QÜESTIIÓ, Vol 20, 2 pp. 167-186, 1996

CRITERIO DE SELECCIÓN DE UN ÁRBOL ÓPTIMO SEGÚN COEFICIENTES DE ASOCIACIÓN DERIVADOS DE χ^2

F.J. CANO SEVILLA*

A. MUNDUATE DEL RIO**

A. PÉREZ PRADOS***

Se analiza en primer lugar la variación que se produce en el valor del coeficiente de contingencia al realizarse un proceso de poda en un árbol de decisión T. Conocido este efecto, se define una cantidad criterio que combina linealmente el coeficiente de contingencia con el índice de simplicidad. A partir de esta cantidad criterio, se propone un método de obtención de un árbol óptimo para cada uno de los distintos valores del parámetro α de la combinación lineal. Para seleccionar el árbol óptimo, entre todos ellos, se utiliza el coeficiente de Tschuprow, dependiente de las dos medidas consideradas para la calidad del árbol.

Criterion for the selection of an optimum tree considering association coefficients obtained from the χ^2

Keywords: Árboles de decisión, proceso de poda, coeficiente de contingencia, coeficiente de Tschuprow, simplicidad, árbol óptimo.

Clasificación AMS: 62H30

* Dep. de Estadística e Investigación Operativa. Universidad Complutense de Madrid.

** Dep. de Física de Materiales. Universidad del País Vasco.

*** Dep. de Estadística e Investigación Operativa. Universidad Pública de Navarra.

- Article rebut el desembre de 1994.
- Acceptat el gener de 1996.

1. INTRODUCCIÓN

Es sabida la existencia de diversos métodos para la construcción de árboles de decisión, obtenidos a partir de un conjunto de datos para las variables cualitativas $\{V^j\}_{j=1,\ldots,J}$ así como para la variable Y, que se relaciona con las anteriores, definidas todas ellas sobre el conjunto de aprendizaje I.

Se conoce también la validez de los coeficientes de asociación obtenidos a partir del estadístico χ^2 (coeficiente de contingencia, coeficiente de Tschuprow y coeficiente de Cramer) para medir la utilidad de un árbol como predictor de la variable criterio Y. Considerando los nodos terminales de T como modalidades de una nueva variable, dichos coeficientes de asociación calculados sobre la matriz que cruza las modalidades de esta nueva variable con las de la variable criterio Y, son una medida de la asociación existente entre T e Y y en consecuencia de la utilidad de T como predictor de Y.

Por otra parte, diferentes criterios para determinar el árbol óptimo entre los construídos a partir de una colección de datos pueden encontrarse en la literatura. Así, por ejemplo, el criterio de complejidad (Breiman, 1984) utiliza una cantidad criterio que combina el error de resustitución y la simplicidad del árbol medida a través del número de sus nodos terminales; el criterio del error esperado (Niblett, 1987) se basa en la probabilidad de error al asignar un nuevo ejemplo a una modalidad de Y en un nodo x de T. El criterio de contribución (Cuesta, 1989) es una generalización del criterio de complejidad considerando la contribución de cada nodo interior a la calidad global del árbol.

Teniendo en cuenta todo ello, y conocida la variación que un proceso de poda produce en el valor del estadístico χ^2 (Pérez Prados y otros, 1994) se estudia la modificación que ésta induce en el coeficiente de contingencia. A la vista de este resultado se plantea la utilización de este coeficiente para la definición de una nueva cantidad criterio para la selección del árbol óptimo.

Esta cantidad criterio combina linealmente dos medidas de la calidad del árbol: el coeficiente de contingencia y la simplicidad y permite construir un proceso de selección del árbol óptimo, de forma que para cada valor del parámetro α de la combinación lineal se obtiene un árbol óptimo. Además el coeficiente de Tschuprow que depende tanto del estadístico χ^2 como del número de nodos terminales de T permite seleccionar entre los anteriores. Con lo cual mediante este planteamiento se consideran dos de los coeficientes de asociación obtenidos a partir de χ^2 como medidas de la calidad, junto con la simplicidad medida a través del número de nodos terminales.

2. CONCEPTOS FUNDAMENTALES

2.1. Estructura en árbol

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

Partiendo de los datos conocidos para las variables anteriores puede construirse lo que se denomina *estructura en árbol*, que es aquella que presenta 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* y una sucesión de arcos consecutivos se llama *camino*.

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 x_0. Para el árbol T, se representa $\tilde{T}$ el conjunto de sus nodos terminales y $\tilde{T}^0$ el de sus nodos interiores.

2.2. Proceso de poda

Dado un nodo x de un árbol T, T_x representa *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*.

2.3. Coeficientes de contingencia y de Tschuprow

Partiendo de la matriz $(\tau, Y)_{\text{card}\,\tau \cdot c}$ que cruza las modalidades de la variable criterio con los nodos terminales, los coeficientes de contingencia y de Tschuprow obtenidos a partir del estadístico χ^2 calculado sobre dicha matriz, supuesto que $\text{card}\,\tau > 1$ y $c > 1$, ya que en otras condiciones existiría un único nodo terminal y una sola modalidad para la variable criterio, son válidos para medir la utilidad del árbol T como predictor de la variable criterio. Estos coeficientes se definen por:

$$\textit{Coeficiente de contingencia:} \quad \text{CP} = \sqrt{\frac{\frac{\chi^2}{n}}{1 + \frac{\chi^2}{n}}}$$

$$\textit{Coeficiente de Tschuprow:} \quad T = \frac{\frac{\chi^2}{n}}{\sqrt{(m-1)(c-1)}}$$

siendo

χ^2: el valor del estadístico calculado sobre $(\tau, Y)_{\text{card}\,\tau \cdot c}$

m: el número de nodos terminales de T.

c: el número de modalidades de la variable criterio Y.

