CENSUS
Data Analysis & Data Mining

SEGMENTACIÓN : UN PROBLEMA DE MINERÍA DE DATOS Y DOS ALGORITMOS

María del Rosario Bruera

Publicado en la revista “En conexión” (IBM ,Data Management Division Software) Año 2, Número 5, Oct-Dic 2001

La segmentación de los datos es la tarea inicial de todo proyecto de Data Mining ya que sirve de soporte a todas las actividades que involucra, por ejemplo, el CRM.

Consiste en asignar a los individuos del universo de análisis a un único grupo (cluster o segmento) de modo **exhaustivo** – todos los individuos tienen un grupo de pertenencia – y mutuamente **exclusivo** – cada individuo pertenece a un único segmento – tratando de maximizar la **homogeneidad** de los individuos **dentro** de cada segmento y las **diferencias entre** segmentos.

Desde un punto de vista metodológico esta segmentación puede realizarse a partir de criterios pre-establecidos - segmentos por deciles de facturación - o a partir del análisis conjunto de todas las variables que definen el comportamiento de la unidad de análisis (el cliente, por ejemplo). Este último enfoque se denomina "*data driven segmentation*", es decir, los datos **mandan** sin recibir supuestos ni restricciones a priori.

Las herramientas de minería que típicamente se utilizan para producir estos agrupamientos "*data driven*" son los **árboles de decisión** y los **algoritmos de clustering**. La elección de la herramienta dependerá de los objetivos que se persigan con la segmentación.

Los árboles se aplican en el caso de contar con una variable **objetivo**, por ejemplo, la tasa de respuesta a un mailing, cuyo comportamiento se quiere explicar. Esta variable es la que orienta el aprendizaje del modelo y por eso se define a estas técnicas como de **aprendizaje supervisado**. La segmentación intentará separar a los individuos en grupos homogéneos según la variable objetivo.

Por otra parte, los algoritmos de clustering , trabajan de un modo **sin supervisión**, es decir sin la guía de ninguna variable en particular. Todos los datos entran al análisis y el éxito consiste en agrupar a los individuos en segmentos que resulten **significativos** para los objetivos del negocio.

Existe una gran variedad de algoritmos de clustering disponibles en las herramientas de minería, con fundamento estadístico o neuronal, pero no siempre la aplicación directa de estos algoritmos conduce a una segmentación efectiva y aplicable.

Un algoritmo de clustering tiene como propósito agrupar individuos **similares**. El problema es ¿cuándo son similares?. Precisamente el beneficio de seleccionar uno u otro algoritmo dependerá de la capacidad de este algoritmo para definir una buena medida de similitud.

En el siguiente ejemplo se analiza la aplicación de dos algoritmos de clustering : **K-Means** (o K Medias) y **Clustering Demográfico**. El primero está presente en muchas herramientas de minería ya que resulta relativamente sencillo su desarrollo computacional. El Clustering Demográfico es una de las herramientas de clustering disponibles en Intelligent Miner.

El problema consiste en segmentar a 33.668 usuarios de una tarjeta de crédito a partir de variables obtenidas de las transacciones del último año : compra promedio, frecuencia de uso de la tarjeta, cantidad de débitos automáticos y saldo promedio mensual.

De la descripción de estas variables se verifica que no siguen una distribución **normal**. Presentan mucha asimetría y **outliers**, es decir, individuos con comportamientos **atípicos** respecto de la gran masa de clientes "promedio" (gastan mucho más, tienen un saldo muy alto, etc.). Por razones propias del negocio no se quiere eliminar a los outliers sino asimilarlos a la segmentación general.

El objetivo de la segmentación se cumplirá si se identifica grupos que muestren un **Pareto efectivo**: el 20% de los clientes resume el 80% de los beneficios. Para este ejemplo se elige como variable de valor el **saldo mensual**.

La aplicación inicial del algoritmo de K-medias conduce a:

Segmento	Suma Saldo Mensual	Cuentas	% Suma Saldos	% cuentas
1	2.772.601	260	9	1
2	13.358.865	3.979	43	12
3	14.648.991	29.429	48	87
Total	30.780.457	33.668	100	100

Esta segmentación desde un punto de vista práctico es totalmente inútil: concentra el 87% de las cuentas en un solo segmento y aísla los "outliers" : 260 cuentas.

Este problema del algoritmo K-medias - debido a las asimetrías de las variables - puede resolverse realizando un pre-procesamiento de los datos mediante una técnica de **multidimensional scaling** disponible en productos de análisis estadístico. Con este tratamiento previo, la posterior aplicación de K-medias conduce a:

Segmento	Suma Saldo Mensual	Cuentas	% Sum	% cuentas
1	4.401.944	13.576	14	40
2	25.833.505	13.291	84	39
3	545.008	6.801	2	20
Total	30.780.457	33.668	100	100

Ahora el 39% de las cuentas acumula el 84% de la suma de saldos mensuales.

Los mismos datos procesados con Clustering Demográfico de Intelligent Miner - sin intervención del usuario para corregir las asimetrías - se agrupan:

Segmento	Suma Saldo Mensual	Cuentas	% Sum	% cuentas
0	962.186	6.017	3	18
1	7.871.636	18.319	26	54
2	21.946.635	9.332	71	28
Total	30.780.457	33.668	100	100

El Pareto se mejora ya que el 28% de las cuentas acumulan ahora el 71% del saldo.

Para el tratamiento de los outliers no fue necesaria la aplicación de ninguna técnica estadística sino solamente seleccionar una de las opciones previstas en el menú de parámetros de la función de minería. La asimetría de la distribución se resuelve mediante transformaciones que realiza por sí mismo Intelligent Miner y que son transparentes al analista.

De este modo el algoritmo de Clustering Demográfico asiste al usuario de manera efectiva para cumplir con el objetivo del problema aliviando las etapas de preparación de los datos, ya de por sí complejas, y que generalmente ocupan el 70% del tiempo total del proyecto y permitiendo su aplicación con buenos resultados a usuarios no necesariamente expertos en metodología de análisis estadístico.

María del Rosario Bruera
Census – Data Analysis & Data Mining
chb@census.com.ar

Bibliografía:

- Data Mining with Neural Networks. Joseph Bigus. Mc Graw Hill, USA, 1996
- Data Mining Techniques for Marketing, Sales and Customer Support. Michael Berry, Gordon Linoff. Wiley, USA, 1997
- Techniques of Cluster Algorithms in Data Mining. Versión 2.0. Johannes Grabmeier, Andreas Rudolph. IBM Deutschland Informationssysteme GmbH, Germany, 1998
- Data preparation for Data Mining. Dorian Pyle. Morgan Kaufmann Publishers Inc. San Francisco, USA, 1999
- Mastering Data Mining. Michael Berry, Gordon Linoff. Wiley, USA, 2000
- Building Data Mining applications for CRM. A. Berson, S. Smith, K. Thearling. Mc Graw Hill, USA, 2000