}=\frac{n_{1+}}{1-p_{0+}}$$

Despejando p_{0+},

$$\widehat{p}_{0+}=\frac{n_{0+}}{n_{++}}$$

se obtiene el estimador $\widehat{p}_{0+}$. Los otros estimadores pueden obtenerse de la misma manera.

Apéndice B

Manual de Usuario de $\mathcal{MOGG}$

B.1 Introducción

Objetivo general del programa: "Obtener modelos gráficos clásicos y modelos gráficos generalizados a partir de archivos de texto con muestras o tablas de contingencia de datos discretos de tipo binario".

B.1.1 ¿Qué es $\mathcal{MOGG}$?

$\mathcal{MOGG}$ (Modelos Gráficos clásicos y Generalizados), es un software realizado bajo la plataforma Windows 98, con el lenguaje de programación C++ Builder 4.0, cuyo proposito es proporcionar al usuario una interfaz gráfica amigable y confiable, con la cuál pueda interactuar para generar modelos gráficos clásicos y generalizados a partir de archivos de texto en formato ASCII, con muestras o con tablas de contingencia de datos discretos de tipo binario.

$\mathcal{MOGG}$ contiene varios métodos que ayudan a obtener modelo gráficos de manera rápida y eficiente, además $\mathcal{MOGG}$ también cuenta con métodos para el preprocesamiento de los datos, a manera de eliminar variables no deseadas en el conjunto de datos, o la conversión de los datos al tipo usado comunmente en minería de datos para la generación de reglas de asociación. Cuenta con métodos para la verificación de modelos gráficos y predicción de variables. También el programa permite copiar, guardar y colocar etiquetas de texto a los modelos gráficos.

B.1.2 Entorno del programa

En la figura (B.1) se muestra el entorno de especificación y construcción de modelos gráficos en $\mathcal{MOGG}$. Como se puede apreciar se utilizan dos tipos de ventanas de interacción.

- Ventana de Información Textual o de Datos.- En esta ventana se muestra información de los diferentes pasos que se van realizando para la construcción de los modelos gráficos y algunos resultados numéricos que pueden ser de interés para el usuario. El usuario no puede escribir en esta ventana de texto, solo podrá ver resultados.

- Ventana de Información Visual.- En esta ventana se muestran los modelos gráficos clásicos o generalizados que se vayan obteniendo. El usuario puede interactuar con los modelos mediante el mouse para especificar el modelo deseado.

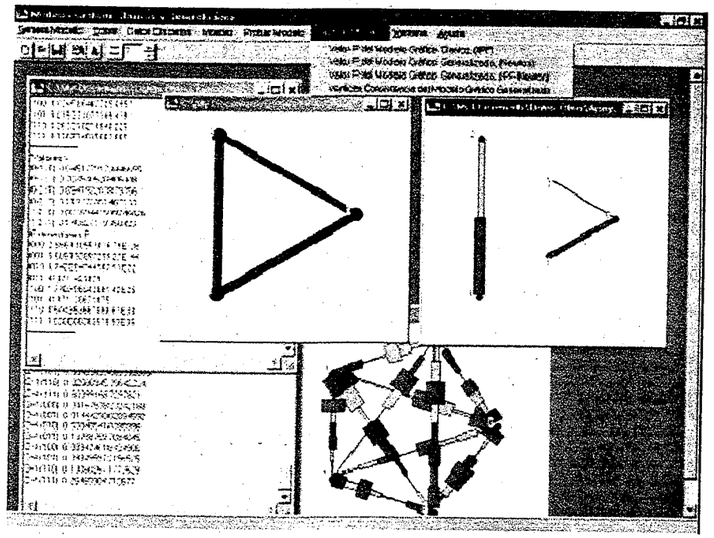


Figura B.1: Entorno de $\mathcal{MOGG}$

$\mathcal{MOGG}$ cuenta con las siguientes opciones en la barra de menu principal:

- Genera Modelo.-
  - Generar un modelo a partir de un archivo con datos.
  - Guardar la información de la ventana de datos en formato texto.
  - Guardar un modelo gráfico en formato de imagen bitmap (bmp).
  - Cerrar una ventana.
  - Cerrar todas las ventanas.
  - Salir de $\mathcal{MOGG}$.
- Editar
  - Copiar texto de la ventana de datos o un modelo gráfico al Clipboard.
  - Agregar texto a una ventana con un modelo gráfico.
  - Ajustar un modelo gráfico a un tamaño de ventana particular.
  - Abrir cuadro de dialogo de Preferencias.
- Datos Discretos

- Convierte datos muestrales binarios al formato típico utilizado por muchos programas para obtener reglas de asociación. La opción de menos de 2000 datos despliega los resultados en ventas de datos. La opción de más de 2000 datos despliega un cuadro de dialogo donde el usuario podrá especificar el nombre y la ruta del archivo donde se almacenarán los datos.
- Elimina una variable de un conjunto de muestras.
- Genera datos muestrales apartir de una serie de asociaciones entre variables especificadas por el usuario.

