DESCUBRIMIENTO DE INFORMACIÓN USANDO SECUENCIAS DE MAPAS PERCEPTUALES

Esther Hochsztain✦✧ Andrómaca Tasistro✦

Departamento de Métodos Cuantitativos✧ Instituto de Computación✦

Facultad de Ciencias Económicas Facultad de Ingeniería

[esther, tasistro]@fing.edu.uy

Universidad de la República

Montevideo - Uruguay

Abstract

En este artículo se utilizan técnicas estadísticas de análisis exploratorio de datos para descubrir información, en particular escalamiento multidimensional y métodos factoriales. El método usado es el análisis de correspondencias, que se utiliza en marketing con el nombre de mapas perceptuales. Éstos constituyen representaciones gráficas que son un buen medio de comunicación, ya que no es necesario ser especialista en estadística para comprender que la proximidad entre dos puntos traduce la semejanza entre los objetos que representan sin que sea necesario comprender la formalización matemática de esta semejanza. Nuestro enfoque se distingue del enfoque tradicional en estadística, en que procura el descubrimiento automático de información.

Palabras clave: bases de datos, data mining, descubrimiento de información, mapas perceptuales.

1. Introducción-

El desarrollo de nuevas técnicas para encontrar la información requerida a partir de enormes cantidades de datos es uno de los principales desafíos para los desarrolladores de software de hoy día. La cantidad de datos en el mundo se duplica cada año, y como una consecuencia sorprendente, la cantidad de información disminuye rápidamente. El hecho de que la cantidad de datos esté creciendo es la razón de la dificultad creciente de encontrar los hechos significativos que buscamos [HST94]. No se puede ver el bosque por los árboles.

Hoy en día la información es considerada también como un factor productivo. En el futuro, la habilidad para leer e interpretar por sí sola no será suficiente para sobrevivir como profesional, científico o como una organización comercial. La producción automática y la reproducción de datos nos fuerzan a adaptar nuestras estrategias y desarrollar métodos automáticos para filtrar, seleccionar e interpretar datos. Las organizaciones que sobresalgan en esta tarea tendrán mejores posibilidades de sobrevivir, y a causa de esto, la información en sí misma se ha convertido en un factor de producción importante.

La explosión de datos en la sociedad moderna y el corolario de la producción automática de datos ha creado la necesidad de proceso automático de datos. La mayor

parte de las organizaciones tienen grandes bases de datos que contienen una riqueza de información potencialmente accesible. Sin embargo, usualmente les es muy difícil acceder a esta información. El desenfrenado crecimiento de datos inevitablemente conduce a una situación en la que es cada vez más difícil acceder a la información deseada: será siempre como buscar una aguja en un pajar, y el tamaño del pajar crece contínuamente.

Con estos antecedentes se entiende el gran interés que ha despertado el área de data mining o Knowledge Discovery in Databases (KDD). El nombre data mining es una metáfora. Como es bien sabido, en minería, enormes cantidades de desechos deben ser removidos para encontrar diamantes u oro. La analogía de que, con un computador, se puede automáticamente encontrar el "diamante de información" entre toneladas de datos de desecho en su base de datos, es por supuesto muy atractiva. No existe una definición única de data mining o KDD. En este artículo se usará la dada por [ADZA96] " la extracción no trivial de conocimiento potencialmente útil , implícito y previamente desconocido a partir de los datos". KDD no es una nueva técnica sino un ámbito multidisciplinario de investigación al que contribuyen áreas como machine learning, estadística, tecnología de bases de datos, sistemas expertos y visualización de datos.

En este artículo se tratará de descubrir información utilizando técnicas estadísticas de análisis exploratorio de datos, en particular escalamiento multidimensional y métodos factoriales. El método usado es el análisis de correspondencias, que se utiliza en marketing con el nombre de mapas perceptuales. Éstos constituyen representaciones gráficas que son un buen medio de comunicación, ya que no es necesario ser especialista en estadística para comprender que la proximidad entre dos puntos traduce la semejanza entre los objetos que representan sin que sea necesario comprender la formalización matemática de esta semejanza. Nuestro enfoque se distingue del enfoque tradicional en estadística, en que procura el descubrimiento automático de información.

