K- MEDIAS AXIAL EN EL ANÁLISIS DE CANASTAS DE PRODUCTOS

Área de investigación: Informática Administrativa

Ana Isabel Tenjo Morales
Departamento de Ciencias Básicas
Universidad de la Salle
Colombia
atenjo@unisalle.edu.co, aitenjom@yahoo.com.mx

XVIII Congreso Internacional de Contaduría Administración e Informática

Octubre 2, 3 y 4 de 2013 ◆ Ciudad Universitaria ◆ México, D.F.

ANFECA
Asociación Nacional de Facultades y Escuelas de Contaduría y Administración

K- MEDIAS AXIAL EN EL ANÁLISIS DE CANASTAS DE PRODUCTOS

Resumen

El análisis de canastas de productos se usa en la optimización de procesos como: ubicación de productos en almacenes o en centros de distribución, organización e implementación de campañas de mercadeo, articulación de planes de abastecimiento, etc. Este análisis generalmente se realiza mediante reglas de asociación que dependiendo de los datos, pueden ser binarias, lineales, secuenciales, etc. El algoritmo a priori brinda buena información para el análisis de canastas, pero genera demasiadas reglas, lo que dificulta la identificación de información relevante para la toma de decisiones. Para superar estas deficiencias se han propuesto diversas técnicas entre las que están métodos de clasificación que han mejorado los resultados pero que requieren de información adicional para su uso. Este trabajo presenta una metodología para el análisis de canastas de productos, mediante la aplicación del método llamado K- medias axial- KMA usado en el análisis de datos textuales; el cual permite clasificar las canastas por tipos de productos y proporciona resultados que se pueden utilizar para mejorar el proceso de construcción de reglas de asociación binarias, sin necesidad de información adicional. Desde los resultados se puede concluir que el uso del KMA mejora y facilita el análisis de canastas de productos.

Palabras clave. Análisis de canastas de productos; K-medias axial; reglas de asociación.

K- MEDIAS AXIAL EN EL ANÁLISIS DE CANASTAS DE PRODUCTOS

Introducción

Tener información organizada de lo que adquieren los clientes, se ha convertido en una herramienta clave para el éxito de muchas empresas, además con el avance tecnológico, la capacidad para almacenar bases grandes de datos con este tipo de información ha ido creciendo a ritmo acelerado. Esta información se puede utilizar optimizar procesos en diferentes áreas como: ubicación de productos en almacenes o en centros de distribución, organización e implementación de campañas de mercadeo, articulación de planes de abastecimiento, etc. (Buitrago 2006). Los métodos existentes para procesar y utilizar toda esta información generalmente no son suficientes ver (Hernández et al. 2005); haciendo necesaria la búsqueda continua de procesos eficientes para responder a las necesidades y exigencias actuales (Berry & Linoff 2004).

Las técnicas de "minería de datos" han mejorado notablemente el manejo de bases grandes, utilizadas en diversas áreas como: inversión, industria, astronomía, medicina, análisis textual, mercadotecnia, etc. (Han & Kamber 2001). Dos importantes métodos de "minería" son las técnicas de agrupación y las reglas de asociación. Las técnicas de agrupación se utilizan para encontrar grupos entre un conjunto de individuos, de tal forma que individuos similares (cercanos) se asignan al mismo grupo, mientras que las reglas de asociación (**RA**), permiten encontrar relaciones existentes y patrones de comportamiento de conjuntos disjuntos de datos en términos de elementos que se "adquieren" usualmente juntos (Liu et al. 1998).

En este trabajo se extiende el uso del método de agrupación K- medias axial (en adelante **KMA**), concebido inicialmente para el estudio de bases de datos documentales y textuales, al análisis de canastas de productos, cuyas bases de datos están conformadas por miles o millones de registros que generalmente se analizan mediante diferentes reglas de asociación. El **KMA** (Lelu 1993), parte de un conjunto de documentos denominado "curpus documental" que se representa mediante una matriz de datos de la forma: "Documentos ×Palabras clave", la cual en el contexto del análisis de canastas de productos se puede interpretar como la matriz "Canastas × productos".

Con la aplicación del **KMA** en el análisis de canastas de productos, se busca proporcionar una herramienta que permita organizar la información por tipologías de productos y a la vez por tipologías de canastas, y que además brinde información relevante que facilite el proceso de búsqueda de reglas de asociación. Para tal propósito se construyen diferentes clases caracterizadas por productos particulares, las cuales se representan cada una, por un vector llamado "eje central de clase". Las coordenadas de los ejes de clases que se destacan como típicas de la clase, permiten identificar simultáneamente los productos que

caracterizan al grupo y los conjuntos candidatos de ítems frecuentes para el proceso de identificación de reglas de asociación.

1. Análisis de canastas

El "análisis de canastas de productos" básicamente se entiende como el estudio de lo que adquieren las personas en una compra o transacción (Narros 2007). Con el avance tecnológico la forma de realizar transacciones se ha ido ampliando y con éste muchos conceptos asociados. En la actualidad el análisis de canastas de productos se puede abordar desde diferentes bases de datos, como los registros de datos de clientes, llamadas de pedidos, registros Web, almacenamiento de mercancía, diferentes puntos de venta, archivos de facturación, etc., éstos conforman una gran fuente de información útil para las empresas. La idea es extraer patrones de conducta de compra sobre tablas grandes de datos de múltiples categorías de productos y sus interrelaciones. Esta información se utiliza para mejorar las estrategias orientadas a lograr el éxito de las empresas, facilitando la planificación y ejecución de acciones de mercadeo y abastecimiento, ajustadas a las necesidades específicas de cada población de interés (Berry & Linoff 2004).

1.1 Estado del análisis de canastas de productos

En las últimas décadas el análisis de canastas de productos básicamente se ha abordado desde dos perspectivas como se presenta en Narros (Narros 2007): mediante metodologías de análisis estadísticos clásicos como tablas de contingencia, modelos de regresión, modelos econométricos, modelos logit, modelos log lineales (para modelización de ventas, identificación de segmentos de compra homogeneidad, estimación de dependencia, identificación de correlaciones, etc.) y en la actualidad gracias al avance tecnológico, se hace cada vez más frecuente el uso de sistemas conocidos como minería de datos, cuyo propósito es extraer información que está implícita en los datos. De éstas últimas, las de uso más habitual en el estudio de canastas, son las redes neuronales y las reglas de asociación. En el análisis de canastas las redes neuronales permiten identificar y representar tipologías de clientes basándose en patrones de conducta comunes en su comportamiento en la canasta de compra (Narros 2007). Las reglas de asociación (que se estudian más en detalle en la siguiente sección) permiten identificar asociaciones de productos en la misma canasta.