2.4. Simplicidad

Si en cada nodo terminal se asignan los elementos de I a la modalidad de la variable criterio Y que en dicho nodo presenta mayor proporción, cada camino que une el nodo raíz con un nodo terminal es una caracterización para la modalidad asignada a dicho nodo terminal. En consecuencia, la determinación de la modalidad de Y que le corresponde a un elemento cualquiera será más simple cuanto menor sea el número de caracterizaciones. Por este motivo se considera como una medida de la calidad del árbol, la simplicidad del mismo calculada a través del número de sus nodos terminales, definiéndose esta simplicidad $\mathcal{M}(T)$ en la siguiente forma: $\mathcal{M}(T) = \text{card}\,\tau$.

3. EFECTO DE UN PROCESO DE PODA EN EL COEFICIENTE DE CONTINGENCIA

Como se ha indicado, este coeficiente de asociación viene dado por:

$$CP = \sqrt{\frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}}}$$

Es inmediado que, si en el árbol T se realiza la poda de la rama engendrada por el nodo interior x, la variación producida en esta medida viene dada por:

$$\Delta CP = (CP)' - CP$$

donde $(CP)'$ es el valor del coeficiente de contingencia para el árbol podado $T' = T - (T_x)^*$ y CP es el correspondiente a T.

Analizando la variación del cuadrado de este coeficiente se tiene:

$$\Delta(CP)^2 = (CP')^2 - (CP)^2 = \frac{\frac{(\chi^2)'}{n}}{1+\frac{(\chi^2)'}{n}} - \frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}} = \frac{\frac{\chi^2+\Delta\chi^2}{n}}{1+\frac{\chi^2+\Delta\chi^2}{n}} - \frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}} =$$

$$= \frac{\left(1+\frac{\chi^2}{n}\right)\frac{\Delta\chi^2}{n} - \frac{\chi^2}{n}\frac{\Delta\chi^2}{n}}{\left(1+\frac{\chi^2}{n}\right)^2 + \frac{\Delta\chi^2}{n}\left(1+\frac{\chi^2}{n}\right)} = \frac{\frac{\Delta\chi^2}{n}}{\left(1+\frac{\chi^2}{n}\right)^2 + \frac{\Delta\chi^2}{n}\left(1+\frac{\chi^2}{n}\right)} =$$

$$= \frac{1}{\frac{\left(1+\frac{\chi^2}{n}\right)^2}{\frac{\Delta\chi^2}{n}} + \left(1+\frac{\chi^2}{n}\right)}$$

(1) $$\Delta(CP)^2 = \frac{1}{\frac{\left(1+\frac{\chi^2}{n}\right)^2}{\frac{\Delta\chi^2}{n}} + \left(1+\frac{\chi^2}{n}\right)}$$

donde se ha supuesto que $\Delta\chi^2 \neq 0$ ya que en caso contrario, resulta que $\Delta(CP) = 0$.

Proposición 1

El coeficiente de contingencia no aumenta al realizarse la poda de un subárbol T_x cualquiera de T

Demostración

De acuerdo con (1) se tiene:

$$\Delta(\mathrm{CP})^2 = \frac{1}{\dfrac{\left(1+\dfrac{\chi^2}{n}\right)^2}{\dfrac{\Delta\chi^2}{n}} + \left(1+\dfrac{\chi^2}{n}\right)}$$

Pero: $$\Delta\chi^2 = \frac{n}{\sum\limits_{s\in\tau_x} n_{s\cdot}} \sum_{k=1}^{c} \left(\frac{1}{n_{\cdot k}} \sum_{\substack{s,s'\in\tau_x \\ s<s'}} \frac{-(n_{s'k}n_{s\cdot} - n_{s'\cdot}n_{sk})^2}{n_{s'\cdot}n_{s\cdot}} \right)$$

al realizarse la poda de un subárbol cualquiera T_x de T (Pérez Prados y otros, 1994) y en consecuencia $\Delta\chi^2 \leq 0$, y en nuestro caso $\Delta\chi^2 < 0$, ya que es por hipótesis $\Delta\chi^2 \neq 0$.

Pero esta reducción que se produce en χ^2 deberá ser en todo caso inferior al valor inicial del estadístico, en consecuencia, $\Delta\chi^2 \geq -\chi^2$, luego:

$$\frac{1+\dfrac{\chi^2}{n}}{\dfrac{\Delta\chi^2}{n}} + 1 \leq \frac{1+\dfrac{\chi^2}{n}}{-\dfrac{\chi^2}{n}} + 1 \leq 0$$

Por lo tanto: $\Delta(\mathrm{CP})^2 \leq 0$, y puesto que:

$$\Delta(\mathrm{CP}) = \frac{\Delta(\mathrm{CP})^2}{(\mathrm{CP}) + (\mathrm{CP})'}$$

y el coeficiente de contingencia es siempre positivo se tiene que $\Delta(\mathrm{CP}) \leq 0$.