• Mostrar

- Muestra en la ventana de datos los cliques que hay en el modelo gráfico.
- Muestra en la ventana de datos las frecuencias marginales del conjunto de datos observados.
- Muestra en la ventana de datos las frecuencias marginales del conjunto de datos observados en los cliques.
- Muestra en la ventana de datos los valores de las medidas de asociación entre variables.
- Muestra en la ventana de datos la tabla de contingencia del conjunto de datos observado.

• Probar Modelo

- Realiza un etiquetado del modelo gráfico clásico.
- Realiza un etiquetado del modelo gráfico generalizado
- Realiza un Backward del modelo gráfico clásico.

• Verificar Modelo

- Obtener el valor de P del modelo gráfico clásico utilizando el método IPF.
- Obtener el valor de P del modelo gráfico generalizado utilizando el método Newton-Raphson.
- Obtener el valor de P del modelo gráfico generalizado utilizando el método IPF como inicializador del Newton-Raphson
- Verificar la consistencia del modelo gráfico generalizado.

• Ventana

- Organiza ventanas en forma de Cascada.
- Organiza ventanas en forma Horizontal.
- Organiza ventanas en forma Vertical.
- Minimiza todas las ventanas.

- Organiza todos los iconos minimizados.

• Ayuda

- Indice de la ayuda
- Acerca de...

$\mathcal{MOGG}$ cuenta con una barra de estado localizada en la parte inferior de la ventana principal. En esta barra se le proporciona al usuario información referente a los métodos aplicados para la construcción de modelos y con ayuda sobre el entorno del programa.

B.1.3 Opciones de inicio

El programa al inicio carga las siguientes opciones predeterminadas a partir de un archivo denominado config.cfg, el cuál se encuentra en el directorio donde esta instalado el programa.

• Orden de lectura de datos.- Existen dos posibilidades de lectura de variables, en orden:

- De izquierda a derecha $(X_0, X_1, ..., X_n)$
- De derecha a izquierda $(X_n, X_{n-1}, ..., X_0)$

Esta opción configura el orden de lectura de las variables en el modelo con respecto a los datos en el archivo de texto.

• Tipo de medida de asociación en los modelos.- Se puede utilizar alguna de las siguientes dos medidas de asociación entre los datos: Medida de correlación (OddsRatio) y medida de covarianza.

• Mostrar información de:

- Estimadores P
- Medidas de Asociación
- Valores P en las aristas
- Residuos

Las opciones que estén activas determinarán lo que se mostrará en la ventana de datos cuando el usuario ejecute alguna acción.

• Umbral de significancia del Valor P para la aceptación de modelos.

Una vez dentro del programa se pueden modificar estas opciones en el cuadro de dialogo *Preferencias*, figura $(B.2)$, que se encuentra en el menu *Editar*.

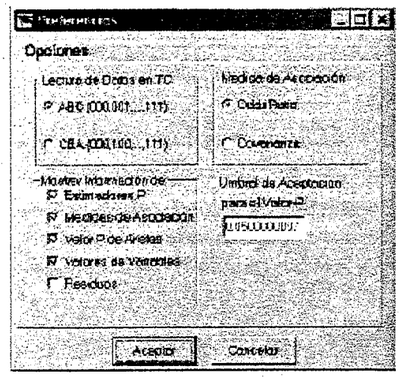


Figura B.2: Cuadro de Dialogo de Preferencias

B.2 Datos Discretos

B.2.1 Características de los archivos de datos

Se puede cargar cualquier tipo de archivo de texto que contenga el formato de datos correcto. Los datos observados se pueden cargar de dos formas distintas: En forma de datos muestrales o en forma de tabla de contingencia.

El archivo necesariamente debe incluir un retorno de linea extra al final del archivo para indicar el fin del archivo, si no existe este retorno o existen dos retornos de linea al final del archivo, el archivo será cargado erroneamente. En este caso se mostrará un mensaje de error de lectura de archivo.

Los datos especificados en forma de muestras deberán ser especificados como se muestra en la figura $(B.3)$. El primer renglón del archivo deberá contener el numero de variables. en este caso $n = 4$. Entre cada valor de las variables deberán haber al menos un espacio que los separe. No importa la alineación de los valores. La forma en que se considera el orden de la lectura para este tipo de archivos es de la forma $(X_0, X_1, ..., X_n)$.