1.2. Reglas de asociación

Aparecieron con el objetivo de estudiar los hábitos de compra de los clientes (Han & Kamber, 2001). Las múltiples aplicaciones en diversas líneas de investigación, han motivado gran número de estudios, que buscan la manera más rápida y eficiente para identificarlas. Existen diversas metodologías para la búsqueda y análisis de RA. Narros (Narros, 2007) presenta la siguiente clasificación de las RA:

- Reglas de asociación apriori.
- Modelos GRI (Generalizad Rule Induction).
- Modelos de reglas secuenciales generadas (Generated Sequence Rules).

1.2.1 Análisis de canastas mediante reglas de asociación.

Para reglas de asociación apriori las diferentes canastas se escriben como una matriz de datos binarios **X** (cuyas filas representan las diferentes canastas): con entradas ij-ésimas iguales a 1 si la canasta i incorpora el producto j y 0 si no lo hace (Hernández, 2005).Dada esta matriz **X**, se representa:

- El conjunto de todos los productos (ítems) en la tabla como (Liu et al. 1998):

$$\mathbf{I} = \{I_1, I_2, \cdots, I_j, \cdots I_p\}$$

- El conjunto de todas las filas de **X** (canastas en la base de datos) como:

$$\mathbf{M} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_i}, \cdots, \mathbf{x_n}\}$$

Definición 1. Dados los conjuntos **I** y **M**, sea $A \subseteq \mathbf{I}$ un conjunto de ítems. Se dice que la canasta $\mathbf{x_i}$ ($\mathbf{x_i} \in \mathbf{M}$) incorpora al conjunto de ítems A sí para todo $I_j \in A$, $x_{ij} = 1$. El conjunto de filas en **M** que incorporan a A se denota por $\mathbf{M}_A$, así:

$$\mathbf{M}_A = \{\mathbf{x_i} \in \mathbf{M} / \text{para todo } I_j \in \text{A}, x_{ij} = 1\} \quad (1.1)$$

Ejemplo 1. Considere la tabla 1.1 compuesta por 10 canastas (o transacciones) y 9 ítems relacionados.

Tabla 1.1. Ejemplo de 10 canastas con 9 ítems

	Productos (ítems)					
Canastas						
$Canasta_1$	I_1,	I_2,	I_7,			
$Canasta_2$	I_3,	I_4,	I_8			
$Canasta_3$	I_5,	I_6				
$Canasta_4$	I_1,	I_2,	I_3,	I_4,	I_7,	I_8
$Canasta_5$	I_1,	I_2,	I_7,	I_8		
$Canasta_6$	I_3,	I_4,	I_8,	I_9		
$Canasta_7$	I_1,	I_5,	I_6,	I_9		
$Canasta_8$	I_1,	I_5,	I_6,	I_7,	I_9	
$Canasta_9$	I_1,	I_2,	I_4,	I_7		
$Canasta_{10}$	I_3,	I_4,	I_8,	I_9		

Para la tabla 1.1, con $A = \{I_1, I_2\}$ se tiene $\mathbf{M}_A = \{\mathbf{x_1}, \mathbf{x_4}, \mathbf{x_5}, \mathbf{x_9}\}$

Definición 2. El soporte de un conjunto de ítems A en **M**, se denota por $Sop(A)$ y se define como:

$$Sop(A) = \frac{\#\mathbf{M_A}}{n} = \tilde{P}(A) \quad (1.2)$$

Donde n es el número de filas de **X**.

Notación. Se antepone # a un conjunto para indicar el número de elementos en él $\#\boldsymbol{M}_A$ indica el número de elementos en $\mathbf{M}_A$. $\tilde{P}(A)$ indica una estimación de la probabilidad de que las canastas en **M** incorporen al conjunto de ítems A.

Para el conjunto A en el ejemplo 1, $Sop(A) = \frac{4}{10}$

Definición 3. Dado un umbral mínimo de soporte $\delta \in [0, 1]$, el conjunto de ítems A se dice soportado, si $Sop(A) \geq \delta$.

Observación. El umbral de soporte δ es una cantidad dada por el usuario y depende de la aplicación.

Definición 4. Dados **I**, **M** y δ para una tabla de datos. La colección de conjuntos de ítems A que sobrepasan el umbral de soporte dado δ se denota y se define como

$$A(\delta) = \{A \subseteq \mathrm{I} / Sop(A) \geq \delta\} \quad (1.3)$$

La colección de conjuntos de ítems $A(\delta)$de tamaño *k* que satisfacen el umbral de soporte δ se escribe como:

$$A_k(\delta) = \{A(\delta) / \ \#A = k\}. \quad (1.4)$$

Las **RA** describen cómo varias combinaciones de productos van apareciendo juntos en las mismas canastas. Establecen relaciones de la forma:

sí la canasta x_i *incorpora* a A entonces también *incorpora* a B (lo que se representa como $A \rightsquigarrow B$) con A y B subconjuntos de ítems disjuntos de I.

1.2.2. Medidas de soporte y confianza para una regla

El conjunto A recibe el nombre de predecesor de la regla o antecedente y B es el sucesor o consecuente (Hernández et al. 2005).

- **Soporte**

El soporte de una regla se define como el porcentaje de instancias que la regla predice correctamente.

$$Sop(A \rightsquigarrow B) = Sop(A \cup B) = \frac{\#\mathbf{M}_{(A \cup B)}}{n} = \tilde{P}(A \cup B) \quad (1.5)$$

$Sop(A \rightsquigarrow B)$ en este contexto se toma como una medida de la posibilidad de que las canastas que conforman la matriz **X** , i*ncorporen* el conjunto de ítems $A \cup B$. Aquí $\tilde{P}(A \cup B)$ es una estimación de la probabilidad de que una canasta incorpore el conjunto de ítems $A \cup B$.

- **Confianza**

La confianza mide el porcentaje de veces que la regla se cumple cuando se puede aplicar.

$$Conf(A \rightsquigarrow B) = \frac{Sop(A \cup B)}{Sop(A)} = \tilde{P}(B \mid A) \quad (1.6)$$

$\tilde{P}(B \mid A)$ indica una estimación de la probabilidad de que dado que una canasta *incorporó* el conjunto de ítems A también incorpore el conjunto de ítems B.

