Selección de variables en la segmentación del mercado: Clustering o Biclustering?

*29 de julio de 2014*

Por [Joel Cadwell](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=https://www.r-bloggers.com/author/joel-cadwell/&usg=ALkJrhhjK8DUlcukVpxgmugYj_UaOwiodg)

[inShare](javascript:void(0);)5

(Este artículo fue publicado por primera vez en la investigación de [**mercado de**](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://joelcadwell.blogspot.com/2014/07/variable-selection-in-market.html&usg=ALkJrhgibioHlxZR4bwIHSmW6S0JlAYN4Q) compromiso, y amablemente contribuyó a [R-bloggers)](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=https://www.r-bloggers.com/&usg=ALkJrhjdW0V7hJGR20qt3Qxt1rhpxoYddA)

**130**

**COMPARTE**

[Compartir](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://www.facebook.com/sharer.php%3Fu%3Dhttps%253A%252F%252Fwww.r-bloggers.com%252Fvariable-selection-in-market-segmentation-clustering-or-biclustering%252F&usg=ALkJrhh7X6mzcwXA76iu7_Hv9VSPJfpEPw)[Pío](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=https://twitter.com/intent/tweet%3Ftext%3DVariable%2520Selection%2520in%2520Market%2520Segmentation%253A%2520Clustering%2520or%2520Biclustering%253F%26url%3Dhttps://www.r-bloggers.com/variable-selection-in-market-segmentation-clustering-or-biclustering/%26via%3DRbloggers&usg=ALkJrhitb5tjClcDhItWYpoEb4IyBkakjA)

**¿Tendrá esa segmentación con uno o dos modos?**

La matriz de datos para la segmentación del mercado viene a nosotros con dos modos, las filas son los consumidores y las columnas son variables. El agrupamiento utiliza todas las columnas para transformar la matriz de datos de dos modos (fila y columnas son diferentes) en una matriz de distancia de un modo (filas y columnas son iguales) directamente como en agrupación jerárquica o indirectamente como en k-means. La carga recae sobre el analista para ser juicioso en la selección de variables, ya que muy pocos se perderán las distinciones que deberían hacerse y demasiadas borrarán las distinciones que deben hacerse.

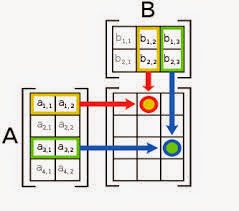
Vale la pena repetir este punto. La mayoría de las matrices de datos tienen dos modos con las filas y columnas que se refieren a diferentes tipos. Por el contrario, las matrices de correlación y distancia tienen sólo un modo con las filas y las columnas que se refieren a las mismas entidades. Aunque una segmentación de mercado coloca sus datos en una matriz bidireccional con las filas como consumidores y las columnas como variables, la intención es agrupar las filas e ignorar las columnas una vez que se han utilizado para definir las distancias entre las filas. Todas las columnas entran en ese cálculo de distancia, por lo que la selección de variables es tan importante en la segmentación del mercado.

Tal selección de variable de todo o nada puede parecer demasiado restrictiva dada la diversidad de productos en muchas categorías. No nos sorprendería descubrir que lo que diferencia un extremo de la categoría de productos no es lo que separa a los consumidores en el otro extremo. Los ejemplos son fáciles de encontrar. El uso de la tarjeta de crédito se puede dividir en los que pagan su cuenta por completo cada mes y ésos que pagan intereses sobre saldos pendientes. Los dos grupos buscan diferentes características de tarjeta de crédito, aunque hay una cierta superposición. Los viajeros de negocios están dispuestos a pagar por los beneficios que no les interesaría a los de vacaciones, una vez más, se encuentra al menos algo en común. Ejemplos adicionales pueden ser generados sin dificultad para muchas categorías de productos que contienen heterogeneidad sustancial tanto en sus usuarios como en sus ofertas.