Los datos especificados en forma de tabla de contingencia deberán ser especificados como se muestra en la figura $(B.4)$. El primer renglón deberá incluir un T para indicar al programa que los datos observados están especificados en forma de una tabla de contingencia. Después de la T se deberá indicar el número de variables. El orden de lectura de la tabla de contingencia podrá cambiarse después dentro del cuadro de dialogo de Preferencias.

B.2.2 Cargar archivos de datos en $\mathcal{MOGG}$

Al cargar un archivo de datos se mostrará un cuadro de dialogo como el de la figura $(B.5)$. El usuario podrá elegir el modelo gráfico inicial. Las primeras dos opciones despliegan un modelo gráfico que no visualiza las aristas o segmentos que pueden considerarse independientes bajo el nivel de significancia especificado en las opciones de inicio. Las ultimas dos mostrarán un

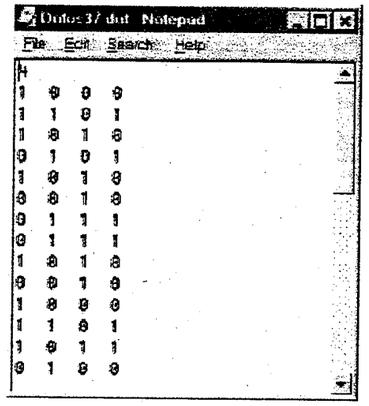


Figura B.3: Ejemplo de archivo de datos muestrales

modelo gráfico indicando con un grosor pequeño de arista o un color gris de segmento, las independencias que posiblemente se pueden considerar en el modelo.

B.2.3 Generación de datos relacionados

En la figura $(B.6)$ se muestra el cuadro de dialogo de generación de datos dado un modelo de asociaciones. Por medio de esta interface el usuario puede especificar lo siguiente: El número de variables que tendrá el conjunto de datos, el número de muestras que desea, las asociaciones entre variables, la probabilidad de tener esa asociación.

La función de generación de datos muestrales para un modelo especificado, utiliza una función de generación de números aleatorios para determinar cuál o cuales asociaciones se tienen dentro de una muestra, dada la probabilidad de cada asociación. Se incluyen automáticamente el término de normalización y los términos β de paridad individual de las variables especificados como se muestra en la ecuación (2.2). La interface permite especificar una probabilidad a cada asociación que se ha ingresado al modelo mediante el boton "Agregar Asociación". El botón "Probabilidades" genera automáticamente valores aleatorios de probabilidad para cada asociación que ha sido agregada.

B.2.4 Opciones de pre-procesamiento de datos

Si en algún momento se desea eliminar una variable de un conjunto de muestras se puede utilizar la opción de eliminar variables del menú Datos Discretos. Esta acción muestra un cuadro de dialogo como el de la figura $(B.7)$. Una vez que se ha especificado la variable que se desea eliminar aparece un cuadro de dialogo de especificación de archivo, donde se deberá indicar el nombre y ruta del archivo del cuál se extraerá esa variable.

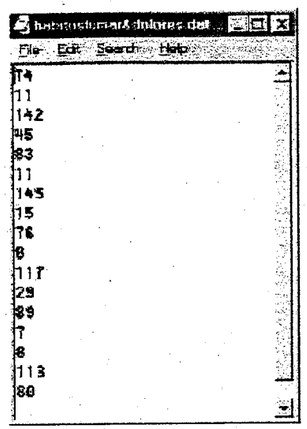


Figura B.4: Ejemplo de archivo con tabla de contingencia

B.3 Modelos gráficos

B.3.1 ¿Cómo obtener Modelos Gráficos Clásicos?

Existen dos formas de obtener un modelo gráfico clásico que sea aceptado bajo un nivel de significancia específico.

La forma más sencilla de lograr esto es utilizando la función Backward del menu "Probar Modelo". Esta acción realiza varios ciclos sobre las aristas del modelo y va seleccionando cada vez la arista que es más suceptible a considerarse fuera del modelo, es decir, la arista que tenga el valor de P más grande. De esta manera se simplifica el modelo automáticamente.

Otra forma de seleccionar un modelo gráfico clásico más sencillo es utilizando la función "Label modelo gráfico clásico" del menu "Probar Modelo". Esta acción en conjunto con la función "Valor P del modelo clásico" genera una selección de modelos utilizando Stepwise. El usuario decide cuales aristas quitar para formular un nuevo modelo. Este proceso se repite hasta no encontrar un modelo más sencillo y que sea aceptable bajo un nivel de significancia especificado por el usuario.