Dados (δ) y (β) umbrales para las medidas de soporte y confianza respectivamente, la regla $A \rightsquigarrow B$ es válida en **M**, sí y sólo sí $Sop\ (A \rightsquigarrow B) \geq \delta$ y $Conf(A \rightsquigarrow B) \geq \beta$.

1.2.3 Algoritmo básico de aprendizaje para RA.

- **Extracción de conjuntos de ítem soportados**.

i. Conformar los conjuntos C_1. Selecciona los soportados $C_1(\delta)$

ii. Conforma los $\binom{\#C_1(\delta)}{2}$ de dos candidatos C_2 y entre estos selecciona los soportados $C_2(\delta)$.

iii. Así sucesivamente se procede de manera incremental. $C_r(\delta)$ resulta de unir dos conjuntos que tienen en común $r-1$ ítems.

- **Generación de reglas a partir de conjuntos establecidos**

i. Para todo $A^* = C(\delta)$ genera todos los posibles subconjuntos (A)

ii. Para cada $A \subset A^*$, genera una regla de la forma:

$A \rightsquigarrow (A^* - A)$ que satisface el criterio mínimo de confianza si:

$$\frac{Sop(A^*)}{sop(A)} \geq nivel\ de\ confianza\ \beta\ .$$

Ejemplo 2. Continuando con el ejemplo de la tabla 1.1. Sea $A^* = \{I_1, I_2, I_7\}$ el conjunto de ítems frecuentes. Los subconjuntos propios de A^* son $\{I_1\}, \{I_2\}, \{I_7\}, \{I_1,\ I_2\}$, $\{I_1,\ I_7\}$ y $\{I_2,\ I_7\}$,. Para todos los posibles subconjuntos de A^* se generan todas las posibles reglas de asociación y para cada una de éstas se calcula su respectiva confianza. Las reglas de asociación resultantes con su respectiva confianza se muestran en la tabla 1.2.

Tabla 1.2. Reglas de asociación entre los subconjuntos de A*

Antecedente		consecuente	confianza
$\{I_1\}$	$\rightsquigarrow$	$\{I_2\}$	4/6=0.67
$\{I_2\}$	$\rightsquigarrow$	$\{I_1\}$	4/4=1
$\{I_2\}$	$\rightsquigarrow$	$\{I_7\}$	4/4=1
$\{I_1\}$	$\rightsquigarrow$	$\{I_7\}$	5/6=0.83
$\{I_7\}$	$\rightsquigarrow$	$\{I_1\}$	5/5=1
$\{I_7\}$	$\rightsquigarrow$	$\{I_2\}$	4/5=0.80
$\{I_1, I_2\}$	$\rightsquigarrow$	$\{I_7\}$	4/4=1
$\{I_1, I_7\}$	$\rightsquigarrow$	$\{I_2\}$	4/5=0.80
$\{I_2, I_7\}$	$\rightsquigarrow$	$\{I_1\}$	4/4=1
$\{I_1\}$	$\rightsquigarrow$	$\{I_2, I_7\}$	4/6=0.67
$\{I_2\}$	$\rightsquigarrow$	$\{I_1, I_7\}$	4/4=1
$\{I_7\}$	$\rightsquigarrow$	$\{I_1, I_2\}$	4/5=0.80

Si el umbral de confianza mínimo es por ejemplo de 90%, sólo 6 de estas 12 reglas alcanzan el umbral de confianza. Para todos los conjuntos frecuentes de 2 o más ítems, se generan todos los subconjuntos y sus respectivas reglas de asociación. Al final las reglas que satisfacen el umbral de soporte de 30% y el de confianza de 90%, son las que se muestran en la tabla 1.3:

Tabla 1.3. Reglas de asociación que satisfacen los umbrales δ = 0,3 y β = 0,9.

No. Regla	antecedente		consecuente	soporte	confianza
1	$\{I_5\}$	$\rightsquigarrow$	$\{I_6\}$	0.3	1
2	$\{I_6\}$	$\rightsquigarrow$	$\{I_5\}$	0.3	1
3	$\{I_2\}$	$\rightsquigarrow$	$\{I_7\}$	0.4	1
4	$\{I_2\}$	$\rightsquigarrow$	$\{I_1\}$	0.4	1
5	$\{I_3\}$	$\rightsquigarrow$	$\{I_4\}$	0.4	1
6	$\{I_3\}$	$\rightsquigarrow$	$\{I_8\}$	0.4	1
7	$\{I_7\}$	$\rightsquigarrow$	$\{I_1\}$	0.5	1
8	$\{I_2, I_7\}$	$\rightsquigarrow$	$\{I_1\}$	0.4	1
9	$\{I_1, I_2\}$	$\rightsquigarrow$	$\{I_7\}$	0.4	1
10	$\{I_3, I_4\}$	$\rightsquigarrow$	$\{I_8\}$	0.4	1
11	$\{I_3, I_8\}$	$\rightsquigarrow$	$\{I_4\}$	0.4	1
12	$\{I_4, I_8\}$	$\rightsquigarrow$	$\{I_3\}$	0.4	1
13	$\{I_2\}$	$\rightsquigarrow$	$\{I_1, I_7\}$	0.4	1
14	$\{I_3\}$	$\rightsquigarrow$	$\{I_4, I_8\}$	0.4	1

2. K- medias axial (KMA)

El **KMA** surgió para dar solución a algunos problemas prácticos que resultan por la utilización de bases extensas de datos documentales, entre los cuales se tienen:

- Cómo resumir el contenido de una base documental y presentarlo a los usuarios.
- Dado un término ó un documento, cómo extraerlo a él y al contexto (es decir la lista.de documentos y palabras clave cercanas).
- ¿Existen contextos diferentes en los cuales puedan aparecer?

La idea de solución nació de una integración de métodos de análisis factorial, análisis factorial esférico y modelos de redes neuronales (Lelu 1993). En este trabajo se trata la versión adaptante del K-medias axial (KMA), que es una variante del K-medias clásico (MacQueen 1967). El KMA se sitúa en el marco de métodos para descripciones cualitativas de documentos. En él se considera la colección de referencias bibliográficas como una nube de puntos en un espacio multidimensional donde cada dimensión corresponde a una palabra clave (Lelu 1993).

