EXTRACTA MATHEMATICAE Vol. **10**, Núm. 1, 21 – 26 (1995)

Efecto de un Proceso de Poda en Algunos Coeficientes de Asociación Derivados del Estadístico χ^2

F.J. Cano Sevilla, A. Munduate del Río and A. Pérez Prados

Dpto. de Estadística e I.O., Fac. Matemáticas, Univ. Complutense de Madrid, 28040-*Madrid, Spain*

Dpto. de Física de Materiales, Fac. de Química, Univ. del País Vasco, Aptdo. 1072, 20080-*San Sebastián, Spain*

Dpto. de Estadística e I.O., Univ. Pública de Navarra, 31006-*Pamplona, Spain*

(Presented by M. Molina)

AMS *Subject Class.* (1991): 62H30 *Received November* 28, 1994

1. Introducción.

Para un conjunto I constituído por n elementos extraídos de una población total $\mathcal{I}$, en el que se dispone de información relativa a un conjunto de variables cualitativas $\{V^j\}_{j=1,\ldots,J}$ y a una variable criterio Y que se relaciona con las anteriores, existen distintos métodos para la construcción, a partir de dichos datos, de árboles de decisión. Los coeficientes de asociación obtenidos a partir del estadístico χ^2 medido sobre la matriz que cruza los nodos de un árbol de decisión con las modalidades de la variable criterio Y, sirven para medir la utilidad de ese árbol como predictor de Y, proporcionando una medida de la calidad del árbol a partir de la cual pueden compararse las utilidades de distintos árboles.

De acuerdo con las propiedades conocidas para el estadístico χ^2 cualquiera que sea la poda realizada en un árbol, en ningún caso produce un aumento en el valor de dicho estadístico, que como consecuencia alcanza su máximo valor, entre los asociados a todos los árboles construídos a partir de un mismo conjunto de datos, en el llamado árbol máximo.

Teniendo en cuenta estos resultados, el presente trabajo se centra en el estudio de la variación en los coeficientes de Tschuprow y de Cramer por efecto de un proceso de poda, debido al interés que estos coeficientes presentan como medidas de la calidad de un árbol válidas para seleccionar el árbol óptimo

construído a partir de un conjunto de datos.

2. Terminología y Conceptualización.

Se considera un conjunto de variables $\{V^j\}_{j=1,\ldots,J}$, llamadas variables explicativas, cuyos valores son conocidos para los n elementos, denominados ejemplos, de un conjunto I, llamado conjunto de aprendizaje, extraído de una población total $\mathcal{I}$. Relacionada con estas variables se considera una variable Y, llamada variable criterio, conocida para los elementos de I; se supone que esta variable es cualitativa y el conjunto de sus modalidades se representa por $\mathcal{Y} = \{\mathfrak{y}_k\}_{k=1,\ldots,c}$.

Partiendo de los datos conocidos para las variables anteriores puede construirse una estructura en árbol, para presentar distintos niveles de asociación de los elementos del conjunto de aprendizaje, correspondientes a diferentes grados de homogeneidad, de acuerdo con la información dada por el conjunto de variables explicativas.

En un árbol T los nodos o vértices se corresponden con subconjuntos de I. Se representa por x un nodo cualquiera de T, siendo n_x el número de ejemplos de I situados en x; cada una de las ramas que partiendo de un vértice llega directamente a otro es un arco. Dados dos nodos x_1 y x_2 de un árbol T, si existe un arco que partiendo de x_1 llega a x_2, se dice que x_1 es nodo generador de x_2 y x_2 es nodo sucesor de x_1. Son nodos terminales de un árbol T aquellos que no tienen nodos sucesores. Todos los demás nodos de T se llaman nodos interiores; en particular, el nodo interior que no tiene nodo generador se llama nodo inicial o nodo raíz y se representa porx_0. Para el árbol T, se representa por $\mathcal{T}$ el conjunto de sus nodos terminales y $\mathcal{T}^0$ el de sus nodos interiores. En particular, el árbol $T_{\text{máx}}$ es aquél en cada uno de cuyos nodos terminales o bien existe un único elemento de I, o bien todos los elementos pertenecen a la misma modalidad de la variable criterio Y.