■

Proposición 2

El valor del coeficiente de contingencia para el árbol $T_{máx}$ viene dado por:

(2) $$(\mathrm{CP})_{T_{\mathrm{máx}}} = \sqrt{1 - \frac{1}{c}}$$

Demostración

Este resultado se obtiene directamente de la definición del coeficiente de contingencia, teniendo en cuenta que $(\chi^2)_{T_{máx}} = (c-1)n$.

$$(\mathrm{CP})_{T_{máx}} = \sqrt{\frac{\frac{(c-1)n}{n}}{1+\frac{(c-1)n}{n}}} = \sqrt{\frac{c-1}{1+(c-1)}} = \sqrt{1-\frac{1}{c}}$$

■

4. CRITERIO DE UTILIDAD RELATIVA SEGÚN COEFICIENTES DE ASOCIACIÓN

4.1. Definiciones y propiedades fundamentales

Dado un árbol cualquiera T se considera la cantidad $S_\alpha(T)$ definida por la siguiente expresión:

$$S_\alpha(T) = \mathrm{CP}(T) - \alpha \mathcal{M}(T) \tag{3}$$

donde $\mathrm{CP}(T)$ es el coeficiente de contingencia; α es un número real positivo o nulo y $\mathcal{M}(T)$ la simplicidad.

Obsérvese que $S_\alpha(T)$ combina una medida de la utilidad de T con su simplicidad.

Por otra parte, se conoce también que el coeficiente de Tschuprow depende tanto de χ^2 como de la simplicidad de T. Por lo tanto podrán combinarse ambas medidas para obtener un árbol óptimo.

Definición 1

Se dice que T' es un subárbol óptimamente podado de T si el valor de la cantidad criterio a él asociado es el mayor entre los correspondientes a todos los subárboles de T.

Definición 2

Dados T_1 y T_2 dos árboles cualesquiera obtenidos a partir de I se dice que T_1 es mejor que T_2 según el coeficiente de Tschuprow, si se verifica que: $T(T_1) > T(T_2)$

Definición 3

Dos árboles cualesquiera T_1 y T_2 obtenidos a partir de I se dice que son equivalentes según el coeficiente de Tschuprow si se verifica que: $T(T_1) = T(T_2)$

Definición 4

Dado un conjunto A de subárboles podados de un árbol T, se dice que $\{T_e\}$ es el conjunto de mejores árboles de A según el coeficiente de Tschuprow, si para todos su elementos se verifica que:

$$T(T_e) \geq T(T_i) \qquad \forall\, T_i / T_i \in A$$

Propiedades

❶ Para el caso particular del árbol trivial T_1 que contiene únicamente el nodo raíz x_0, se tiene:

$$S_\alpha(T_1) = \mathrm{CP}(T_1) - \alpha \mathcal{M}(T_1) = -\alpha$$

❷ Si el árbol T corresponde al árbol $T_{\text{máx}}$ en el sentido de que todos los elementos de cada nodo terminal pertenecen a una misma modalidad de Y, entonces:

$$S_\alpha(T_{\text{máx}}) = \mathrm{CP}(T_{\text{máx}}) - \alpha \mathcal{M}(T_{\text{máx}}) = \sqrt{1 - \frac{1}{c}} - \alpha \,\mathrm{card}\, \mathfrak{T}_{\text{máx}}$$

Según el resultado obtenido en la proposición 1, el máximo valor del coeficiente de contingencia se alcanza en el árbol $T_{\text{máx}}$, por ser máximo χ^2 en dicho árbol; en consecuencia, cualquier árbol obtenido a partir de $T_{\text{máx}}$ mediante un proceso de división de uno o varios de sus nodos terminales reducirá el valor de la utilidad $S_\alpha(T)$, salvo en el caso $\alpha = 0$ en el cual no se producirá ninguna modificación. En este último caso:

$$S_0(T) = \mathrm{CP}(T) \leq \sqrt{1 - \frac{1}{c}}$$

Para cada α será un árbol óptimamente podado de $T_{\text{máx}}$ según este criterio, cualquiera que maximice el valor de $S_\alpha(T)$.

Dado un árbol cualquiera T, si mediante un proceso de poda del nodo x, se obtiene el árbol T', la variación producida en la utilidad $S_\alpha(T)$ será:

$$\begin{aligned} \Delta_x S_\alpha(T) &= S_\alpha(T') - S_\alpha(T) = \mathrm{CP}(T') - \mathrm{CP}(T) + \alpha(\mathrm{card}\, \mathfrak{T} - \mathrm{card}\, \mathfrak{T}') = \\ &= \Delta_x \,\mathrm{CP}(T) + \alpha(\mathrm{card}\, \mathfrak{T}_x - 1) \end{aligned} \tag{4}$$

donde $\Delta_x \text{CP}(T) \leq 0$ cualquiera que sea el nodo x, de acuerdo con los resultados de la proposición 1.

En consecuencia, el incremento producido en la utilidad $S_\alpha(T)$ al introducirse la poda de un nodo x, consta de dos términos, el primero de ellos es negativo o nulo y el segundo positivo, salvo en el caso $\alpha = 0$ en el cual se anula.

4.2. Sucesión de árboles según los valores de α

Se trata de obtener ahora una sucesión de árboles que optimicen, para los distintos valores de α, la utilidad $S_\alpha(T)$. Las demostraciones de los lemas se hallan en el apéndice.

Lema 1

«El único árbol óptimamente podado de $T_{\text{máx}}$ según $S_0(T)$ es el propio $T_{\text{máx}}$».

Por lo tanto, de acuerdo con este lema, el primer elemento de la secuencia que queremos obtener es: $T_0 = T_{\text{máx}}$.

Proceso de obtención de la secuencia de subárboles óptimamente podados del árbol máximo

- ***Paso inicial*** $\alpha = 0$

  En este caso según (4)

$$\Delta_x S_0(T) = \Delta_x \text{CP}(T) \leq 0$$

cualquiera que sea el nodo x elegido. Por tanto serán árboles óptimamente podados de $T_{\text{máx}}$ para $\alpha = 0$ todos aquellos que correspondan al mismo valor del coeficiente de contingencia que dicho árbol $T_{\text{máx}}$.