Este artículo está organizado de la siguiente forma: la sección 2 describe la motivación de este trabajo presentando un ejemplo de aplicación, la sección 3 define los conceptos básicos utilizados en las áreas de data mining y estadística, la sección 4 presenta un algoritmo para descubrimiento de información usando mapas perceptuales. En la sección 5 se presentan las conclusiones del trabajo.

2. *Motivación a través de un ejemplo-*

La marca E de galletitas después de realizar un estudio de mercado decidió hacer una campaña orientada a cambiar su imagen. En ese estudio se encontró que la marca E estaba fundamentalmente asociada a la mejor presentación. Se desea que el público la asocie con las más ricas y la mayor variedad de sabores, ya que el mismo estudio indicó que estas dos últimas características son las determinantes en la compra.

A partir del momento del lanzamiento de la campaña se desea ir evaluando periódicamente sus resultados en función de los datos de encuestas realizadas mientras dura la campaña publicitaria. En la figura 1 se presenta la pregunta básica efectuada en estas encuestas:

"Yo le voy a leer algunas frases y le pido que para cada frase leída me indique qué marca (A, B, C, D, E, F o G) reúne mejor o se asocia más a esa característica."

	A	B	C	D	E	F	G
La mayor variedad de sabores							
Hechas con productos naturales							
Las más ricas							
Los mejores precios							
Hechas con productos de primera calidad							
La mejor presentación							
Se encuentran en todas partes							
Los mejores envases							

Figura 1: Parte del formulario de la encuesta

Para visualizar y analizar los datos se utilizan los mapas perceptuales que se definen en la siguiente sección. En la figura 2 se muestra el mapa perceptual obtenido con los datos de 800 encuestas antes de comenzar la campaña publicitaria.

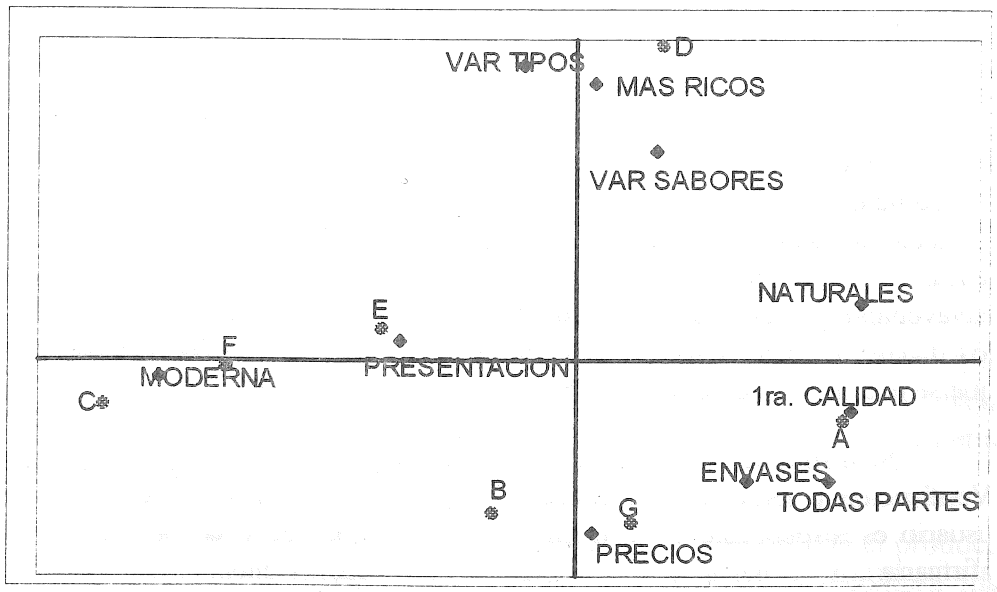

Figura 2: mapa perceptual antes de comenzar la campaña

La proximidad entre los puntos indica que están asociados en la percepción del público. Por ejemplo, se ve que las marcas B y G son percibidas como las que tienen mejores precios, en cambio la marca E es percibida como la que tiene la mejor presentación.