Biclustering ofrece una alternativa de dos modos que permite definir diferentes clusters por diferentes variables. El "bi" se refiere a la agrupación conjunta de filas y columnas al mismo tiempo. Todas las variables permanecen disponibles para la selección por cualquier grupo, pero cada grupo ejerce su propia opción para incorporar o ignorar. Entonces, ¿por qué no se refieren a biclustering como clustering de dos modos o clustering simultáneo o co-clustering? [Lo hacen](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://en.wikipedia.org/wiki/Biclustering&usg=ALkJrhjORn7LtKd1VrgiuVkrFeaCJ3SsVA) . Los nombres proliferan cuando la misma técnica es redescubierta por diversas disciplinas o cuando se utilizan diferentes modelos y algoritmos. Como es de esperar, R ofrece muchas opciones (incluyendo [biclust](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://ro.uow.edu.au/commpapers/2506/&usg=ALkJrhijvILKjJFuHzlBblud4yAmQxilNg" \t "_blank) para biclustering basado en distancia). Sin embargo, me centraré en [NMF](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://nmf.r-forge.r-project.org/vignettes/NMF-vignette.pdf&usg=ALkJrhiZMZeNub8tfkYscFbg9f__4EDfsQ) para la factorización basada en biclustering, un paquete que ha demostrado ser útil con una serie de mis proyectos de investigación de marketing.

**Revisitando K-means como Factorización de Matriz**

Para muchos de nosotros el agrupamiento k-means está estrechamente asociado con nubes de puntos en el espacio bidimensional. Les pido que se olviden de ese diagrama de dispersión de puntos y que lo reemplacen con el siguiente cuadro de [multiplicación matricial](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=https://www.khanacademy.org/math/algebra2/alg2-matrices/matrix-multiplication-alg2/v/multiplying-a-matrix-by-a-matrix&usg=ALkJrhjSqoCfBgO_Q6xfXFJkM5A0yCdqLA) de [Wikipedia](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://en.wikipedia.org/wiki/Matrix_multiplication&usg=ALkJrhi714FjRXNJZKsqCDLXiAGW7wY5ew) :

[](https://i2.wp.com/2.bp.blogspot.com/-c-pkexFwF1w/U9bIZNk44xI/AAAAAAAAAxg/Erpt5FGL4p4/s1600/matrix+multiplication.jpg?w=456)

¿Qué tiene que ver esto con k-means? La matriz no marcada en la parte inferior derecha es la matriz de datos con filas como individuos y columnas como variables (4 personas con 3 medidas). El círculo verde es la puntuación de la tercera persona en la tercera variable. Se calcula como (1,1) \* b (1,3) + a (3,2) \* b (2,3). En un k-medio A es la matriz de pertenencia con K columnas como indicadores de clusters. Cada fila tiene una sola entrada con una que indica su membresía de clúster y ceros K-1 para los otros clústeres. Si la Persona # 3 ha sido clasificada como el grupo # 1, entonces su tercera variable se reduce a b (1,3) ya que a (1,1) = 1 y a (3,2) = 0. De hecho, la fila entera para este individuo es simplemente una copia de la primera fila para B, que contiene el centroide para el primer racimo.

Con dos grupos, k-means te da uno de los dos perfiles de puntuación. Usted me da el número de clúster diciéndome qué columna de A es una, y voy a leer las puntuaciones que debe obtener del perfil de cluster en la fila correcta de B. El mismo proceso se repite cuando hay más grupos, Y la matriz de datos reproducidos contiene sólo K patrones únicos. La pertenencia a clústeres significa perder su identidad y adoptar el centroide de clúster como su perfil de datos.