Los pasos para obtener un modelo de esta forma consisten en los siguientes: Realizar un Label del modelo gráfico clásico, elegir la arista o aristas con un valor de P pequeño o aristas con un grosor pequeño para que salgan del modelo, obtener el valor de P del modelo especificado y comprobar que sea aceptado bajo un nivel de significancia. Este procedimiento de selección continua hasta que el usuario obtiene un modelo sencillo y aceptable. En la figura (B.8) se muestra el procedimiento seguido para realizar este tipo de selección.

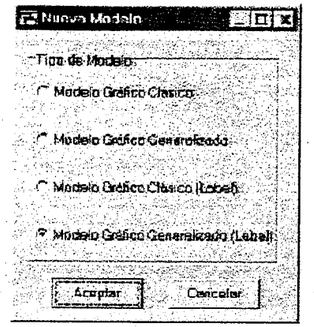


Figura B.5: Cuadro de Dialogo para un nuevo modelo

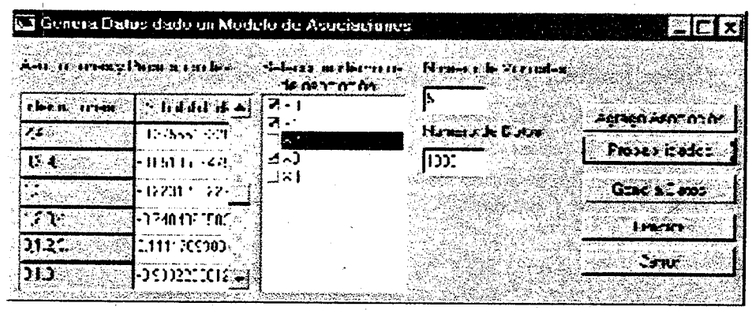


Figura B.6: Cuadro de Dialogo para generar datos con asociaciones entre variables

B.3.2 ¿Cómo obtener Modelos Gráficos Generalizados?

Al igual que para modelos gráficos clásicos, la obtención de modelos gráficos generalizados se puede realizar mediante la selección de modelos utilizando Stepwise. Si el valor de P es mayor al nivel de significancia especificado por el usuario el color del segmento es de color gris. En este caso la selección no se realiza sobre las aristas sino sobre los segmentos en color gris que puedan salir del modelo. En la figura (B.9) se muestra el procedimiento para construir un modelo gráfico generalizado.

Para ahorrar un poco de trabajo una estrategia es utilizar el método Backwise para obtener el modelo gráfico clásico más simple y despues aplicar el método de etiquetado generalizado para seleccionar ahora los segmentos que pueden salir del modelo.

Una vez que se ha obtenido un modelo gráfico generalizado que es aceptado bajo un nivel de significancia se debe verificar que el modelo obtenido sea un modelo que no tenga independencias implicadas, es decir, que sea un modelo válido. Para hacer esto el usuario puede utilizar la

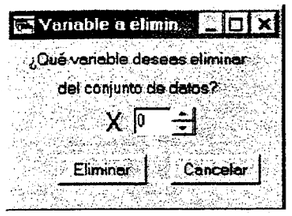


Figura B.7: Cuadro de Dialogo para eliminar una variable

función "Verificar Consistencia del Modelo Grafico Generalizado" del menu "Verificar Modelo". Esta función mostrará un mensaje al usuario indicando si el modelo es válido o no. Si el modelo no es válido se mostrarán textualmente en la ventana de datos las independencias implicadas. Además visualmente aquellas independencias implicadas se mostrarán como segmentos con un contorno de color amarillo, como se muestra en la figura $(B.10)$.

B.3.3 ¿Cómo realizar predicciones sobre variables?

Una vez que se tienen los estimadores de máxima verosimilitud de un modelo dado, es fácil determinar para alguna variable, cual es la probabilidad de que tenga un valor dado, si se conocen los valores de las variables restantes, por ejemplo dadas tres variables X_i, X_j, X_k, la probabilidad de que $X_i = 0$ dado que $X_j = 1$ y $X_k = 0$, se puede obtener de la siguiente manera,

$$p(X_i = 0 \mid X_j = 1, X_k = 0) = \frac{p(X_i = 0, X_j = 1, X_k = 0)}{p(X_j = 1, X_k = 0)}$$

entonces de manera general se tiene que para $X = (X_0, ..., X_n)$ variables aleatorias,

$$p(X_i = x_i \mid X_0 = x_0, ..., X_n = x_n) = \frac{p(X_i = x_i, X_0 = x_0, ..., X_n = x_n)}{p(X_0 = x_0, ..., X_n = x_n)}$$

Ejemplo

Para el modelo gráfico generalizado de la figura (3.9), la probabilidad de que un accidente ocurra en una tacleada $X_2 = 0$ dado que el accidente fue en defensa y no aventando el balón, $X_0 = 0$ y $X_1 = 1$, es de 0.7582. en la figura $(B.11)$ se muestra la obtención de estos valores de predicción en el programa $\mathcal{MOGG}$.