3. *Descubrimiento de información usando mapas perceptuales-*

En esta sección se presentan los conceptos básicos de descubrimiento de información y la técnica de mapas peceptuales utilizada para la visualización de los datos.

3.1. Descubrimiento de información

En las últimas dos décadas se ha registrado un gran aumento en la cantidad de información o datos almacenados electrónicamente. Se estima que la cantidad de información en el mundo se duplica cada 20 meses y el tamaño y cantidad de bases de datos crece más rápido aún. [ADZA96].

Actualmente el almacenamiento de datos es cada vez más fácil y más barato. Además, hay más capacidad de procesamiento lo que permite nuevos métodos de análisis. Al análisis de datos tradicional en estadística se agregan nuevos métodos de machine learning para representación de conocimiento basados en programación lógica.

Los Database Management System permiten el acceso a los datos almacenados, pero esto es sólo una pequeña parte de lo que se puede obtener de los datos. Los OLTPs (On-Line Transaction Processing systems) almacenan datos en forma rápida, segura y eficiente. Las técnicas de Data Mining o Knowledge Discovery in Databases (KDD) en cambio, tienen por objetivo analizar los datos buscando su significado.

El proceso de Data Mining comienza con los datos brutos y termina con el conocimiento adquirido como resultado de las siguientes etapas: Selección (segmentación de los datos de acuerdo a algún criterio), Preprocesamiento (limpieza de datos donde se elimina información innecesaria y se reconfiguran los datos para lograr un formato consistente), Transformación (los datos se hacen útiles y navegables), Data Mining propiamente dicho (extracción de información relevante de los datos), Interpretación (la información obtenida se interpreta como conocimiento que se utilizará como soporte para la toma de decisiones).

Se distinguen dos modelos de Modelos de Data Mining [AMST96]: Modelo de Verificación (Recibe una hipótesis del usuario y testea su validez en los datos. El usuario es responsable de formular la hipótesis y hacer la consulta de los datos para afirmarla o negarla) y Modelo de Descubrimiento (El sistema descubre automáticamente información importante escondida en los datos. Los datos se examinan en búsqueda de patrones, tendencias y generalizaciones sin la intervención o guía del usuario. Se procura encontrar un gran número de hechos relativos a los datos en el menor tiempo posible). Nuestro trabajo se enmarca dentro del segundo modelo descripto.

El data mining no es una técnica nueva sino un ámbito de investigación multidisciplinario al que contribuyen las áreas de machine learning, estadística, bases de datos, sistemas expertos y visualización de datos. [ADZA96]. Nuestra propuesta trabaja dentro de la estadística con análisis exploratorio de datos (Exploratory Data Analysis) que permite resumir y visualizar grandes cantidades de datos complejos. En particular, se utiliza la técnica descriptiva de mapas perceptuales para visualizar la información.

3.2. Mapas perceptuales

Desde hace una veintena de años, los métodos para el análisis de datos han probado ampliamente su eficacia en el estudio de grandes masas complejas de información [ESPA92]. Se trata de métodos llamados multidimensionales por oposición a los métodos de estadística descriptiva que no tratan más que de una o dos variables a la vez. Por tanto, permiten la confrontación entre numerosas informaciones, lo que es infinitamente más rico que su examen por separado. Las representaciones simplificadas de grandes tablas de datos que estos métodos permiten obtener se han manifestado como un instrumento de síntesis notable. Extraen las tendencias más sobresalientes de datos demasiado numerosos para ser aprehendidos directamente, los jerarquizan y eliminan los efectos marginales o puntuales que perturban la percepción global de los hechos.