Dado un nodo x de un árbol T, se representa por T_x la rama de T generada por x o subárbol engendrado por el nodo x en el árbol T, es decir el árbol formado por la parte de T que contiene a x y a todos sus nodos sucesores hasta llegar a los correspondientes nodos terminales; $(T_x)^*$ representa dicha rama eliminado el nodo x. Se llama poda del subárbol T_x de T al hecho de considerar en T el nodo x como terminal eliminando toda su rama engendrada. El árbol así obtenido se representa por $T' = T - (T_x)^*$ y se llama subárbol podado de T.

3. Efecto de un Proceso de Poda en el Coeficiente de Tschuprow.

El coeficiente de Tschuprow se obtiene a partir de χ^2 según la expresión:

$$T = \frac{\chi^2/n}{\sqrt{(m-1)(c-1)}}$$

siendo $m = \text{card } \mathcal{T}$ y $c = \text{card } \mathcal{Y}$.

Proposición 1. *La variación producida en el coeficiente de Tschuprow por la poda de la rama T_x del árbol T viene dada por:*

$$\Delta T = \frac{\chi^2/n}{\sqrt{m-1}\sqrt{c-1}} \left(\frac{1}{\sqrt{1+\frac{\Delta m}{m-1}}} - 1 \right) \left(1 + \frac{\Delta\chi^2}{\chi^2 \left(1 - \sqrt{1+\frac{\Delta m}{m-1}} \right)} \right)$$

siendo $\Delta\chi^2$ y Δm las variaciones en el estadístico χ^2 y en el número de nodos terminales de T, respectivamente.

Demostración. Al realizar la poda de la rama T_x la variación producida en este coeficiente, suponiendo $(m + \Delta m) > 1$, es:

$$\begin{aligned}
\Delta T &= T' - T \\
&= \frac{(\chi^2)'/n}{\sqrt{(m'-1)(c-1)}} - \frac{\chi^2/n}{\sqrt{(m-1)(c-1)}} \\
&= \frac{1}{n\sqrt{c-1}} \left[\frac{\chi^2 + \Delta\chi^2}{\sqrt{(m-1)+\Delta m}} - \frac{\chi^2}{\sqrt{m-1}} \right] \\
&= \frac{\chi^2/n}{\sqrt{m-1}\sqrt{c-1}} \left(\frac{1}{\sqrt{1+\frac{\Delta m}{m-1}}} - 1 \right) \left(1 + \frac{\Delta\chi^2}{\chi^2 \left(1 - \sqrt{1+\frac{\Delta m}{m-1}} \right)} \right)
\end{aligned}$$

donde $\Delta m = \text{card } \mathcal{T}_x - 1$. ∎

Proposición 2. *El coeficiente de Tschuprow para el árbol $T_{\text{máx}}$ es:*

$$T_{T_{\text{máx}}} = \sqrt{\frac{c-1}{m-1}}.$$

Demostración. Puesto que $\chi^2_{T_{\text{máx}}} = (c-1)n$ y $\Delta m = -\text{card}\,\mathcal{T}_{\text{máx}} + 1$, se sigue que

$$T_{T_{\text{máx}}} = \frac{c-1}{\sqrt{(m_{T_{\text{máx}}} - 1)\,(c-1)}} = \sqrt{\frac{c-1}{\text{card}\,\mathcal{T}_{\text{máx}} - 1}}.$$

∎

De acuerdo con este resultado, como en cada uno de los nodos terminales de $T_{\text{máx}}$ todos los elementos pertenecen a una determinada modalidad, el número de nodos debe ser mayor o igual que el de modalidades de Y; por tanto, el valor del coeficiente de Tschuprow es menor o igual que la unidad.