- ***Pasos sucesivos*** $\alpha > 0$ y creciente

  A partir de T_0 cualquier poda que se introduzca dará lugar a una reducción del valor del estadístico χ^2 y, como consecuencia, a una reducción del valor del coeficiente de contingencia; pero como α es creciente, el término positivo de $\Delta_x S_\alpha(T)$ crece, hasta que al llegar a un determinado valor de α que se indica α_1, el valor de $\Delta_x S_{\alpha_1}(T)$ es nulo, lo que indica que para ese valor α_1 existe al menos una poda posible tal que el nuevo árbol posee una utilidad S_{α_1} igual que la que corresponde a T_0; por lo cual, para α_1, cualquiera de los subárboles obtenidos mediante estas podas serán subárboles óptimamente podados. Entre ellos se elige T_1 como el mejor de acuerdo con el criterio de Tschuprow.

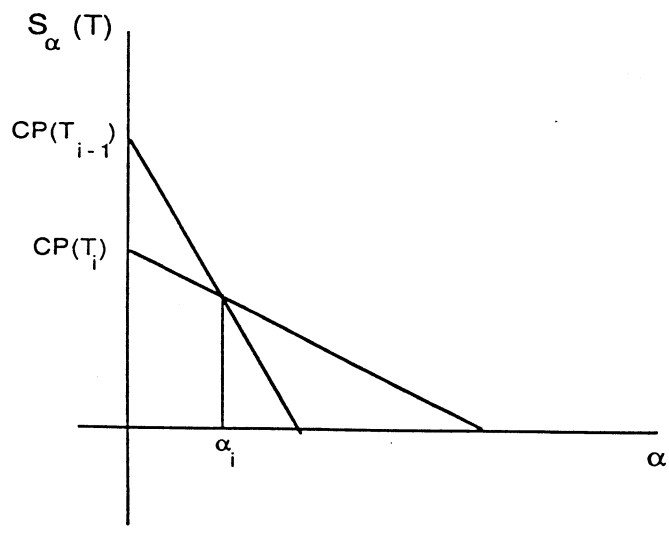
Se continúa con el crecimiento de α hasta llegar a un nuevo valor α_2 para el cual el valor de la utilidad $S_{\alpha_2}(T_1)$ coincide con el valor de la misma correspondiente a otro u otros árboles podados de $T_{\text{máx}}$. Entre éstos se elige el mejor según el criterio de Tschuprow y éste será el nuevo árbol de la sucesión.

Repitiendo el proceso se obtendría la sucesión completa de árboles $T_0 = T_{\text{máx}}$, $T_1, T_2, \ldots, T_r = \{x_0\}$ todos ellos podados de $T_{\text{máx}}$, junto con los valores de α asociados.

Lema 2

«Los árboles de la sucesión $T_0 = T_{\text{máx}}$, $T_1, T_2, \ldots, T_r$ anterior verifican: $\text{card}\,\tau_i < \text{card}\,\tau_{i-1} \quad \forall\, i = 1, \ldots, r$».

Puede verse la interpretación gráfica de este lema en la figura siguiente, teniendo en cuenta que la pendiente de la recta $S_\alpha(T) = \text{CP}(T) - \alpha\,\text{card}\,\tau$ es $m = -\text{card}\,\tau$.



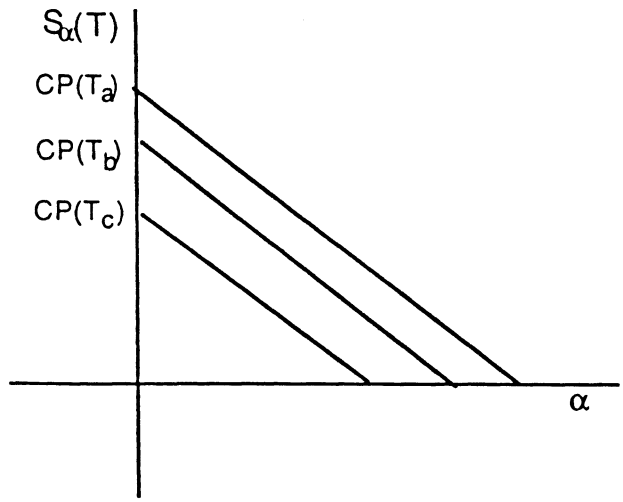
Por lo tanto, de acuerdo con este criterio de utilidad relativa según coeficientes de asociación, en el que se trata de obtener la sucesión de subárboles óptimamente podados de $T_{\text{máx}}$ para valores crecientes de α, será preciso en primer lugar, calcular el conjunto de todos los subconjuntos podados de $T_{\text{máx}}$ y sus correspondientes valores del coeficiente de contingencia CP(T) así como el número de sus nodos terminales $\text{card}\,\tau$, valor este último que verificará $1 \leq \text{card}\,\tau \leq \text{card}\,\tau_{\text{máx}}$; a partir de estos resultados se realiza una partición del conjunto de subárboles en función del valor de

card$\mathbf{T}$. Esta partición constará por lo tanto de card$\mathbf{T}_{máx}$ elementos de la forma:

$$C^i = \{T \leq \mathrm{p}T_{máx} / \mathrm{card}\,\mathbf{T} = c^i\}$$

donde $T \leq \mathrm{p}T_{máx}$ indica que T se obtiene mediante un proceso de podas realizadas a partir de $T_{máx}$.

