8.3. ANÁLISIS DE CLUSTER Y CART

Álvaro Roel
aroel@tyt.inia.org.uy
Instituto Nacional de Investigación Agropecuaria (INIA), Uruguay

ANÁLISIS DE CLUSTERS (GRUPOS)

El término análisis de clusters agrupa a una serie de algoritmos de clasificación. Una pregunta que muchas veces se realizan investigadores de diferentes áreas es: ¿cómo organizar bases de datos en estructuras que tengan una utilidad? Para este tipo de preguntas es que comúnmente se utiliza el análisis de grupos o clusters.

Básicamente, el análisis de "clusters" consiste en una serie de algoritmos que lo que realizan es el agrupamiento de datos en grupos. Normalmente, estas técnicas son utilizadas cuando no tenemos a priori ninguna hipótesis referente al problema en estudio y se está en la etapa de exploración de datos.

El rol de estas técnicas es el agrupamiento de los datos en "clusters" teniendo en cuenta algún tipo de medición de similitud dentro de los grupos conformados y de diferencia entre los mismos. El algoritmo se optimiza buscando la conformación de diferentes grupos, de manera tal, de minimizar la varianza dentro de los grupos y maximizar la varianza entre los grupos, moviendo datos de un grupo a otro.

Una aplicación de esta técnica en agricultura de precisión ha sido en la delimitación de zonas de comportamiento productivos diferentes dentro de una chacra o lote (Roel y Plant, 2004a). Muchas veces sucede que, cuando se posee una serie de mapas de rendimiento de una misma chacra, el productor o técnico desea analizar cuales fueron las zonas de la chacra que rindieron en forma aceptable y cuales tuvieron un comportamiento deficiente. Esto, que tal vez se obtenga simplemente mirando una serie de mapas de rendimiento, puede llegar a ser muy complejo cuando se tiene comportamientos erráticos o serie muy grande de mapas. La delimitación de zonas con comportamientos productivos diferentes de una serie de mapas constituye el entendimiento de la variabilidad espacial y temporal del rendimiento dentro de la misma. Este es uno de los primeros pasos que deben estudiarse dentro de una chacra o lote cuando se intentan implementar estrategias de AP en la misma. Es por lo tanto fundamental, contar con un método objetivo que permita realizar estas delimitaciones y no puede quedar al "buen ojo" del investigador, por esto consideramos importante el uso de estas técnicas estadísticas. En este caso, se debe inicialmente proceder a la estandarización de los rendimientos de las diferentes zafras, de manera de poder hacer comparación entre ellos. Paso seguido, se procede a correr el algoritmo descrito. Hoy en día, la mayoría de los paquetes estadísticos cuentan con la opción del análisis de grupo o cluster.

CART

Los árboles de calcificación y regresión (CART, del inglés Classification And Regression Trees, Breiman et al. 1984) son modernas técnicas estadísticas que permiten tanto modelar como explorar la existencia de múltiples relaciones causa-efecto, tanto en el tiempo como en el espacio, dentro de una misma base de datos. A diferencia de las técnicas normalmente utilizadas en los análisis estadísticos tradicionales, donde lo que se intenta es buscar un modelo general de relación entre

variables explicativas y de respuesta, CART divide en forma sucesiva el espacio multidimensional generado por las variables explicativas entre zonas que son lo más uniformes posibles en términos de la variable de respuesta. En vez de identificar una sola estructura dominante en la base de datos, lo que comúnmente realizan la mayoría de las técnicas estadísticas tradicionales, CART está diseñado para trabajar con base de datos que puedan tener múltiples estructuras a diferentes escalas espaciales y temporales (Roel y Plant, 2004b). Esto, lo hace sumamente útil en el caso de las aplicaciones en la agricultura de precisión donde muchas veces dentro de una misma chacra o lote pueden existir una serie de variables explicando la variabilidad de rendimiento a diferentes niveles y zonas dentro de la misma.

CART, es hoy, uno de los métodos más utilizados en medicina para intentar relacionar la presencia de síntomas en pacientes con una serie muy amplia de factores predisponentes (historia clínica). Es también comúnmente utilizado, en aplicaciones meteorológicas; en estudios de biodiversidad de especies; en general, cuando existen base de datos que contengan una serie amplia de variables observadas.