Para un conjunto de documentos determinado, el KMA parte de una matriz de datos **X**, de *n* filas (documentos), *p* columnas (palabras clave), con entradas x_{ij} iguales a 1, sí el documento *i* tiene asociado la palabra *j* y cero si no la tiene. Las filas de la matriz **X**, son vectores del espacio R^p de la forma $\mathbf{x_i}=(x_{i1}, x_{i2}, \dots, x_{ip})'$ con $i = 1, \dots, n$ y las columnas vectores de R^n de la forma $\mathbf{X_j}=(x_{1j}, x_{2j}, \dots, x_{nj})'$, con $j = 1, \dots, p$. Los vectores $\mathbf{x_i}$ y $\mathbf{X_j}$ son tales que $\|\mathbf{x_i}\|^2$ es el número de palabras asociadas al *i* ésimo documento y $\|\mathbf{X_j}\|^2$ es el número de documentos que asumen el j-ésimo atributo.

	Palabras Clave				
Documentos	x_{11}		x_{1j}		x_{1p}
			$\vdots$		
	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
			$\vdots$		
	x_{n1}		x_{nj}		x_{np}

Figura 2.1. Documentos × Palabras clave.

2.1 Algoritmo KMA - versión adaptante -

Dada una matriz **X**, de n filas, p columnas y con entradas x_{ij} , En el KM (K-medias usual) se comienza por definir k centros de gravedad $\mathbf{u}_0(k)$ para cada una de las clases que se desea agrupar, donde $k = 1, 2, ...$ K (pueden ser las primeras K filas x_i ó K seleccionadas al azar). En (Morato L 1999) se presenta la fórmula que propuso Velasco en 1999 para el cálculo de esta cantidad *K*. Estos centros se recalculan cada vez que entra un dato nuevo, con el fin de situar a los centros en el espacio de forma que los datos con características similares pertenezcan al mismo centro. Después de situar correctamente los prototipos, se compara cada dato nuevo con éstos y se asocia a aquél que sea el más próximo, en términos de una distancia que se elige previamente (normalmente se usa la distancia euclidiana) (MacQueen 1967). El **KMA** no parte directamente de la matriz binaria **X** definida al comienzo de esta sección, sino de la matriz **X** con las filas normalizadas. Esta matriz normalizada se representa como **W**. En la figura 2.2 se presenta la versión adaptante del algoritmo **KMA** ajustada a la notación de este documento.

1. Se inicializa con K ejes $\mathbf{u_k^0}$ (igual que en el **KM**), para $k = 1, ..., K$. Aquí se requiere además que $\|\mathbf{u_k^0}\| = 1$
2. Para cada fila $\mathbf{w_i}$ de **W**, con $\|\mathbf{w_i}\| = 1$, $i = 1, ..., n$. Se calculan las coordenadas de las K proyecciones de las $\mathbf{w_i}$ sobre los ejes $\mathbf{u_k}$:

$$\eta_{k(i)} = < \mathbf{w_i}, \mathbf{u_k} > \tag{2.1}$$

3. Se incorpora $\mathbf{w_i}$ al grupo (o grupos) para el cual la componente de esta proyección sea máxima.
4. Se pone al día la posición del eje $\mathbf{u_k}$:

$$\mathbf{u_k^t} = \mathbf{u_k^{t-1}} + \left(\frac{\eta_{k(i)}}{\tau_k^t}\right)\left(\mathbf{w_i} - \eta_{k(i)}\mathbf{u_k^{t-1}}\right) \tag{2.2}$$

   para: $\tau_k^t = \tau_k^{t-1} + \eta_k^2$, con $\tau_k^0 = 0$
5. Se normaliza el eje $\mathbf{u_k^t}$
6. Después de realizar el procedimiento con las n filas: fin. La ecuación 2.2 es la regla de aprendizaje de Oja (Hertz et al. 1995)

Figura 2.2. Algoritmo **KMA** adaptante

Ejemplo 3. Asumiendo la tabla 1.1 como datos que corresponden a documentos y palabras clave, como se ve en la tabla compuesta por 10 documentos y 9 palabras clave.

Tabla 2.1. Tabla de datos con documentos y palabras clave (matriz **X**).

	Descriptores								
Documentos	Pal_1	Pal_2	Pal_3	Pal_4	Pal_5	Pal_6	Pal_7	Pal_8	Pal_9
$\mathbf{x_1}$	1	1	0	0	0	0	1	0	0
$\mathbf{x_2}$	0	0	1	1	0	0	0	1	0
$\mathbf{x_3}$	0	0	0	0	1	1	0	0	0
$\mathbf{x_4}$	1	1	1	1	0	0	1	1	0
$\mathbf{x_5}$	1	1	0	0	0	0	1	1	0
$\mathbf{x_6}$	0	0	1	1	0	0	0	1	1
$\mathbf{x_7}$	1	0	0	0	1	1	0	0	1
$\mathbf{x_8}$	1	0	0	0	1	1	1	0	1
$\mathbf{x_9}$	1	1	0	1	0	0	1	0	0
$\mathbf{x_{10}}$	0	0	1	1	0	0	0	1	1

Las filas de la tabla 2.1 se normalizan y el algoritmo trabaja sobre la nueva matriz normalizada que se denomina **W.** A continuación se describe el algoritmo KMA aplicado a esta matriz **W.**

1. De **W** se seleccionan al azar (ó directamente) *K* filas, estas filas se toman como ejes iniciales. Para el caso con *K* = 3 y tomando las primeras tres filas de la tabla normalizada de datos, se tienen los tres ejes iniciales que aparecen en la tabla 2.2.

2. La primera fila $\mathbf{w_1}$ de **W** se proyecta sobre cada uno de los ejes iniciales (tabla 2.2). La proyección se escribe como $\eta_{k(1)} =< \mathbf{w_1} \ , \mathbf{u_k^t} >$, tabla 2.3.

Tabla 2.2. Ejes iniciales.

	Palabras claves								
Ejes	Pal_1	Pal_2	Pal_3	Pal_4	Pal_5	Pal_6	Pal_7	Pal_8	Pal_9
$\mathbf{u_1^0}$	0.58	0.58	0.00	0.00	0.00	0.00	0.58	0.00	0.00
$\mathbf{u_2^0}$	0.00	0.00	0.58	0.58	0.00	0.00	0.00	0.58	0.00
$\mathbf{u_3^0}$	0.00	0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00

Tabla 2.3. Proyección de la primera fila sobre los ejes iniciales.