Estos métodos en principio fueron utilizados para en dominios científicos como la ecología, la lingüística, la economía, etc. Pero actualmente se aplican en marketing, seguros, banca, etc. El primer objetivo es el de conservar las informaciones y poder consultarlas fácilmente. No obstante, es evidente que para explotar los datos almacenados, cuya recopilación a sido a menudo costosa, es necesario disponer de útiles estadísticos adaptados a las mismas.

El análisis factorial ocupa un lugar primordial entre los métodos de análisis de datos. Esto es debido en parte a las representaciones geométricas de los datos que transforman en distancias euclideanas las proximidades estadísticas entre elementos. Permiten usar las facultades de percepción cotidianamente utilizadas. Sobre los gráficos del análisis factorial se ven, literalmente, agrupaciones, oposiciones, tendencias, imposibles de discernir directamente sobre una gran tabla de números, incluso luego de un examen prolongado.

Hay diversos tipos de análisis factorial. Cuando las variables que intervienen son categóricas se denomina análisis factorial de correspondencias. Los mapas perceptuales son la aplicación del análisis factorial de correspondencia al marketing para efectuar estudios de imagen[GCS89].

Una imagen de marca consta de todas las cosas que se asocian con el producto o que se perciben acerca de él. Es esta imagen lo que la gente piensa que compra, más que la realidad [EVAN59]. La imagen de una marca se mide en función de muchas características. En nuestro caso las marcas de galletitas se comparan a través de 8 características, por lo que para representar esta información se necesitarían 8 dimensiones.

La escala multidimensional (Multidimensional Scaling, MDS) trata el problema general de dar posición a objetos en un espacio perceptivo. Gran parte de la administración de la mercadotecnia está relacionada con la asignación de posiciones. ¿Con quién competimos? ¿Cómo se nos compara con nuestros competidores? ¿Sobre qué dimensiones? ¿Qué posición estratégica debe seguirse? Éstas y otras cuestiones se tratan en MDS [AADA91].

La MDS involucra básicamente dos problemas. Primero deben identificarse las dimensiones sobre las cuales los clientes perciben o evalúan objetos (organizaciones, productos o marcas). Por ej., los consumidores deben evaluar galletitas en términos de su calidad, costo, etc. Sería conveniente trabajar sólo con dos dimensiones, puesto que las marcas entonces podrían representarse gráficamente. Sin embargo, no siempre es posible trabajar con dos dimensiones, puesto que algunas veces se necesitan dimensiones adicionales para representar las percepciones y las evaluaciones de los clientes. Segundo, las marcas necesitan ser posicionadas con respecto a estas dimensiones y a esto se denomina mapa perceptual.

4. *Algoritmo para descubrimiento de información-*

En esta sección se muestra la metodología propuesta mostrando su aplicación a un caso de estudio. Inicialmente se tiene un mapa perceptual originado en una visión de los datos en el momento del tiempo t0, antes de comenzar la campaña, tal como se vio en la Figura 2. Se hace una interpretación de ese mapa perceptual. Este tipo de interpretación no es trivial, ya que es necesario dar un significado a los ejes y analizar que puntos proyectados cerca realmente están próximos.

Periódicamente se van realizando encuestas y éstas originan mapas perceptuales tal como se ve en la figura 3. Al momento ti le corresponde el mapa mpi.

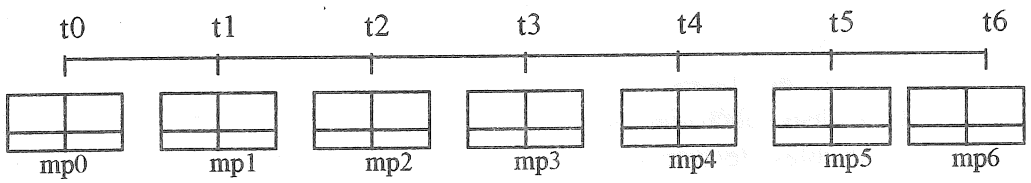

Figura 3: sucesión de mapas perceptuales.