Las representaciones $(\alpha, S_\alpha(T))$ para los árboles de C^i pueden observarse en la figura siguiente:



Lema 3

«En cada uno de los subconjuntos C^i de subárboles podados de $T_{máx}$ anteriores, el conjunto de árboles de dicho conjunto que pueden ser subárboles óptimamente podados, o bien está formado por un único elemento o bien los elementos que lo forman son de igual utilidad para cualquier valor de α y equivalentes según Tschuprow».

Se representa T^i el elemento de C^i que verifica las condiciones para poder ser subárbol óptimamente podado de $T_{máx}$, que evidentemente es el subárbol óptimo entre los del conjunto C^i.

De acuerdo con el lema anterior la selección de subárboles óptimamente podados de $T_{máx}$ deberá realizarse en el conjunto de subárboles T^i anteriores, conjunto que contará con un número de elementos igual a card$\mathbf{T}_{máx}$ y que se representa $\mathcal{S}$. En consecuencia, se tienen card$\mathbf{T}_{máx}$ ecuaciones de la forma:

$$S_\alpha(T^i) = \mathrm{CP}(T^i) - \alpha\,\mathrm{card}\,\mathbf{T}^i \quad T^i \in \mathcal{S} \quad i = 1, 2, \ldots, \mathrm{card}\,\mathbf{T}_{máx}$$

Para $\alpha = 0$ se tiene que:

$$S_0(T_{\text{máx}}) = \text{CP}(T_{\text{máx}}) = \sqrt{1 - \frac{1}{c}}$$

$$T_0 = T_{\text{máx}}$$

a partir de este árbol se calculan los valores de α para los cuales se verifica:

$$S_\alpha(T^i) = S_\alpha(T_{\text{máx}})$$

para cada uno de los árboles $T^i \in \mathcal{S}$. Esto significa:

$$\text{CP}(T^i) - \alpha^{i0} \,\text{card}\,\mathcal{T}^i = \sqrt{1 - \frac{1}{c}} - \alpha^{i0} \,\text{card}\,\mathcal{T}_{\text{máx}} \rightarrow$$

$$\rightarrow \alpha^{i0} = \frac{\sqrt{1 - \frac{1}{c}} - \text{CP}(T^i)}{\text{card}\,\mathcal{T}_{\text{máx}} - \text{card}\,\mathcal{T}^i}$$

El mínimo de ellos será α_1 y T_1 el árbol correspondiente del conjunto $\mathcal{S}$, que será el segundo elemento de la sucesión de árboles óptimos.

$$\alpha_1 = \text{mínimo} \left\{ \alpha^{i0} / \alpha^{i0} = \frac{\sqrt{1 - \frac{1}{c}} - \text{CP}(T^i)}{\text{card}\,\mathcal{T}_{\text{máx}} - \text{card}\,\mathcal{T}^i} \right\}$$

$$T_1 \in \mathcal{S} / S_{\alpha_1}(T_1) = S_{\alpha_1}(T_0)$$

Se opera análogamente con T_1, para la obtención de α_2 y T_2, teniendo en cuenta ahora que según el lema 2 se verifica que $\text{card}\,\mathcal{T}_2 < \text{card}\,\mathcal{T}_1$.

$$S_\alpha(T^i) = S_\alpha(T_1)$$

para cada uno de los árboles $T^i \in \mathcal{S} / \text{card}\,\mathcal{T}^i < \text{card}\,\mathcal{T}_1$ y por lo tanto:

$$\text{CP}(T^i) - \alpha^{i1} \,\text{card}\,\mathcal{T}^i = \text{CP}(T_1) - \alpha^{i1} \,\text{card}\,\mathcal{T}_1 \rightarrow$$

$$\rightarrow \alpha^{i1} = \frac{\text{CP}(T_1) - \text{CP}(T^i)}{\text{card}\,\mathcal{T}_1 - \text{card}\,\mathcal{T}^i}$$

El mínimo de ellos será α_2, siendo T_2 su árbol correspondiente.

$$\alpha_2 = \text{mín} \left\{ \alpha^{i1} / \alpha^{i1} = \frac{\text{CP}(T_1) - \text{CP}(T^i)}{\text{card}\,\mathcal{T}_1 - \text{card}\,\mathcal{T}^i} \right\}$$

$$T_2 \in S/S_{\alpha_2}(T_2) = S_{\alpha_2}(T_1)$$

Continuando con el proceso se llega a la obtención del árbol T_r formado únicamente por el nodo raíz. La forma general del proceso en su fase j–ésima será:

$$S_\alpha(T^i) = S_\alpha(T_{j-1}) \quad T^i \in S/\text{card}\,\tau^i < \text{card}\,\tau_{j-1} \rightarrow$$

$$\rightarrow \quad \text{CP}(T^i) - \alpha^{i(j-1)}\,\text{card}\,\tau^i = \text{CP}(T_{j-1}) - \alpha^{i(j-1)}\,\text{card}\,\tau_{j-1} \rightarrow$$

$$\rightarrow \quad \alpha^{i(j-1)} = \frac{\text{CP}(T_{j-1}) - \text{CP}(T^i)}{\text{card}\,\tau_{j-1} - \text{card}\,\tau^i}$$

$$\alpha_j = \text{mín}\left\{\alpha^{i(j-1)}/\alpha^{i(j-1)} = \frac{\text{CP}(T_{j-1}) - \text{CP}(T^i)}{\text{card}\,\tau_{j-1} - \text{card}\,\tau^i}\right\}$$

$$T_j \in S/S_{\alpha_j}(T_j) = S_{\alpha_j}(T_{j-1})$$

Es decir, que para ese valor de α ambos subárboles corresponden a la misma utilidad.