CART es un método no-paramétrico, lo que de alguna manera flexibiliza el problema que muchas veces se plantea con los métodos paramétricos, que exigen una distribución normal de los datos y la presencia de un determinado nivel de homogeneidad de varianza entre las variables. El no cumplimiento de estas exigencias puede acarrear problemas a la hora de analizar los datos por métodos tradicionales.

Básicamente, CART funciona en base a un algoritmo de partición recurrente que considera una serie de variables explicativas $\mathbf{X_1}$, $\mathbf{X_2}$, ... $\mathbf{X_N}$, y una variable de respuesta Y. Si la variable Y es nominal u ordinal (por ejemplo, nivel de enmalezamiento alto, medio o bajo) se aplica el método de clasificación. En caso que la variable Y sea numérica continua (por ejemplo, rendimiento) se aplica el método de la regresión.

- **Paso 1.** De todas las variables **Xi**, busca el nivel de esa variable que le permite dividir la variable **Y** en los dos grupos más homogéneos entre sí y más diferente entre ellos.
- **Paso 2.** Para cada grupo, en caso de ser completamente homogéneo, finaliza el proceso, de lo contrario vuelve al Paso 1.

La manera que posee este método para estimar el error de clasificación es utilizando 9/10 del total de los datos y con el 1/10 restante verifica si son correctamente asignados, con el propósito de generar un nivel de probabilidad. Repite este procedimiento 10 veces con diferentes porciones de la base de datos.

CART utiliza una estructura de árbol de decisión para desplegar la relación entre las variables explicativas y la o las variables de respuesta. A modo de ejemplo, en la **Figura 8.17**, se presenta el árbol generado al aplicar este procedimiento a una base de datos de 125 puntos que son localidades en diferentes chacras de arroz ubicadas en la zona Este del Uruguay, donde se midieron una serie muy amplia de posibles variables explicativas y el rendimiento de arroz en cada uno de ellos. El objetivo de este trabajo era tener una ponderación objetiva de las variables que estaban incidiendo en las diferencias de rendimientos observadas en estas 125 localidades.

En la **Figura 8.17**, podemos observar que el promedio (AVG) de rendimiento de todas las localidades fue 5999 kg/ha, con un desvío de +/- 1158 kg/ha. El algoritmo identificó a la variable RIEGO como la primer variable que permitió generar dos grupos de puntos (nodos ó nodes): el primero de ellos, con 16 observaciones y un promedio de rendimiento de 4253 kg/ha; y el segundo, conformado por 109 observaciones y un promedio de rendimiento de 6255 kg/ha. A su vez, identificó el nivel de esta variable (riego > o < 2,5) que permitió formar estos dos grupos de rendimientos. Para este caso en particular, el valor de la variable riego variaba de 1 a 5, siendo

1 el peor riego y 5 el mejor riego. Estos grupos tienen la propiedad de que los valores de rendimiento **dentro** de cada uno de ellos son lo más similares entre sí, pero a su vez, lo más diferentes **entre** estos dos grupos. Es decir, que esta técnica permitió identificar que la variable riego al nivel de 2,5 permitía conformar dos grupos de rendimientos muy diferentes. Aquellas localidades cuyo riego era mejor que 2,5 tenían un rendimiento considerablemente superior (6255 kg/ha) que las que tuvieron una peor calidad de riego (4253 kg/ha).

Continuando con el análisis de la **Figura 8.17**, podemos observar que a su vez, dentro del grupo de localidades que estuvieron "bien regadas" la variable CONTROL, por control de malezas, permitió dividir este grupo de 109 localidades en dos, uno conformado por 41 localidades que presentaron un rendimiento promedio de 5758 kg/ha y el otro, conformado por 68 localidades con un rendimiento promedio de 6554 kg/ha. El nivel de control de malezas que determinó estas dos agrupaciones de localidades fue de 3,5. En este caso en particular, la variable control de malezas variaba entre 1 y 5, siendo 1 el peor control y 5 el mejor.