La variación del coeficiente de Tschuprow por efecto de un proceso de poda depende tanto de la variación que dicho proceso produce en el estadístico χ^2 como en el número de nodos terminales. Por tanto, el máximo valor del coeficiente no es necesariamente el correspondiente al árbol $T_{\text{máx}}$. Así, al realizarse una poda en un árbol, el coeficiente χ^2 no aumenta, mientras que el valor de card $\mathcal{T}$ disminuye, pudiendo presentarse el caso de que una poda no suponga disminución del estadístico χ^2, con lo cual el valor del coeficiente de Tschuprow aumenta.

4. Efecto de un Proceso de Poda en el Coeficiente de Cramer.

El coeficiente de Cramer se define mediante la expresión:

$$CR = \frac{\chi^2/n}{\min(m-1, c-1)}.$$

Proposición 3. *La variación producida en el coeficiente de Cramer por efecto de la poda de la rama engendrada por el nodo x del árbol T viene dada por:*

$$\Delta(CR) = \begin{cases} -\frac{\Delta m}{(m-1)(m-1+\Delta m)}\frac{\chi^2}{n}\left(1 + \frac{\Delta\chi^2}{\chi^2}\frac{m-1}{-\Delta m}\right) & \text{if} \quad m \leq c \\ \frac{c-m-\Delta m}{(c-1)(m-1+\Delta m)}\frac{\chi^2}{n}\left(1 + \frac{\Delta\chi^2}{\chi^2}\frac{c-1}{c-m-\Delta m}\right) & \text{if} \quad m + \Delta m < c < m \\ \frac{\Delta\chi^2}{n(c-1)} & \text{if} \quad c \leq m + \Delta m \end{cases}$$

Demostración. La variación producida en este coeficiente por efecto de la poda es: $\Delta(CR) = (CR)' - CR$ donde $(CR)'$ es el valor del coeficiente de Cramer para el árbol $T' = T - (T_x)^*$ y CR es el correspondiente al árbol T.

Por la misma definición se tiene:

$$\Delta(CR) = (CR)' - CR = \frac{(\chi^2 + \Delta\chi^2)/n}{\min(m + \Delta m - 1, c - 1)} - \frac{\chi^2/n}{\min(m - 1, c - 1)}.$$

Se supone que $(m + \Delta m) > 1$, y se analizan las tres situaciones siguientes:

1) Si $m \leq c$ entonces $\min(m + \Delta m - 1, c - 1) = m + \Delta m - 1$ y $\min(m - 1, c - 1) = m - 1$, con lo cual:

$$\begin{aligned}
\Delta(CR) &= \frac{(\chi^2 + \Delta\chi^2)/n}{m + \Delta m - 1} - \frac{\chi^2/n}{m - 1} \\
&= \frac{\chi^2}{n}\left(\frac{1}{m + \Delta m - 1} - \frac{1}{m - 1}\right) + \frac{\Delta\chi^2}{n(m + \Delta m - 1)} = \\
&= \frac{\chi^2}{n}\frac{-\Delta m}{(m - 1)(m - 1 + \Delta m)} + \frac{\Delta\chi^2}{n}\frac{1}{(m - 1) + \Delta m} \\
&= \frac{\chi^2}{n}\frac{-\Delta m}{(m - 1)(m - 1 + \Delta m)}\left(1 + \frac{\Delta\chi^2}{\chi^2}\frac{m - 1}{-\Delta m}\right).
\end{aligned}$$

