Cluster Analysis

José María García
Director de Estadística y Procesamiento de Datum Internacional

El Análisis Cluster en la investigación de mercados es usado para la segmentación de mercados; comprensión del comportamiento del comprador (identificación de grupos de compradores homogéneos para analizar el comportamiento de cada grupo por separado); identificar oportunidades para nuevos productos, seleccionar mercados de prueba, reducción de datos con el fin de facilitar el manejo de la información.

El análisis Cluster es un conjunto de técnicas utilizadas para clasificar los objetos o casos en grupos homogéneos llamados conglomerados (clusters) con respecto a algún criterio de selección predeterminado. Los objetos dentro de cada grupo (conglomerado), son similares entre sí (alta homogeneidad interna) y diferentes a los objetos de los otros conglomerados o clusters (alta heterogeneidad externa). Es decir, que si la clasificación hecha es óptima, los objetos dentro de cada cluster estarán cercanos unos de otros y los cluster diferentes estarán muy apartados. Por eso, este análisis trata de establecer una realidad que no conseguiríamos visualizar a simple vista por ser una representación multidimensional de la realidad.

PASOS DEL ANÁLISIS DE CONGLOMERADOS

a) Formulación del problema

Lo más importante de la formulación del problema, es la selección de las variables en las que se basará la agrupación. El conjunto de variables seleccionado debe describir la similitud entre los objetos en términos relevantes para el problema de investigación de mercados. Estas variables se seleccionan en base a investigaciones anteriores, la teoría o una consideración de las hipótesis que se prueban.

b) Selección de una medida de similitud

Como el conglomerado agrupa objetos similares, se necesita una medida para evaluar las diferencias y similitudes entre objetos.

La Similaridad (similitud) es una medida de correspondencia o semejanza entre los objetos que van a ser agrupados. Lo más común es medir la equivalencia en términos de la distancia entre los pares de objetos. Así, los objetos con distancias reducidas entre ellos son más parecidos entre sí que aquellos con distancias mayores y se agruparán por lo tanto, dentro del mismo cluster.

Los tres métodos usados en la medición de la similitud son: las medidas de correlación y las medidas de distancia (usadas cuando se tienen variables métricas) y las medidas de asociación (usadas para variables categóricas).

c) Estandarización de datos

Como las medidas de distancia son sensibles a la diferencia de escalas o de magnitudes hechas entre variables es necesaria la estandarización de datos para evitar que las variables con una gran dispersión tengan un mayor efecto en la similaridad.

La forma de estandarización más común es restarle a cada observación la media de la variable y este resultado dividirlo entre su desviación estándar. Lo que se consigue con ello es eliminar las diferencias introducidas por la diferencias de escalas de las distintas variables (atributos) usados en el análisis.

Luego de seleccionar las variables y calcular las similaridades, se empieza con el proceso de agrupación, lo primero es seleccionar el algoritmo de agrupación para formar los grupos (clusters) y luego determinar el numero de grupos que se van a formar. Estos dos procedimientos dependerán de los resultados que se obtengan y la interpretación derivada de ellos.

Los dos tipos de procedimientos de agrupación son los jerárquicos y los no jerárquicos.

El conglomerado jerárquico se caracteriza por el desarrollo de una jerarquía o estructura de árbol (dendograma). De este modo, los clusters están formados solamente por la unión de los grupos existentes, así cualquier miembro de un cluster puede trazar su relación en un irrompible sendero que comenzaría con una simple relación. Los métodos jerárquicos pueden ser por Aglomeración o por División. Los métodos de conglomerados más usados en la investigación de mercados son el método de Enlace, método de varianza y el método Centroide.

Entre los métodos de conglomerados no jerárquicos más usados se conocen como Agrupación K medias e incluyen a los métodos Umbral secuencial, Umbral paralelo y la división para la optimización.

NÚMERO DE CONGLOMERADOS A CONSIDERAR

El problema para seleccionar el número de clusters, es que no existe un procedimiento de selección objetivo, una guía útil en el caso del análisis cluster jerárquico podría ser calcular distintas soluciones de aglomeración para después decidir entre las soluciones alternativas con ayuda de un criterio prefijado de antemano. Estas distancias reciben a menudo el nombre de medidas de variabilidad del error.

Para el análisis cluster no jerárquico, se puede trazar un gráfico que compare el número de grupos con la relación entre la varianza total de los grupos y la varianza entre los grupos. El punto del gráfico donde se presente un cambio marcado indicará el número apropiado de grupos.

Otro problema que puede presentarse es la presencia de grupos unipersonales, que podrían ser valores atípicos (outliers) no detectados en el proceso de depuración de la fuente de datos. Si se presentara este caso, el analista debe determinar si representa una estructura válida en la muestra o debe ser retirada de la misma, lo cual implicaría volver a definir los grupos.

INTERPRETACIÓN Y PERFIL DE LOS GRUPOS

Comprende el análisis de los centroides de grupo (valores medios de los objetos que contiene el grupo en cada una de las variables).