Acerca de...

En la figura $(B.12)$ se muestra el cuadro de dialogo Acerca de.

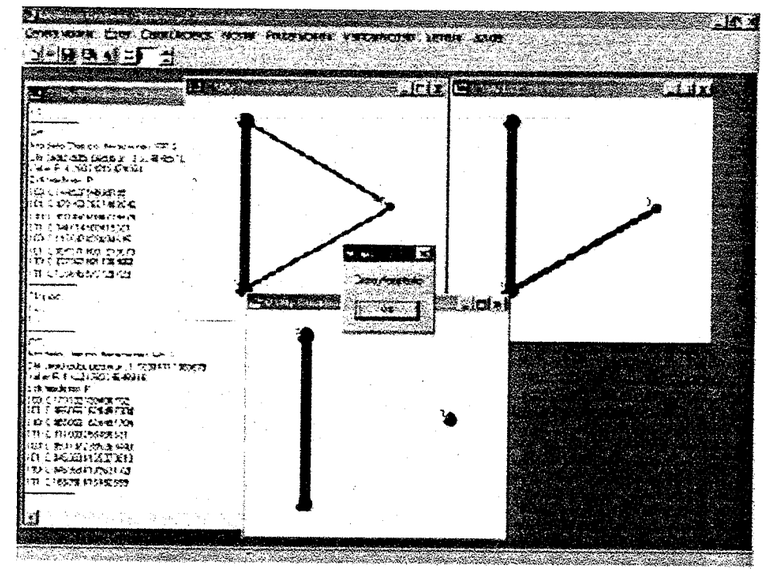


Figura B.8: Ejemplo de construcción de un modelo gráfico clásico

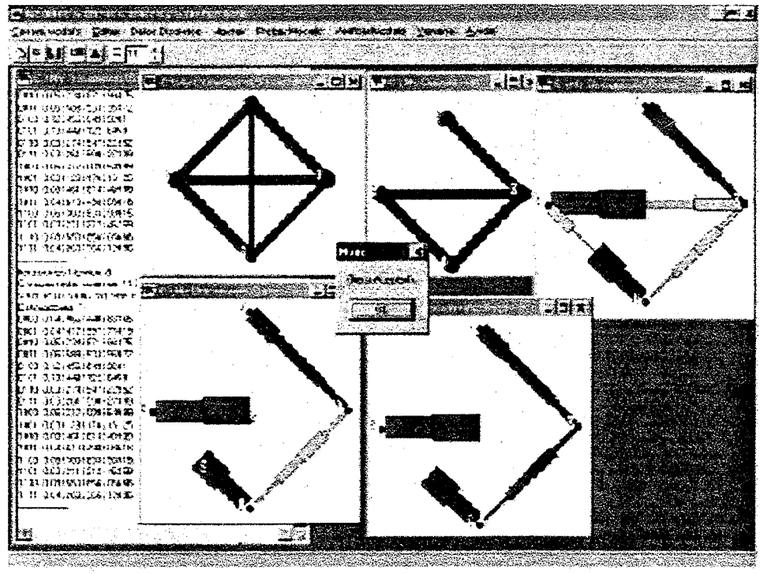


Figura B.9: Ejemplo de construcción de modelo gráfico generalizado

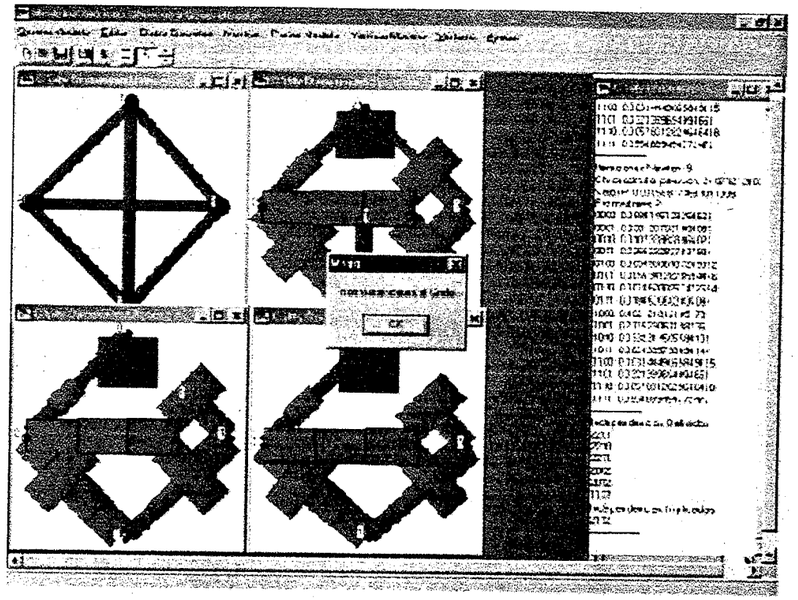


Figura B.10: Ejemplo de Inconsistencia en un modelo gráfico generalizado

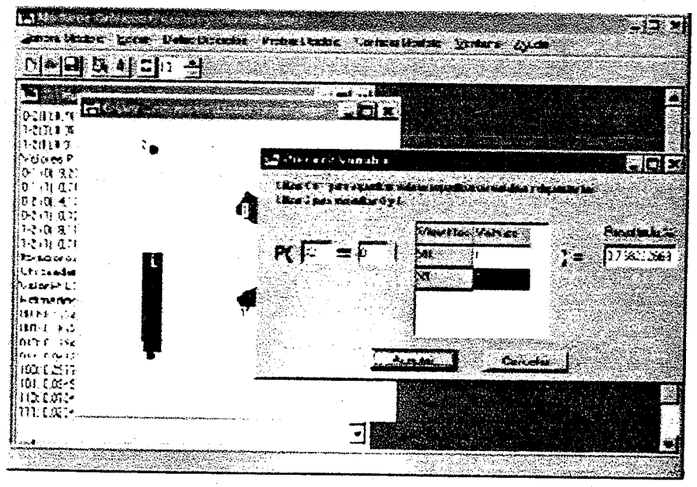


Figura B.11: Ejemplo de predicción sobre una variable

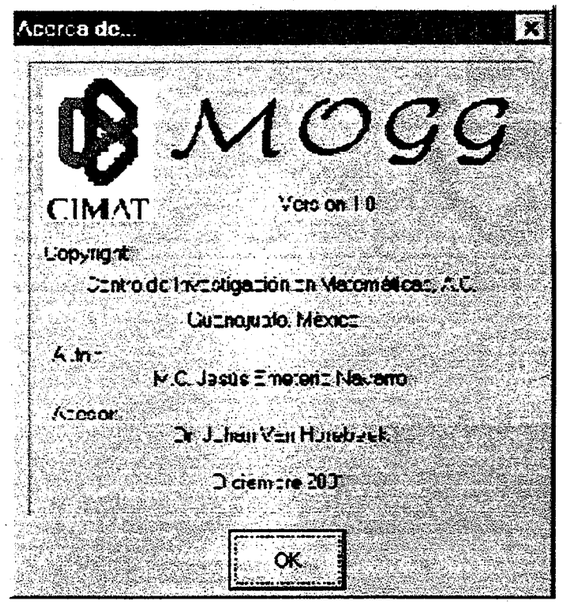


Figura B.12: Acerca de...

Bibliografía

[1] Agrawal, R., Imielinsky, T. y Swami, A. "Mining association rules between sets of items in large databases", In Proceedings of ACM SIGMOD Conference, 207–216, 1993.

[2] Andersen, L. R., Olesen, K., Jensen, F. y Jensen, F., "HUGIN - a shell for building Bayesian belief universes for expert systems", Proceedings of the Eleventh International Joint conference on Artificial Intelligence (IJCAI), Detroit, MI, 1989.

[3] Bartlett, M. S., "Contingency Table Interaction", Journal of the Royal Statistical Society, Suppl. 2,: 248–252, 1935.

[4] Benedetti, J. K. y Brown, M. B., "Strategies for the Selection of Log-Linear Models", Biometrics, 34: 680-686, 1978

[5] Bickel, P. J., Hammel, J. W., y O'Connell, J. W., "Sex bias in graduate admissions: Data from Berkley" Science, 187: 398-403, 1975.

[6] Birch, M. W., " Maximum likelihood in three-way contingency tables", Journal of the Royal Statistical Society, B25, 220-233.

[7] Bishop, Y. M. M., Fienberg, S. E. y Holland, P. W., "Discrete Multivariate Analysis: Theory and Practice", Second edition, MIT Press, Cambridge, Massachusetts, 1975.