Proyecciones
$Proy_{u_1} w_1 = \eta_{1(1)} = 1$
$Proy_{u_2} w_1 = \eta_{2(1)} = 0$
$Proy_{u_3} w_1 = \eta_{3(1)} = 0$

3. Se identifica el eje para el cual la proyección es máxima y se asigna la fila 1 a la clase que corresponde a dicho eje. Para este caso la proyección máxima fue sobre el eje 1, por lo cual se asigna la fila $\mathbf{w_1}$ a la clase con eje central $\mathbf{u_1}$, luego haciendo $\tau_1^0 = 0$, se calcula $\tau_1^1 = 0 + 1 = 1$ y se transforma el eje 1:

$\mathbf{u_1^1}=\mathbf{u_1^0}+\frac{\eta_{1(1)}}{\tau_1^1}\left(\mathbf{w_1}-\eta_{1(1)}\mathbf{u_1^0}\right)=\mathbf{u_1^1}=\mathbf{u_1^0}+(\mathbf{w_1}-\mathbf{u_1^0})$**.** Luego **s**e normaliza el eje $\mathbf{u_1^1}$**.**

4. La fila 2 de W, w_2 se proyecta sobre cada uno de los ejes u_1^1 , u_2^0 y u_3^0. La proyección se escribe como $\eta_{k(2)}=<w_2, u_k^t>$. Tabla 2.4. La proyección máxima fue sobre el eje 2, por lo cual se asigna la fila $w_2$al eje u_2, y con $\tau_2^0=0$, se calcula $\tau_2^1=0+1=1$ y se recalcula el eje 2:

Tabla 2.4. Proyección de la segunda fila sobre los ejes.

Proyecciones
$Proy_{u_1^1} w_2=\eta_{1(2)}=0$
$Proy_{u_2^0} w_2=\eta_{2(2)}=1$
$Proy_{u_3^0} w_2=\eta_{3(2)}=0$

$u_2^1=u_2^0+\frac{\eta_{2(2)}}{\tau_2^1}\left(\mathbf{w_2}-\eta_{2(2)}\mathbf{u_2^0}\right)=\mathbf{u_2^1}=\mathbf{u_2^0}+(\mathbf{w_2}-\mathbf{u_2^0})$**,** se normaliza el eje u_2^1.

5. Se repite este mismo procedimiento hasta proyectar todas las filas de **W** sobre los 3 ejes (que se van transformando a medida que entra una nueva fila a la clase). La tabla 2.5 contiene los ejes finales, después de que han pasado todas las filas:

Tabla 2.5. Ejes finales (matriz **uk**).

	Palabras clave								
Ejes	Pal_1	Pal_2	Pal_3	Pal_4	Pal_5	Pal_6	Pal_7	Pal_8	Pal_9
$\mathbf{u_1^4}$	0.54	0.54	0.09	0.23	0.00	0.00	0.54	0.23	0.00
$\mathbf{u_2^4}$	0.12	0.12	0.55	0.55	0.00	0.00	0.12	0.55	0.18
$\mathbf{u_3^3}$	0.32	0.00	0.00	0.00	0.62	0.62	0.16	0.00	0.32

la tabla 2.6 resume los resultados de las proyecciones en cada paso. La coordenada $ik-ésima$, es la proyección de la fila *i* sobre el eje *k*.

Las clases se definen a partir de las proyecciones sobre los *K* ejes, tomando como criterio, la proyección máxima de cada documento sobre los diferentes ejes. Estas proyecciones máximas se resaltan con negrilla en la tabla 2.6.

Con las proyecciones se identifican los documentos que conforman cada una de las clases. Los resultados se resumen en la tabla 2.7.

Tabla 2.6. Proyecciones (matriz **M**).

	Ejes		
Documentos	$\mathbf{u_1}$	$\mathbf{u_2}$	$\mathbf{u_3}$
$\mathbf{w_1}$	**1,00**	0.00	0.00
$\mathbf{w_2}$	0.00	**1,00**	0.00
$\mathbf{w_3}$	0.00	0.00	**1,00**
$\mathbf{w_4}$	**0,71**	**0,71**	0.00
$\mathbf{w_5}$	**0,91**	0.55	0.00
$\mathbf{w_6}$	0.27	**0,82**	0.00
$\mathbf{w_7}$	0.27	0.15	**0,71**
$\mathbf{w_8}$	0.48	0.19	**0,80**
$\mathbf{w_9}$	**0,87**	0.46	0.24
$\mathbf{w_{10}}$	0.27	**0,92**	0.16

Note que un documento puede pertenecer a más de una clase (clasificación difusa), como es el caso del documento 4 que está simultáneamente en las clases 1 y 2, lo que indica que este documento contiene palabras asociadas a dos contextos diferentes ver tabla 2.7.

Tabla 2.7. Clasificación.

Clase	Eje principal	Documentos asociados
1	$\mathbf{u}_1^4$	1, 4, 5, 9
2	$\mathbf{u}_2^4$	2, 4, 6, 10
3	$\mathbf{u}_3^3$	3, 7, 8

3. K- medias axial en el análisis de canastas de productos

Con la adaptación del **KMA** se pretende brindar una herramienta que optimice el proceso de obtención de información de grandes bases asociadas al análisis de canastas que supere las deficiencias que presentan otros métodos usados, aquí básicamente se tienen tres propósitos fundamentales (que permiten resumir el contenido y presentarlo a los usuarios de esta información) que son:

i. Agrupar los datos simultáneamente por tipos de canasta y por tipos de productos. Dado un producto o una canasta permite identificar el tipo al que corresponde (es decir, la lista de canastas y productos cercanos) y otros tipos diferentes en los cuales puede aparecer.

ii. En un sólo paso por los datos identificar conjuntos de productos frecuentes altamente correlacionados. Estos conjuntos disminuyen el número de registros que se deben verificar para la identificación de reglas de asociación.

iii. Determinar una medida de ponderación con relación al grupo para cada uno de los productos en la tabla. Dicha medida se puede utilizar para identificar reglas de asociación verdaderamente relevantes.

En el caso de análisis de canastas el **KMA** considera un conjunto de datos como una nube de puntos en un espacio geométrico donde cada dimensión corresponde a un producto. Las clases se representan por medio de vectores apuntando hacia las zonas de nubes de alta densidad. La figura 3.1 muestra lo que podría ser un ejemplo de un conjunto de canastas caracterizadas por tres productos I_1, I_2 y I_3. Estos productos (o ítems) definen el espacio $\boldsymbol{R}^3$. La pertenencia de una canasta a una clase $\mathbf{W}_k$ se determina en función del valor de su proyección sobre el eje ut k que representa la clase. En la figura 3.1 (derecha) se observa que la proyección de la canasta *ii* en el eje $\mathbf{u}_k^t$ es mayor que la de la canasta *i* sobre este mismo eje. El valor de la proyección corresponde con un orden de "tipicidad", las canastas de una clase provienen de un tipo ideal de canasta que se coloca exactamente en el eje imaginario de la clase en el espacio geométrico (Domenges,1979).