Si en alguno de los casos el árbol T_j que verifica la condición anterior para α_j no es único, se elige entre ellos el de menor número de nodos terminales, ya que para $\alpha_j < \alpha < \alpha_{j+1}$ este árbol será el óptimo.

Como consecuencia de todo ello, si se consideran todos los valores de $\alpha \geq 0$ se obtiene una secuencia de árboles óptimos, de forma que para cada intervalo (α_i, α_{i+1}) el subárbol podado de $T_{\text{máx}}$ que es óptimo de acuerdo con este criterio de utilidad es T_i. Según se ha realizado la construcción de la sucesión de árboles, para los valores de α_i se verifica:

$$S_{\alpha_i}(T_{i-1}) = S_{\alpha_i}(T_i)$$

con lo cual para los valores de α correspondientes a los extremos de los intervalos (α_i, α_{i+1}), se elige entre los dos subárboles correspondientes al mismo valor de la utilidad S_α, el que sea mejor según el criterio de Tschuprow. En estas condiciones se puede considerar una asociación entre los valores de α y los árboles óptimos T en la siguiente forma:

$\alpha = 0 \rightarrow T = T_0 = T_{\text{máx}}$

$\alpha \in (\alpha_i, \alpha_{i+1}) \rightarrow T = T_i$ término i–ésimo de la sucesión $T_0, T_1, \ldots, T_r$

$\alpha = \alpha_i \rightarrow T = T_{i-1}$ si $T(T_{i-1}) > T(T_i)$

$T = T_i$ si $T(T_{i-1}) < T(T_i)$

$\alpha > \alpha_r \rightarrow T = T_r = \{x_0\}$

4.3. Selección de un árbol entre los de la sucesión

A partir de los resultados anteriores, si el valor del parámetro α es conocido, el árbol podado de $T_{\text{máx}}$ óptimo de acuerdo con los criterios S_α y T queda perfectamente determinado. Esto, sin embargo, no ocurre si el valor de α no es conocido. En este caso, es necesario seleccionar uno entre los árboles de la sucesión $T_0, T_1, T_2, \ldots, T_r$. Para ello es preciso, en primer lugar, conocer la existencia o no de posibles limitaciones bien en cuanto a la importancia relativa de las medidas de calidad utilizadas, o bien en cuanto a la forma del árbol; la presencia o no de condiciones puede dar lugar a los siguientes casos:

❶ Se presentan limitaciones en cuanto a los valores de α, en cuyo caso la selección se realizará únicamente entre los árboles de la sucesión $T_0, T_1, T_2, \ldots, T_r$ que correspondan a esos valores de α, pudiendo incluso darse el caso particular en que los valores posibles de α lleven asociado un único árbol, con lo cual el árbol óptimo queda directamente determinado. En caso de existir varios, se selecciona uno, como en el caso general, pero entre los de la subsecuencia obtenida.

❷ Se presentan limitaciones en cuanto al número de nodos terminales del árbol, que no deben exceder un valor determinado. En este caso, se produce una reducción en el número de árboles posibles de la sucesión de subárboles óptimos de $T_{\text{máx}}$. La situación es por lo tanto, desde el punto de vista de la selección, análoga a la del apartado anterior.

❸ No existen limitaciones y, en consecuencia, se presenta el caso general, en el que hay que seleccionar un árbol entre los de la sucesión $T_0, T_1, T_2, \ldots, T_r$. Para ello, puede considerarse el valor del coeficiente de Tschuprow. Se elige entre los árboles óptimos según el criterio de utilidad que combina el coeficiente de contingencia con la simplicidad, el mejor según el coeficiente de Tschuprow.

5. CONCLUSIONES

Teniendo en cuenta los resultados anteriores puede concluirse la importancia de los coeficientes de asociación en la selección del árbol óptimo.

Partiendo del criterio que combina linealmente el coeficiente de contingencia con la simplicidad, según la cantidad criterio $S_\alpha(T) = \text{CP}(T) - \alpha\mathcal{M}(T)$, la sucesión de árboles óptimos para valores crecientes de $\alpha \geq 0$ se inicia con el árbol T_0, máximo en el sentido de que en cada nodo terminal todos los elementos pertenecen a la misma

modalidad de la variable criterio Y y entre ellos el que tenga un menor número de nodos terminales. A partir de él, mediante sucesivas podas pueden obtenerse subárboles podados de T_0 con número de nodos terminales decrecientes; agrupando éstos según el número de nodos terminales, en cada uno de los conjuntos formados, únicamente un árbol, el que corresponde al mayor valor del coeficiente de contingencia, puede ser subárbol óptimamente podado. Considerando estos árboles, a medida que los valores de α crecen, de acuerdo con la cantidad criterio, se determina para cada valor de α el que corresponde al mayor valor de la utilidad, teniendo en cuenta que el número de nodos terminales de cada árbol de la sucesión $T_0, T_1, T_2, \ldots, T_r$ es necesariamente menor que el que le corresponde al árbol que le precede en la sucesión.

Obtenida la sucesión de árboles, teniendo en cuenta las posibles restricciones existentes, bien en cuanto al valor del parámetro o del número de nodos terminales del árbol, se selecciona el que proporciona un mayor valor para el coeficiente de Tschuprow.

Entre las numerosas aplicaciones para las que los resultados anteriores son válidos está en estudio una referida a índices de ocupación, en particular en el campo de la enseñanza. Considerando como variables tanto las calificaciones en sucesivos cursos escolares, como las que hacen referencia a otras actividades extraescolares, familiares, etc. medidas sobre alumnos de una etapa escolar en sucesivos años, y conocidos los resultados que esos alumnos obtienen y la «plaza» que en cada momento ocupan, pueden determinarse árboles de clasificación entre los que se seleccionará el óptimo según el método planteado, que permitirá preveer las necesidades para cursos posteriores y planificar sobre ellas. Podrán obtenerse también otras consecuencias que posibiliten una mejora de la calidad de enseñanza y orientación del alumnado.