Con una membresía dura de todo o nada del grupo, todos en el mismo grupo deberían tener el mismo patrón de puntuaciones, excepto para la variación aleatoria. Esto no sería el caso con la membresía de clúster suave, es decir, las filas de la matriz de membresía A aún sumarían a uno, pero las entradas serían probabilidades de pertenecer a un clúster que varían de cero a uno. De forma similar, B no necesita ser el centroide de agrupamiento. Las filas de B podrían representar un arquetipo, un patrón extremo o inusual de partituras. [El análisis arquetípico](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://joelcadwell.blogspot.com/2012/07/archetypal-analysis.html&usg=ALkJrhj5x7iQuD7HWEY-Lp1LtgwMiZ4tsA) adapta tal enfoque y también lo hace la [factorización](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://joelcadwell.blogspot.com/2014/07/taking-inventory-analyzing-data-when.html&usg=ALkJrhiiuik8ZeputRvQ8aAG7MmWQzixrg) no [negativa de la matriz](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://joelcadwell.blogspot.com/2014/07/taking-inventory-analyzing-data-when.html&usg=ALkJrhiiuik8ZeputRvQ8aAG7MmWQzixrg) (NMF), aunque las filas de B tienen diferentes restricciones. Ambas técnicas se resumen en la Sección 14.6 del libro en línea [Elements of Statistical Learning](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf&usg=ALkJrhi7jahevR-qck4dsCtIKbcNY_rElg" \t "_blank) .

Teniendo en cuenta el título de este post, es posible que desee saber lo que tiene que ver con la selección de variables. La no negativa en NMF restringe todos los valores de las tres matrices a ser cero o un número positivo: la matriz de datos contiene recuentos o cantidades, la pertenencia a cluster se transforma a menudo para parecer una probabilidad que varía entre 0 y 1, y los clústeres se definen Siempre añadiendo variables o excluyéndolas enteramente con un coeficiente cero. Como resultado, encontramos en aplicaciones reales que muchos de los coeficientes en B son cero, indicando que la variable asociada ha sido excluida. Lo mismo ocurre con la matriz A que sugiere una co-agrupación simultánea de las columnas y filas que forman sub-matrices de alta densidad reordenando estas filas y columnas para que aparezcan como bloques homogéneos. Usted puede ver esto en el [heatmaps](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://nmf.r-forge.r-project.org/demo-heatmaps.html&usg=ALkJrhhCZ-sZsvSgbqh3tVvle2Ooqar-wA" \t "_blank) del paquete de NMF R con las [porciones](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://nmf.r-forge.r-project.org/demo-heatmaps.html&usg=ALkJrhhCZ-sZsvSgbqh3tVvle2Ooqar-wA) de regiones de baja densidad y [solamente](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://nmf.r-forge.r-project.org/demo-heatmaps.html&usg=ALkJrhhCZ-sZsvSgbqh3tVvle2Ooqar-wA) algunos bloques de alta densidad.

**Ampliar la definición de un cluster**

Biclustering desvía nuestra atención del diagrama de dispersión y lo concentra directamente en la matriz de datos, específicamente, cómo se puede descomponer en componentes y luego recomponer un bloque de construcción a la vez. Los racimos ya no son patrones descubiertos en la nube de puntos trazados en algún espacio de alta dimensión y observados por partes 2 o 3 dimensiones a la vez. Tampoco son clases en grupos o particiones en mezclas de diferentes distribuciones. Los clústeres se han convertido en componentes que se combinan de forma aditiva para reproducir la matriz de datos original.

En la figura anterior, las filas de B definen los clústeres en términos de las variables observadas en las columnas de B, y A muestra cuánto aporta cada grupo a cada fila de la matriz de datos. Por ejemplo, un consumidor nos dice qué información buscan al seleccionar un hotel. Biclustering ve esa fila de la matriz de datos como una combinación lineal de un pequeño conjunto de estrategias de búsqueda de información. Los consumidores pueden tener una membresía parcial en más de un cluster, y lo que entendemos por pertenecer a un cluster es adoptar una estrategia de búsqueda de información en particular. Un purista se basa en una sola estrategia para que su fila en la matriz de membresía de clúster tendrá un valor cercano a uno. Otros consumidores adoptarán una mezcla de estrategias con membresía extendida a través de dos o más entidades de fila.

Un [post anterior proporciona](https://translate.googleusercontent.com/translate_c?depth=1&hl=es&prev=search&rurl=translate.google.com.mx&sl=en&sp=nmt4&u=http://joelcadwell.blogspot.com/2014/07/uncovering-preferences-shaping-consumer.html&usg=ALkJrhj9RfmM1cjPxPdCFcnaqAhUCZZ9jA) más detalles, y habrá más por venir.

**130**

**COMPARTE**