Se procurará ir detectando la ocurrencia de cambios a lo largo del tiempo. Esto ayuda a definir si es necesario seguir con la campaña, en función de si ésta está logrando sus objetivos.

Para detectar si se han producido cambios se define una forma de comparación entre mapas. Como nos interesa la posición de la marca E respecto de las características ser las más ricas y tener la mayor variedad de sabores, mediremos en los distintos mapas la distancia euclideana entre los puntos que representan la marca E y estas dos características. Cuanto más próximos estén mayor será la asociación entre ellos.

Al realizar un nuevo mapa perceptual, en general, éste debe interpretarse nuevamente, ésto implica un costo en cuanto al tiempo y la necesidad de contar con asesoramiento en estadística. En nuestro caso, esto no será necesario ya que, los mapas son "parecidos" al del momento t0, que ya está interpretado. Entonces, se puede observar la magnitud y la orientación del cambio en el nuevo mapa sin la intervención de un experto, utilizando solamente la noción de distancia que se ha definido.

En el caso de nuestro estudio, el usuario (gerente de marketing) pudo observar fácilmente que las distancias definidas se iban acortando y por lo tanto la campaña estaba produciendo los efectos deseados.

5. Conclusiones-

Hemos presentado una metodología que integra la visualización multidimensional, la tecnología de exploración de datos y el descubrimiento de información. Nuestra propuesta es una forma de descubrir información en marketing utilizando secuencias de mapas perceptuales. Permite estudiar no sólo la imagen de las marcas en un momento dado, sino también analizar su evolución a lo largo del tiempo. Es un instrumento sumamente útil para la adopción de decisiones.

Pese a que nuestra propuesta no es totalmente automática, ya que es necesario que un experto en estadística realice la primera interpretación, se ha señalado [ELPR96] que es difícil concebir que todo el proceso pueda ser alguna vez automatizado. La creciente automatización no ha evitado a los investigadores la necesidad de pensar en términos estadísticos, entre ellos la búsqueda de interpretabilidad. Sin embargo, las técnicas propuestas permiten al analista pensar en el problema a un nivel más alto y colaboran en el objetivo de encontrar formas claras, fáciles y fluídas de suministrar información útil.

En futuros trabajos pensamos analizar nuevas formas de definir la similitud entre mapas perceptuales, de modo que la decisión de que dos mapas perceptuales son "parecidos" pueda tomarse en forma automática.

6. Bibliografía

[AADA91] A. Aacker, P. Day, "Investigación de Mercados". Mc. Graw Hill, México, 1991.

[ADZA96] P. Adriaans, Dolf Zantinge "Data Mining" Addison Wesley Longman. England 1996.

[AMST96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo "Fast Discovery of Association Rules" en Advances in Knowledge Discovery and Data Mining U. Fayyad et al. AAAI Press USA 1996.

[ELPR96] J.F. Elder, D. Pregibon "A Statistical Perspective on Knowledge Discovery in Databases" en Advances in Knowledge Discovery and Data Mining U. Fayyad et al. AAAI Press USA 1996.

[ESPA92] B. Escofier, J. Pagès. "Análisis factoriales simples y múltiples". Dunod, París, 1992.

[EVAN59] F.B. Evans. "Factores psicológicos de la predicción sobre la elección de marcas; Ford y Chevrolet". Journal of Bussiness, vol XXXII, núm 4 , octubre 1959.

[GCS89] P.Green, F.Carmone, S.Smith. "Multidimensional Scaling: concepts and applications". Allyn and Bacon, USA, 1989.

[HST94] E. Hochsztain, H. Steffen, A. Tasistro: *A Database Tool for Information Discovery Based on Pattern Recognition*. XIV International Conference of the Chilean Computer Science Society (SCCC Conference), Concepción, Chile, 1994.