Los centroides permiten dar un nombre a cada grupo. El objetivo de esta etapa es, esencialmente, examinar la variación de los clusters para asignar etiquetas que describan de un modo veraz su naturaleza. Resulta útil elaborar el perfil de los grupos en términos de las variables utilizadas para el conglomerado, como los datos demográficos, los psicográficos, uso del producto, uso de los medios u otras variables.

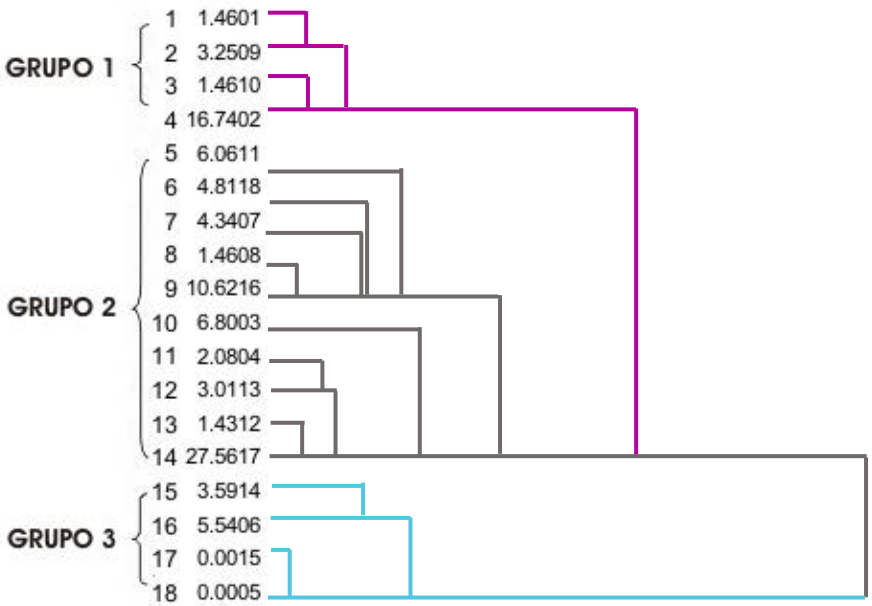
Generalmente, en investigación de mercados se utilizan variables no métricas, y como los métodos de clasificación clásicos han sido desarrollados para variables métricas, antes de hacer la clasificación es necesario convertir los datos en cuantitativos el cual puede hacerse usando un análisis factorial.

Para finalizar, se presenta como ejemplo un caso presentado por Ildefonso Grande y Elena Abascal (*), cuyo objetivo es conocer cómo pasan su tiempo libre los jóvenes de una comunidad. Para ello se realizó una encuesta a una muestra de 18 jóvenes, en la que se les pedía indiquen para cada actividad: discoteca, lectura, música, bares y deportes si la realiza con mucha frecuencia, poca o ninguna. Como variables ilustrativas se recogieron las características de sexo y niveles de estudio.

Al aplicar un análisis de correspondencia múltiple, se encuentra que el primer factor es un factor que opone aquellas personas que leen y escuchan música en grado de "mucho" y no acuden a los bares y discotecas frente al resto. El segundo factor opone los jóvenes que se dedican mucho a las actividades sociales y nada a la lectura a los que hacen un poco de todo. En ninguno de los factores se observan diferencias significativas entre hombres y mujeres pero sí respecto al nivel de estudio.

A partir de estos ejes factoriales se obtendrá una clasificación de los jóvenes de acuerdo a las actividades que prefieren en sus ratos de ocio. El objetivo entonces es buscar grupos de jóvenes que tienen un comportamiento semejante en sus ratos de ocio. Usando el procedimiento de clasificación jerárquico, se obtiene el siguiente dendograma de clasificación:

Dendograma de clasificación



Observando el dendograma se decidió obtener una partición de 3 clases. Los grupos 1 y 3 son muy homogéneos mientras que el segundo grupo es más heterogéneo.

Finalmente, se obtienen tres grupos de clasificación:

Primer Grupo - Los sociables: Todos los jóvenes de este grupo han elegido las modalidades discoteca y bares mucho, y lectura nada. Se caracterizan por tener estudios elementales y predominan las mujeres.

Segundo Grupo - Los adocenados: Formado por jóvenes que han elegido las modalidades discoteca poco, lectura poco, mucho y nada. y el 86% bar poco. No todos los de este grupo realizan actividades en ese grado (poco o nada). EL 80% de la clase ha elegido discoteca poco. En este grupo se encuentran las personas de estudios medios, aunque sólo sean el 60% del grupo. La mitad del mismo no escucha música, y el resto, poco.

Tercer Grupo - Los individualistas o intelectuales: Todos han elegido las modalidades lectura mucho, discoteca nada y un 75% bares nada. La mitad de la clase hace poco deporte. Toda esta clase tiene estudios superiores.

BIBLIOGRAFÍA

(*) Abascal, E. y Grande, I, "Fundamentos y Técnicas de Investigación Comercial", Esic, 1994.

Gondar, E. "Análisis Cluster",

Johnson, R. y Wichern, Dean, "Applied Multivariate Statistical Análisis" Prentice-Hall International, Inc, 1982..

Kinnear y Taylor "Investigación de Mercados", McGraw-Hill Interamericana S.A, 1998.