APÉNDICE

Demostración lema 1

Sea T un árbol cualquiera podado de $T_{\text{máx}}$, será T un árbol óptimamente podado si verifica:

$$S_0(T) = S_0(T_{\text{máx}}) = \text{CP}(T_{\text{máx}}) = \sqrt{1 - \frac{1}{c}}$$

es decir, si $\Delta_x S_0(T_{\text{máx}}) = 0$.

Pero:

$$\Delta_x S_0(T_{máx}) = \Delta_x CP(T_{máx}) = \frac{1}{CP(T_{máx}) + CP(T)} \frac{\frac{\Delta\chi^2}{n}}{\left(1 + \frac{\chi^2}{n}\right)^2 + \frac{\Delta\chi^2}{n}\left(1 + \frac{\chi^2}{n}\right)}$$

Luego $\Delta_x S_0(T_{máx}) = 0$ exige $\Delta\chi^2 = 0$, pero por las propiedades de χ^2 se conoce que no existe ninguna poda posible en $T_{máx}$ que verifique esta condición, y en consecuencia no existe ningún árbol podado de $T_{máx}$ que mantenga el valor del coeficiente de contingencia, luego el único árbol óptimamente podado de $T_{máx}$ para el criterio $S_0(T)$ es el mismo $T_{máx}$.

■

Demostración lema 2

Si T_i y T_{i-1} son dos árboles cualesquiera de la sucesión de árboles óptimos anterior, se verifica que:

$$S_\alpha(T_i) < S_\alpha(T_{i-1}) \quad \text{si } 0 \leq \alpha < \alpha_i$$

$$S_\alpha(T_i) = S_\alpha(T_{i-1}) \quad \text{si } \alpha = \alpha_i$$

$$S_\alpha(T_i) > S_\alpha(T_{i-1}) \quad \text{si } \alpha > \alpha_i$$

Considerando el caso particular $\alpha = 0$ se obtiene:

$$CP(T_i) < CP(T_{i-1})$$

Pero si $\alpha = \alpha_i > 0 \rightarrow$

$$\rightarrow \quad CP(T_i) - \alpha_i \operatorname{card} \mathbf{T}_i = CP(T_{i-1}) - \alpha_i \operatorname{card} \mathbf{T}_{i-1} \rightarrow$$

$$\rightarrow \quad \alpha_i = \frac{CP(T_i) - CP(T_{i-1})}{\operatorname{card} \mathbf{T}_i - \operatorname{card} \mathbf{T}_{i-1}} > 0;$$

en consecuencia:

$$\operatorname{card} \mathbf{T}_i - \operatorname{card} \mathbf{T}_{i-1} < 0 \rightarrow \operatorname{card} \mathbf{T}_i < \operatorname{card} \mathbf{T}_{i-1}$$

■

Demostración lema 3

Sean T_1 y T_2 dos subárboles de C^i, por tanto:

$$S_\alpha(T_1) = CP(T_1) - \alpha \operatorname{card} \mathbf{T}_1 = CP(T_1) - \alpha c^i$$

$$S_\alpha(T_2) = CP(T_2) - \alpha \operatorname{card} \mathbf{T}_2 = CP(T_2) - \alpha c^i$$

Luego si para algún α_1 es T_1 un subárbol óptimamente podado de $T_{máx}$:

$$S_{\alpha_1}(T_1) \geq S_{\alpha_1}(T_2) \rightarrow \text{CP}(T_1) \geq \text{CP}(T_2) \rightarrow$$

$$\rightarrow S_\alpha(T_1) \geq S_\alpha(T_2) \quad \text{cualquiera que sea } \alpha$$

es decir, pueden presentarse dos posibilidades:

❶ $S_\alpha(T_1) > S_\alpha(T_2)$ cualquiera que sea α, con lo cual T_2 no puede ser un subárbol óptimamente podado de $T_{máx}$, es decir, T_1 es el único subárbol óptimamente podado de $T_{máx}$ entre los de C^i.

❷ $S_\alpha(T_1) = S_\alpha(T_2)$ con lo cual para cualquier valor de α la utilidad de T_1 y T_2 es la misma. Además:

$$T(T_1) = \frac{\chi^2(T_1)}{n\sqrt{(c^i-1)(c-1)}} = \frac{\chi^2(T_2)}{n\sqrt{(c^i-1)(c-1)}} = T(T_2)$$

Luego T_1 y T_2 son equivalentes según el criterio de Tschuprow.

■

BIBLIOGRAFÍA

[1] **Breiman, L., Friedman, J.H., Olshen, R.A.** y **Stone, Ch.J.** (1984). *Classification and Regression Trees.* Wadsworth & Brooks. Monterey, California.

[2] **Ciampi, A., Chang, C.H., Hogg, S.** y **McKineey, S.** (1987). *Recursive partition: A versatile method for exploratory data analysis in biostatistics.* In Proceedings from Joshi Festschrift, G. Umphrey (ed), 23–50. Amsterdam: Nort-Holland.

[3] **Ciampi, A.** (1989). «Generalized Regression Trees». *Computational Statistics and Data Analysis*, **12, 1**, 732–764.

[4] **Cuesta, P.** (1989). *Inducción en bancos de datos cualitativos.* Tesis Doctoral. Facultad de Matemáticas. Universidad Complutense de Madrid.