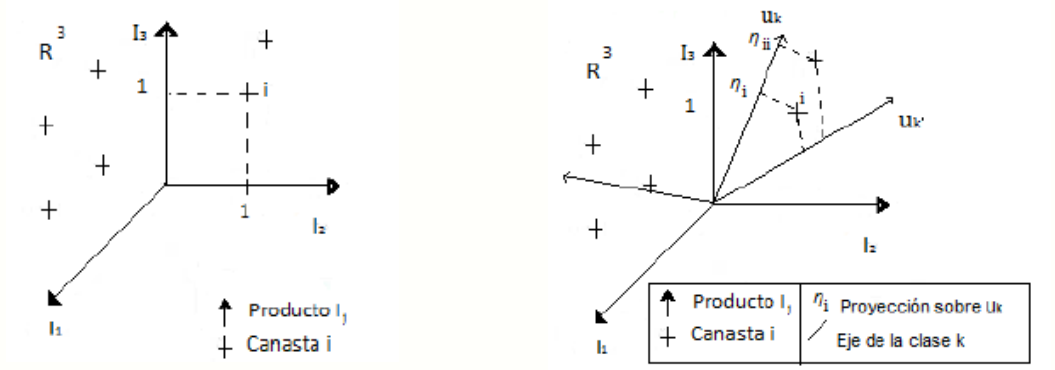


Figura 3.1. A la izquierda canastas caracterizadas por tres productos y a la derecha proyecciones de las canastas sobre los ejes.

3.1 Definiciones generales para el uso del KMA en el análisis de canastas de productos.

Dada una matriz **X** binaria asociada a una tabla datos, la matriz **W** dada en la sección 2, **I** el conjunto de ítems en la tabla de datos y **M** el conjunto de filas de **X** (como están definidos en la sección 1) se tiene:

Definición 5. Para cada clase conformada mediante el método **KMA** se define el $k-$ ésimo eje central de clase $k = 1,2, \dots, K$, como un vector de R^p que se obtiene (eje inicial de la clase k modificado) cuando pasa el algoritmo **KMA** por todas las filas de la matriz **W**; $\mathbf{u_k^t} = \left\{u_{k_1}^t, u_{k_2}^t, \dots, u_{k_p}^t\right\}, t = 1,2, \dots$

Definición 6. Dado $\mathbf{u_k^t}$ el eje de central de la clase *k*, se define su *j-ésima* coordenada $u_{k_j}^t$como un indicador continuo de la medida de la relación del *j-ésimo* ítem con la $k-$ $ésima$ clase y de la relación entre las diferentes clases de canastas.

Definición 7. Se define la matriz de ejes centrales de clase **uk** como la matriz de tamaño $K \times p$ cuyas filas son los ejes centrales de clase obtenidos después del paso del **KMA** por todas las filas de **W**.

Definición 8. Sea $A \subseteq \mathbf{I}$ un conjunto de ítems. Se dice que el conjunto A_k caracteriza la $k-ésima$ clase, si para todo $I_j \in A_k$, $u_{k_j}^t$ es mayor o igual al $\varphi\%$ los valores de las componentes del eje central de la clase.

Observación. El umbral $\varphi\%$ indica a partir de cuál valor un ítem se considera importante para la clase, este umbral lo selecciona el usuario.

Definición 9. Dada la matriz de ejes centrales de clase **uk**, se define la matriz binaria **D** de ítems relevantes o característicos de las clases como la matriz de tamaño $K \times p$ cuya $kj-ésima$ coordenada es igual a 1 si el $j-ésimo$ ítem alcanza el umbral φ, y es igual a cero sí no lo alcanza.

Notación. C_{r_k} representa la colección de conjuntos de ítems de tamaño *r*, relevantes en la *k-ésima* clase,$r = 2, \dots, p$ y $k = 1, \dots, K$

Definición 10. Sea **X** la matriz binaria asociada a una tabla de datos, se define la matriz de transformación $\mathbf{T_k}$ a los ítems relevantes de la *k-ésima* clase como: la matriz diagonal, cuya diagonal principal es la *k-ésima* fila de la matriz binaria de ejes característicos de clase *k*.

Definición 11. Sea **X** la matriz binaria asociada a una tabla de datos y A_k el conjunto de ítems que caracteriza la *k- ésima* clase dada por el **KMA** para estos datos; se define $\mathbf{X_{A_k}} = \mathbf{XT_k}$ la matriz restringida a los ítems de A_k, como una matriz del mismo tamaño que **X**, cuya *j-ésima* columna es igual a la *j-ésima* columna de **X** si I_j pertenece a A_k, y es una columna nula si I_j no pertenece a A_K, $j = 1, \dots, p$.

La matriz binaria de tamaño $n \times K$ obtenida de la matriz de proyección, cuya componente *ik-ésima* es igual a 1 si la proyección de la canasta *j-ésima* es máxima sobre el eje *k*, y cero en caso contrario se denomina la matriz de pertenencia y se denota **M1**.

3.2 Método propuesto

A partir de las definiciones dadas en este documento se describen los pasos a seguir para aplicar el **KMA** en el análisis de canastas de productos:

1. Normalizar la matriz binaria **X** para obtener la matriz **W**.
2. Aplicar el algoritmo **KMA** sobre **W**.
3. De los resultados dados por el **KMA**, seleccionar la matriz de pertenencia **M1** y la matriz de ejes centrales de las clases **uk**.
4. **Resultado 1**: con la matriz de pertenencia obtener las listas las canastas asociadas a cada una de las clases.
5. A partir de la matriz de ejes centrales de las clases **uk**, construir la matriz de ítems relevantes **D** para un umbral φ específico.
6. **Resultado 2**: de la matriz **D** obtener las listas de los productos asociados a cada una de las clases (también se pueden tener las listas de las clases asociadas a cada uno de los productos).
7. Construir las matrices de transformación $\mathbf{T_k}$ para $k = 1, 2, \dots, K$.
8. Construir las matrices $\mathbf{X_{A_k}}$ restringidas a las clases k, para $k = 1, 2, \dots, K$.
9. **Resultado 3**:Aplicar el algoritmo apriori sobre cada una de las matrices de datos restringidas $\mathbf{X_{A_k}}$, con el fin de generar las reglas de asociación para los ítems asociados a cada una de las clases k, para $k = 1, 2, \dots, K$.