2) Si $m + \Delta m < c < m$ entonces $\min(m + \Delta m - 1, c - 1) = m + \Delta m - 1$ y $\min(m - 1, c - 1) = c - 1$. Por tanto:

$$\begin{aligned}
\Delta(CR) &= \frac{(\chi^2 + \Delta\chi^2)/n}{m + \Delta m - 1} - \frac{\chi^2/n}{c - 1} \\
&= \frac{\chi^2}{n}\left(\frac{1}{m + \Delta m - 1} + \frac{\Delta\chi^2}{\chi^2(m + \Delta m - 1)} - \frac{1}{c - 1}\right) \\
&= \frac{\chi^2}{n}\left(\frac{c - m - \Delta m}{(m + \Delta m - 1)(c - 1)} + \frac{\Delta\chi^2}{\chi^2(m + \Delta m - 1)}\right) \\
&= \frac{\chi^2}{n}\frac{1}{(m + \Delta m - 1)(c - 1)}\left(c - m - \Delta m + \frac{\Delta\chi^2}{\chi^2}(c - 1)\right).
\end{aligned}$$

3) Si $c \leq m + \Delta m$ entonces $\min(m + \Delta m - 1, c - 1) = c - 1$ y $\min(m - 1, c - 1) = c - 1$. En consecuencia:

$$\Delta(CR) = \frac{(\chi^2 + \Delta\chi^2)/n}{c - 1} - \frac{\chi^2/n}{c - 1} = \frac{\Delta\chi^2}{n(c - 1)}.$$

De donde se concluye la tesis enunciada en la proposición. ∎

PROPOSICIÓN 4. *Si el número de nodos terminales del árbol podado* $T - (T_x)^*$ *es mayor o igual que el de modalidades de la variable* Y*, entonces el coeficiente de Cramer no aumenta al realizarse la poda del nodo* x.

Demostración. Según lo obtenido en la proposición anterior, si $c \leq m + \Delta m$, se tiene:

$$\Delta(CR) = \frac{\Delta\chi^2}{n(c-1)}$$

y puesto que $\Delta\chi^2 \leq 0$, se verifica que $\Delta(CR) \leq 0$. ∎

PROPOSICIÓN 5. *El valor del coeficiente de Cramer para el árbol* $T_{\text{máx}}$ *es la unidad, es decir,*

$$(CR)_{T_{\text{máx}}} = 1.$$

Demostración. Para el árbol $T_{\text{máx}}$ es $c \leq m$; luego:

$$(CR)_{T_{\text{máx}}} = \frac{(\chi^2)_{T_{\text{máx}}}}{(c-1)n} = 1.$$

∎

En consecuencia para el coeficiente de Cramer la variación por efecto de una poda depende de la relación existente entre el número de nodos terminales del árbol T y de su correspondiente podado $T - (T_x)^*$ y el número de modalidades de la variable criterio Y. En el caso de mayor interés, que es aquél en que el número de nodos terminales del árbol podado es mayor o igual que el de modalidades de Y, la poda no aumenta el valor de este coeficiente.

BIBLIOGRAFÍA

[1] BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. AND STONE, CH.J., "Classification and Regression Trees", Wadsworth & Brooks, Monterey, California, 1984.

[2] CIAMPI, A., Generalized regression trees, *Computational Statistics and Data Analysis,* **12** (1) (1991), 57–78.

[3] CUESTA, P., "Inducción en Bancos de Datos Cualitativos", Tesis Doctoral, Univ. Complutense de Madrid Madrid, 1989.

[4] GOODMAN, L. AND KRUSKAL, W., Measures of Association for Cross Classifications, *JASA,* **49** (1954), 732–764.

[5] HARTIGAN, J.A., "Clustering Algorithms", John Wiley & Sons, New York, London, Sidney, Toronto, 1975.

[6] MATUSITA, K., Decision rule, based on the distance for the classification problem, *Annals Inst. Statist. Math.,* **8** (1956), 67–77.

[7] MUNDUATE, A., "Cuestiones Notables en la Construcción y Comparación de Arboles de Decisión", Tesis Doctoral, Dpto. de Estadística e Investigación Operativa, Univ. Pública de Navarra, Pamplona, 1993.

[8] QUINLAN, J.R., Induction of decision trees, *Machine Learning,* **1** (1986), 81–106.