[8] Borgelt, C., "Data Mining with Graphical Models", Ph.D thesis,Otto-von-Guerocke-Univesitat Magdeburg, July 2000.

[9] Borges y Levene, "Data Mining of User Navigation Patterns", To appear in Lecture Notes in Artificial Intelligence Springer-Verlag, Berlin, 2000.

[10] Bron, C. y Kerbosch, J., "Finding All Cliques of an Undirected Graph", Algorithm 457 de CACM, Eindhoven, The Netherlands, 1971.

[11] Brown, M. B., "Screening Effects in Multidimensional Contingency Tables", Applied Statistics, 25: 37-46, 1976.

[12] Buchner, Baumgarten, Anand. Mulvenna y Hughes, "Navigation Pattern Discovery from Internet Data"

[13] Buckley, W., "Concussions in football: a multivariate analysis ", American Journal of Sport Medicine 16, 609-617, 1988.

[14] Castillo, E., Gutiérrez, J. M. y Hadi, A. S., "Expert Systems and Probabilistic Network Models", Springer Verlag, New York, 1997.

[15] Chao, Lincoln L., "Introducción a la Estadística"

[16] Cooley, Tan y Srivastava, "Discovery of Interesting Usage Patterns from Web Data", WEBKDD, 1999.

[17] Darroch, J.N., Lauritzen, S. L. y Speed, T. P., "Markov fields and log-linear interaction models for contingency tables", Annals of Statistics 8(3): 522-539 ,1980.

[18] Conjunto de Datos "Creencia en vida después de la vida", General Social Survey, 1991.

[19] Deming, W. E. y Stephan, F. F. , "On a least squares adjustement of a sampled frequency table when the expected marginal totals are known", Ann. Math Statist. ,11: 427-444, 1940.

[20] Edwards, D., "Hierarchical interaction models", Journal of the Royal Statistical Society, Series B 52(1): 3-20, 1990.

[21] Edwards, D.,"Introduction to Graphical Modelling, Springer Texts in Statistics", 1995.

[22] Edward, D., y kreiner, S., "The Analysis of Contingency Tables by Graphical Models", Biometrika, 70: 553-565, 1983.

[23] Erick, Nelson y Schmidt, "Graphical analysis of Computer Log Files" In Communications of the ACM, volume 37, pages 50-56 December 1994.

[24] Estivill-Castro, V. y Houle, M. E., "Robust Distance-Based Clustering with Applications to Spatial Data Mining, Special Isue of Algoritmica,1999.

[25] Estivill-Castro, V. y Yang, J., "Clustering Web Visitors by Fast, Robust and Convergent Algorithms", International Journal of Foundations of Computer Science, 2001.

[26] Fisher, R. A., "On the interpretation of χ^2 from contingency tables, and calculation of P", Journal of the Royal Statistical Society, Series A, 58: 87-94, 1922.

[27] Fowlkes, E. B., Freeny A. E. y Landwehr, J. M., "Evaluating Logistic Models for Large Contingency Tables", Journal of the American Statistical Association, September 1988.

[28] Friendly, M., "Visualizing Categorical Data", SAS Institute, CAry, NC, 2000.

[29] Frydenberg, M. y Lauritzen, S. L. "Decomposition of maximum likelihood in mixed graphical interaction models", Biometrika 76: 539-555, 1989.

[30] Fu, Sandhu y Shih, "Clustering of Web Users Based on Access Patterns", In Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, CA. Springer-Verlag, in press.

[31] Goodman, L. A., "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications", Technometrics, 13: 33-61, 1971.

[32] Grinstead, Charles M. y Snell, J. L., "Introduction to Probability", American Mathematical Society, Providence, Rhode Island, second revised edition, 1997.

[33] Hammersley, J. M. y Clifford, P. E., "Markov fields on finite graphs and lattices", Unpublished manuscript, 1971.

[34] Hartigan, J. A. y Kleiner, B. "Mosaics for contingency tables", Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, 268-273, Springer Verlag, New York, NY, 1981.

[35] Hofmann, H. y Wilheilm A., "Visual Comparison of Association Rules", Ausburg University, 2001.

[36] Hojsgaard Soren, "Split Models for Contingency Tables", Tesis de Doctorado, Research Centre Foulum, Dinamarca, 1998.

[37] Hojsgaard, S. y Thiesson, B. "BIFROST - Block recursive models Induced From Relevant knowledge, Observations, and Statistical Techniques", Journal of Computational Statistics and Data Analysis 19: 155-175, 1995

[38] Van Horebeek, J. J. L., Teugels, J. L., "Generalized Graphical Models for Discrete Data", Statistics and Probability Letters, 38: 41-47, 1998.