Figura 3.2. Pasos del método propuesto para el análisis de canastas de productos

4. Aplicación

En esta sección se ilustra el uso del método propuesto sobre una tabla de datos real y al final se comparan los resultados con las reglas de asociación sin el uso del KMA. En la aplicación se utiliza la tabla de datos que se llamó "canastas" la cual contiene el registro de 47 productos relacionados en 1000 compras realizadas un supermercado local (Bogotá, Colombia) en el año 2009, figura 4.1. La base original además contiene algunos datos de información socio-demográfica de los clientes como: sexo, estado civil, estrato, profesión, situación laboral, tiempo de permanencia en el trabajo y edad. A continuación se presentan los resultados del **KMA** sobre los datos de canasta (en el programa **R** se aplicaron las funciones **kms.a** y **kms.ae**, que se definen adelante)

P01	pan empacado	P02	champú cabello
P03	golosinas	P04	productos faciales
P05	café instantáneo	P06	medicamentos para el dolor
P07	esmalte para uñas	P08	chicles
P09	pasa bocas en paquetes	P10	comida para perros y gatos
P11	complementos vitamínicos y calcio	P12	absorbentes para incontinencia
P13	jabón de tocador	P14	colonias
P15	agua envasada	P16	jugos envasados
P17	pañales desechables	P18	maquillaje de ojos
P19	cerveza	P20	queso para untar
P21	tónico de limpieza facial	P22	galletas no dulces
P23	helados	P24	bebidas derivados lácteos
P25	pasa bocas dulces	P26	postres envasados
P27	té	P28	lustra muebles
P29	toallitas limpieza íntima	P30	chocolatinas
P31	café molido o entero	P32	aderezo para ensaladas
P33	polvos faciales	P34	cereal
P35	aceite de cocina	P36	aguardiente
P37	endulzantes	P38	verduras encurtidas o enlatadas
P39	desodorante antitranspirante	P40	bebida energizante
P41	maltas	P42	tampones
P43	gaseosas	P44	productos para limpieza
P45	refrescos en polvo	P46	gelatina
P47	atún enlatado		

Figura 4.1. Productos relacionados en "*canastas*"

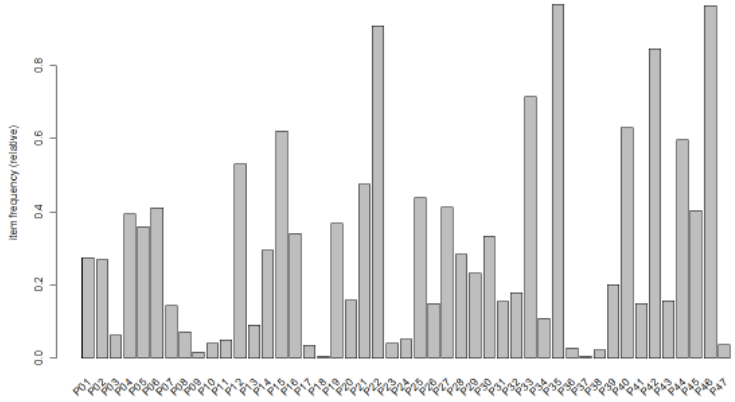


Figura 4.2. Frecuencia de aparición de los productos en la tabla de datos.

4.1 Funciones del programa estadístico R

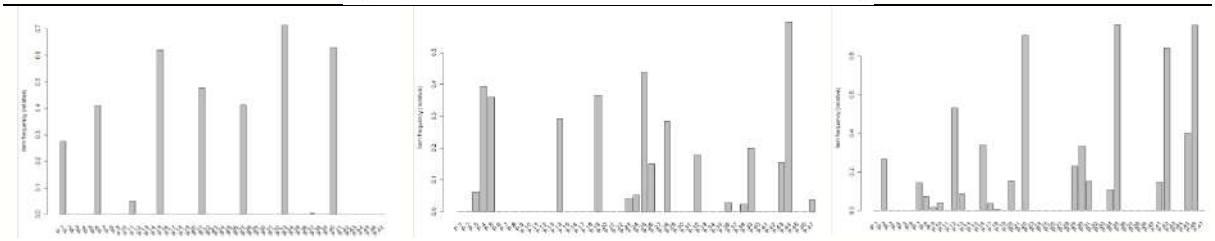
El algoritmo principal está dentro de la función llamada **kms.a**, que depende de los parámetros de la matriz binaria normalizada **W** y de los *K* ejes iniciales $\mathbf{uk}^0$. Ésta función produce las cuatro matrices **M**, **M1**, τ y **uk**. Adicional a la función principal, se crearon en **R** algunos objetos matemáticos adicionales que se describen en el procedimiento. En este trabajo también se usaron las funciones del paquete ***arules*** de **R**, para la generación de reglas de asociación.

4.2 Agrupación de canastas

El **KMA** generó tres clases: la por primera conformada 450 canastas, la segunda por 135 canastas y la tercera por 414 canastas. En los procesos para tomar de decisiones, las listas de grupos de canastas similares se pueden tomar como parte del marco de trabajo (Vercellis 2009) (en el programa **R** esta lista se obtiene mediante el uso de la función as (,"*transactions*") de **arules**). Los ítems que satisfacen un umbral de pertenencia de $\varphi =$ 30%.

Tabla 4.1. Ítems característicos en las clases

No. clase	Ítems
1	1 6 11 15 21 27 33 37 40
2	3 4 5 14 19 23 24 25 26 28 32 36 38 39 43 44 47
3	2 7 8 9 10 12 13 16 17 18 20 22 29 30 31 34 35 41 42 45 46



a. b. c.

Figura 4.3. Frecuencia de aparición de los productos característicos de las clases 1, 2 y 3 respectivamente como a, b y c.

Tabla 4.2. Productos de mayor frecuencia en las clases

No. clase	Productos
1	*agua envasada*($P15$), *polvos faciales*($P33$) y *bebidas energizantes*($P40$)
2	*productos faciales*($P04$), *café instantáneo*($P05$), *cerveza*($P19$), *pasabocas dulces*($P25$) y *productos para la limpieza*($P44$)
3	*gelatina*($P46$), *tampones*($P42$), *aceite de cocina*($P35$) y *galletas no dulces*($P22$)

Las tablas 4.1 y 4.2, junto en la figura 4.3 muestran el comportamiento de los productos característicos de las clases.

4.3 Alguna reglas de asociación para la clase uno con soporte de 0.05 y confianza 0.7.

Con la matriz de ítems restringidos, el algoritmo apriori trabaja sólo sobre las columnas de la matriz que corresponden a los ítems de la clase y las canastas relacionadas con ésta.

Tabla 4.3. 10 de las 43 reglas de asociación para los productos de la clase uno.