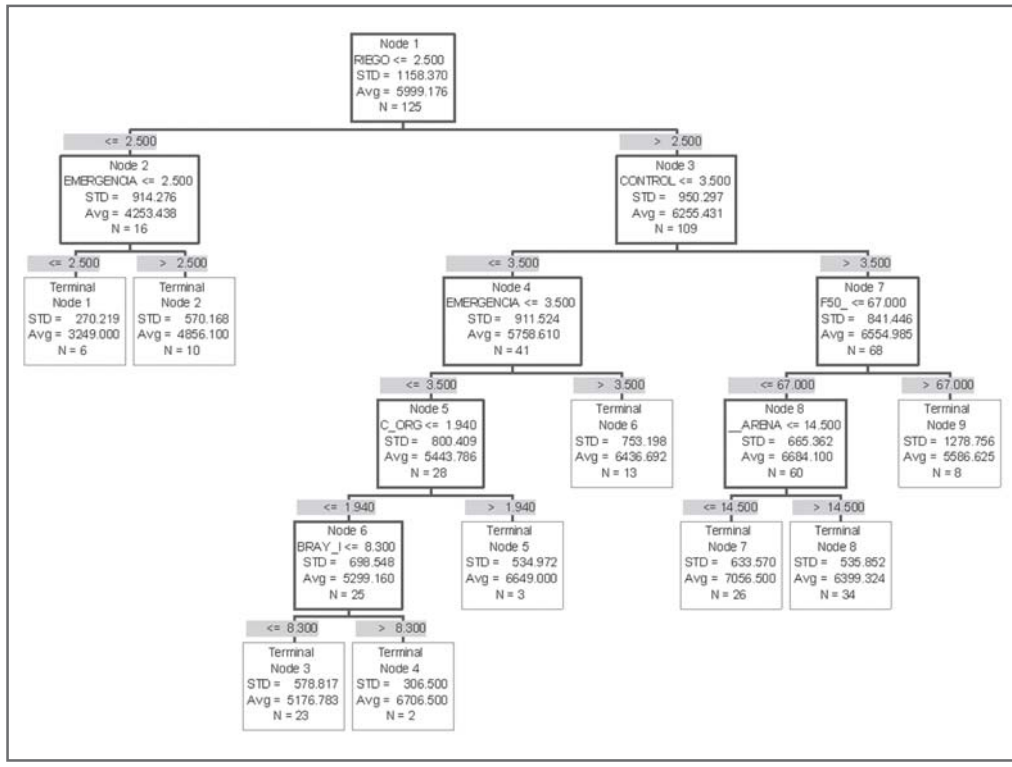

Figura 8.17: Modelo CART aplicado a 125 chacras de arroz en Uruguay.

Por lo tanto, si continuamos con el análisis del árbol generado por CART, para este caso en concreto, podemos observar que el conjunto de localidades con mayor rendimiento (n=26, 7056 kg/ha) dentro de estos 125 casos estudiados fueron aquellos que poseían la siguientes características, eran localidades con un nivel de riego superior al promedio (>2,5), poseían un nivel de control de malezas superior al promedio (>3,5), a su vez habían florecido (F50) antes de 67 días después del primero de enero y estaban ubicados en suelos con niveles de arena (ARENA) inferiores a 14,5. De la misma manera, podemos identificar que las localidades de menor rendimiento (n=6, 3249 kg/ha) fueron aquellas que presentaron, niveles de control de malezas inferiores al promedio (<3,5) y a su vez, niveles de EMERGENCIA menores a 2,5. Un aspecto importante a tener en cuenta, es que si toda esta información se encuentra georeferenciada, permite analizar si estas localidades se encuentran agrupadas en zonas específicas de lotes o chacras o si poseen ningún tipo de patrón espacial. Esta es una información importante a la hora de poder delimitar posibles zonas de manejo. Esta metodología nos permitiría, por lo tanto, determinar la variable y el valor de la misma, que deberían ser ajustados en cada una de estas potenciales zonas, en caso de que existieran.

REFERENCIAS

Breiman, L.; Friedman, J.H.; Olsen, R.A. y Stone, C.J. 1984. Classification and Regression Trees. Chapman & May, Inc., New York, NY. 358p.

Roel, A. y Plant, R. 2004a. Spatiotemporal Analysis of Rice Variability in Two California Fields. Agronomy Journal 96:77-90.

Roel, A. y Plant, R. 2004b. Factors Underlying Yield Variability in Two California Rice Fields. Agronomy Journal 96: 1481-1494.