[5] **Goodman, L.** y **Kruskal, W.** (1954). «Measures of Association for Cross Classifications». *JASA*, **49**, 732–764.

[6] **Hartigan, J.A.** (1975). *Clustering Algorithms.* Wiley Publication.

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

[8] **Munduate, A.** (1993). *Cuestiones notables en la construcción y comparación de árboles de decisión*. Tesis Doctoral. Departamento de Métodos Estadísticos. Universidad Pública de Navarra.

[9] **Pérez Prados, A., Munduate del Río, A.** y **Cano Sevilla, F.J.** (1994). «El estadístico χ^2 en la selección del árbol óptimo (I): Proceso de poda». *Cuadernos de Bioestadística y sus Aplicaciones Informáticas*, **12, 1**, 5–17.

[10] **Pérez Prados, A., Munduate del Rio, A.** y **Cano Sevilla, F.J.** (1994). «El estadístico χ^2 en la selección del árbol óptimo (II): Criterio de Selección». *Cuadernos de Bioestadística y sus Aplicaciones Informáticas*, **2, 1**, 18–38.

[11] **Quinlan, J.R.** (1986). «Induction of Decision Trees». *Machine Learning*, **1**, 81–106.

[12] **Quinlan, J.R.** (1988). «Decision trees and multi-valued attributes». *Machine Intelligence*, **11**, 305–319.

ENGLISH SUMMARY:

CRITERION FOR THE SELECTION OF AN OPTIMUM TREE CONSIDERING ASSOCIATION COEFFICIENTS OBTAINED FROM THE χ^2

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

There are various recognised methods for the construction of decision trees, starting from a set of data for $\{V^j\}$ variables and the Y variable related to the previous series, and defined on the learning group Y as a whole. Furthermore, the validity of association coefficients obtained from the χ^2 statistic (contingency coefficient, Tschuprow's coefficient and Cramer's coefficient) is also recognised for measuring the usefulness of a tree as a predictor of the Y variable criterion.

Taking this into account, we propose a criterion for selecting an optimum tree, considering a complexity criterion which combines the contingency coefficient whit the simplicity of the tree measured by the number of its terminal nodes.

Therefore, starting from the variation that a pruning process produces in the value of the statistic χ^2 (Pérez Prados *et al.*, 1994), the modification that this induces in the contingency coefficient is studied. The result obtained indicates that under no circumstances does this coefficient increase as a result a pruning process, the maximum value corresponding to the $T_{\text{máx}}$ tree, where in each of the terminal nodes all the elements belong to the same Y modality.

This behaviour of the contingency coefficient being known, the following complexity criterion is considered:

$$S_\alpha(T) = \text{CP}(T) - \alpha\mathcal{M}(T)$$

$\text{CP}(T)$ being the contingency coefficient, α a real positive or null number and $\mathcal{M}(T)$ the simplicity obtained based on the number of terminal nodes. A method of obtaining an optimum tree for each of the different values of parameter α is developed in accordance with this complexity criterion $S_\alpha(T)$ together with Tschuprow's coefficient. By establishing an α value any subtree which maximises the $S_\alpha(T)$ value would be an optimally pruned one.

The variation produced in the complexity criterion through the process of pruning the x node of the T tree is:

$$\Delta_x S_\alpha(T) = \Delta_x \text{CP}(T) + \alpha(\text{card}\,\mathbf{T}_x - 1)$$

where $\Delta_v CP(T)$ is the variation produced in the contingency coefficient. It therefore consists of two terms, the first is negative or null and the second positive, except in the case of $\alpha 00$, in which it is cancelled.

In accordance with this result a series of optimally pruned subtrees $T_{máx}$ is obtained for increasing values of α. For $\alpha = 0$ the only optimally pruned subtree $T_{máx}$ is itself, so T_0 is the first element of the succession. On the basis of this element, any pruning done will lead to a reduction in the contingency coefficient value. Furthermore, by considering growing α values the positive term of $\Delta_v S_\alpha(T)$ grows. Therefore, on reaching a certain α level (shown as α_1), the value of $\Delta_v S_{\alpha_1}(T)$ is null. This indicates that there is at least one possible pruning for α_1, so the maximum subtree. Among the subtrees which verify this condition, which will all be optimally pruned subtrees $T_{máx}$ for α_1, the one which corresponds to the highest value of Tschuprow's coefficient is chosen.

The growth of α is continued until a new value (α_2) is reached, for which the value of the $S_\alpha(T)$ utility coincides with the value of the one corresponding to another or other $T_{máx}$ pruned trees. The best is chosen among theses according to Tschuprow's criterion and this will be the new tree of the succession.

Repeating the process, the following succession of trees is obtained: $T_0, T_1, T_2, \ldots, T_r$ (all pruned from $T_{máx}$), together with the associated α value. These trees have a decreasing number of terminal nodes, T_r being the one consisting exclusively of the root node of $T_{máx}$.

If, by taking $T_{máx}$, successive prunings determine all the possible pruned subtrees, and grouping them according to the number of terminal nodes they have, in each of the sets created only one tree (corresponding to the higher contingency coefficient) can be the optimally pruned subtree.

Considering these trees, and bearing in mind that in the $T_0, T_1, T_2, \ldots, T_r$ succession, $\text{card}\,\mathbf{T}_i < \text{card}\,\mathbf{T}_{i-1}\ \forall i = 1, \ldots, r$ is determined for growing α values the tree corresponding to a higher utility value for each α value is necessarily lower than the one corresponding to the preceding tree.

If the value of the α parameter is known, the optimum tree is the one from the succession to which this value corresponds, otherwise one is selected from among all the trees, taking the value of Tschuprow's coefficient into account for each one.