[39] Joachims, F., Mitchell y Armstrong, "WebWatcher: A tour guide for the World Wide Web", Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995.

[40] Joshi, K.P., Joshi, A., Yesha, Y., y Krishnapuram, R, "Warehousing and Mining Web Logs",Proceedings of the second international workshop on on Web information and data management, 63-68, 1999.

[41] Kiiveri, H. T., Speed, T. P. y Carlin, J. B., "Recursive causal models", Journal of the Australian Mathematical Society, Series A 36: 30-52, 1984.

[42] Lan, B., Bressan, S. y Ooi, B. C., "Web Servers Pushier",Proceedings Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), Aug. 1999.

[43] Lauritzen, S. L., "Causal Inference from Graphical Models", University of Aalborg, Research Report, 1999.

[44] Lauritzen, S. L., "Lectures on Contingency Tables", University of Aalborg, 3ra. Ed., 1989.

[45] Lauritzen, S. L. y Wermuth, N. "Graphical Models for associations between variables, some of which are qualitative and some quantitative", Annals of Statistics 17(1): 31-57, 1989.

[46] Murray, D y Durrell, D. "Inferring Demographic Attributes of Anonymous Internet Users".

[47] Nasraoui, O., Frigui, H., Joshi, A., y Krishnapuram, R., "Mining Web Access Logs Using Relational Competitive Fuzzy Clustering", Eight International Fuzzy Systems Association World Congress - IFSA 99, Taipei, August 1999.

[48] Padmanabhan, V., y Mogul, J, "Using predictive prefetching to improve World Wide Web latency", In Proceedings of the ACM SIGCOMM '96 Conference on Communications Architectures and Protocols, pp. 22–36, July 1996.

[49] Perkowitz, M. y Etzioni, O., "Adaptive Web sitios: Conceptual Cluster Mining", In Proceedings of the 15th International Joint Conference on AI (IJCAI-97), Nagoya, Japan, August 23-29, 1997.

[50] Read, T. R. C., Cressie, N. A. C., "Goodness-of-Fit Statistics for Discrete Multivariate Data", Springer Verlag, 1988.

[51] Riedwyl, H. y Schupbach, M., "Siebdiagramme: Graphische darstellung von kontingenztafeln", Technical Report 12, Institute for Mathematical Statistics, University of Bern, Switzerland, 1983.

[52] Schechter, Krishnan y Smith, "Using path profiles to predict HTTP request", Proceedings of the World Wide Web Conference, 1998.

[53] Silani, "Data Mining for e-intelligence Understanding Customer Behavior on the Web".

[54] Snee, R. D., "Graphical display of two-way contingency tables", The American Statistician, 28: 9-12, 1974.

[55] Spiliopoulou, M., Pohle, C. y Faulstich, L., "Improving the Effectiveness of a Web sitio with Web Usage Mining", in Proceedings of the Workshop on Web Usage Analysis and User Profiling, WEBKDD '99, pp. 51-56, San Diego, California, 1999.

[56] Teugels, J. L. y Van Horebeek, J. J. L., "Algebraic description of nominal discrete data", 1995

[57] Van Alstyne, D. J. y Gottfredson, M. R., "A multidimensional CT analysis of parole outcome: new methods and old problems in criminological prediction" Journal of Research in Crime and Delinquency, 15, 172-193, 1978.

[58] Wermuth, N. "Model Search Among Multiplicative Models", Biometrics, 32: 253-263, 1976.

[59] Wermuth, N. y Lauritzen, S. L. "On substantive research hypotheses, conditional independence graphs and graphical chain models", Journal of the Royal Statistical Society, Series B 52: 21-72, 1990.

[60] Whittaker, J., "Graphical Models in Applied Multivariate Statistics", Wiley Series in Probability and Mathematical Statistics, 1989.

[61] Whittaker, J., y Aitkin, M., "A Flexible Strategy for Fitting Complex Log-Linear Models", Biometrics, 34: 487-495, 1978.

[62] Wilson, E. B.y Hilferty, M. M., "The distribution of chi-square", Proceedings of the National Academy of Sciences, Washington, 17: 684-688, 1931.

[63] Wu, K. L., Yu, P. S. y Ballman A., "SpeedTracer: A Web usage mining and analysis tool", IBM Systems Journal, 37: 1, 1998.

[64] Zelen, M. y Severo, N. C., "Probability Function", Handbook of Mathematical Function, U.S. Department of Commerce, Applied Mathematical Series, SS: 925-995, 1965.