No. Regla	antecedente		consecuente	soporte	confianza
1	$\{\}$	$\rightsquigarrow$	$\{P33\}$	0.713	0.7130
2	$\{P06\}$	$\rightsquigarrow$	$\{P33\}$	0.289	0.7032
3	$\{P21\}$	$\rightsquigarrow$	$\{P15\}$	0.362	0.7605
4	$\{P21\}$	$\rightsquigarrow$	$\{P33\}$	0.351	0.7374
5	$\{P40\}$	$\rightsquigarrow$	$\{P33\}$	0.452	0.7175
6	$\{P15\}$	$\rightsquigarrow$	$\{P33\}$	0.461	0.7435
7	$\{P01, P21\}$	$\rightsquigarrow$	$\{P15\}$	0.108	0.7606
8	$\{P21, P27\}$	$\rightsquigarrow$	$\{P15\}$	0.157	0.7659
9	$\{P06, P21\}$	$\rightsquigarrow$	$\{P15\}$	0.177	0.8389
10	$\{P06, P21\}$	$\rightsquigarrow$	$\{P33\}$	0.152	0.7204

De forma similar se pueden ver gráficamente las reglas asociadas a cada clase.
Para el grupo uno se conformó 43 reglas de asociación. Las canastas características de esta clase son las que incluyen polvos faciales (P33). Las reglas muestran las formas de asociación entre el producto polvos faciales (P33) y otro productos como medicamento para el dolor (P06), bebidas energizantes (P40), productos para la limpieza facial (P04), agua envasada (P15) o té (P27). Las reglas 1, 5 y 6 son las de mayor soporte y confianza, indican que los productos que están más asociados a los polvos faciales (P33) son las bebidas energizantes (P40) y el agua envasada (P15). Los ítems que tienen menos frecuencia de aparición generan reglas de asociación que tienen bajo nivel de soporte.
Para la clase dos se conformaron 26 reglas de asociación. Este grupo está caracterizado por incluir productos para la limpieza (P44). Otros productos café instantáneo (P05), lustra muebles (P28), desodorante anti-transpirante (P39), productos faciales (P04), pasa-bocas dulces (P25) y cerveza (P19). La regla de mayor nivel de soporte y confianza en este grupo es la que indica que el café instantáneo (P05) implica la inclusión en la canasta de productos para la limpieza (P44).
En la clase tres se conformaron 1588 reglas de asociación de tamaños 1 a 7. Los productos más representativos de la clase tres son la gelatina, tampones (P42), aceite de cocina (P35) y galletas-no dulces (P22). Otros que tienen menos frecuencia de aparición son queso-crema (P20), jugos envasados (P16), refrescos en polvo (P45), maltas, café molido o entero (P31), cereal (P34), absorbentes para incontinencia (P12), jabón de tocador (P13), chicles (P08), esmalte de uñas (P07) o champú (P02).

La tabla 4.4 resume el número de reglas con diferentes umbrales de soporte y confianza dados por el algoritmo apriori sobre el conjunto de datos brutos y sobre las clases conformadas por el KMA.

Tabla 4.4. Número de reglas KMA versus datos brutos

	Sop =0.05 Conf=0.60	Sop =0.10 Conf=0.60	Sop =0.20 Conf=0.60	Sop =0.05 Conf=0.70	Sop =0.10 Conf=0.70	Sop =0.20 Conf=0.70
Clase1	120	76	26	43	29	11
Clase2	46	14	4	26	8	2
Clase3	642	642	197	1588	640	197
Total clases	**808**	**732**	**227**	**1657**	**677**	**210**
Total datos	**74437**	**17876**	**3211**	**60443**	**14430**	**2507**

Como se puede evidenciar por los resultados dados hasta el momento, el KMA no solo disminuye notablemente el número de reglas de asociación que se deben analizar, sino que aporta información adicional que facilita análisis de éstas, para su uso en el estudio del comportamiento de canastas de productos.

5. Conclusiones

Desde los resultados se puede concluir que el uso del **KMA** es una buena alternativa que mejora, agiliza y facilita el análisis de canastas de productos, ya que permite obtener:

- Listas de canastas caracterizadas por incluir grupos de productos particulares. Una canasta puede pertenecer a más de una clase.
- Listas de productos característicos de clases de canastas particulares. Un producto puede estar en más de una clase de canasta.
- Listas de clases que incluyen un producto particular.
- Indicadores continuos de relación de un producto con la clase.
- Clasificación sencilla del conjunto de datos. El **KMA** en un sólo paso por los datos genera una muy buena clasificación de éstos.
- Ponderación de los ítem en las clases, lo cual se puede utilizar para identificar reglas con mayor importancia que otras (CAI,1998).
- Conjuntos de productos frecuentes altamente correlacionados, lo que disminuye tanto el proceso de búsqueda de conjuntos de ítem frecuentes, como el número de reglas de asociación que se generan y se analizan.

6. Referencias

- [Domenges,1979] Domenges, D. Analyse factorielle sphérque: une exploration. Comptes trimestriels de la Direction des Synthèses économiques de I'INSEE, vol 35, pág 3-43, Paris, 1979

- [Lelu, 1993] Lelu, A. Modeles Neuronaux Pour l'Analyse de Donnees Documentaires et Textuelles. Tesis de Doctoral. Spécilité Mathématique Statistique. Universite Paris 6,1993.
- [Hertz,1995] Hertz, J. and Krogh, A. and Palmer,R. Introduction to the Theory of Neural Computation, vol 1. Santa Fe Institute Studies in the Sciences of Complexity. Addison Wesley, 1995.

- [Morato,1999] Morato L, J.. Análisis de relaciomes cienciométricas y linguisticas en un entorno automatizado. Tesis de Doctoral.Facultad de humanidades, comunicación y documentación. Madrid,1999.

- [Han, 2001] Han, J. and Kamber, M. Data Mining,Concepts and Techniques. Morgan Kaufmann Publisher. New York, 2001

- [Hernández, 2005] Hernández,J. and Ramírez, M. and Ferri, C. Introducción a la Minería de Datos .Prentice Hall, 2005.
- [Narros,2007] Narros, J. Segmentación de mercados de consumo con criterios relacionales: aplicación a la compra de alimentación en hipermercados. Universidad Complutense de Madrid. Tesis de Doctoral. Facultad de Ciencias Económicas y Empresariales. Departamento de Comercialización e Investigación de Mercados, Madrid, 2007.
- [Vercellis, 2009] Vercellis, C. Business intelligence: Data mining and optimization for decision making, Wiley,UK, 2009