

Sistema Híbrido para Resolver Problemas de Clusterización en un Proceso de Data Mining Empleando Lógica Difusa y Redes Neuronales Auto-organizativas

Arturo Palacios-Aguilera¹
Nelly Condori-Fernández²
Michel Quintana Truyenque³

Instituto de Informática de la Universidad Nacional de San Agustín
Parque Industrial S/N, Arequipa-Perú

{artpagui,miquitru}@hotmail.com^{1,3}
nelly@unsa.edu.pe²

Resumen.-

Este trabajo tiene como propósito aprovechar las bondades de la lógica difusa, para optimizar la clusterización en un proceso tradicional basado en redes neuronales para un proceso de Data Mining. Para ello, se implementa un método difuso denominado Fuzzy C-Means que unido al clasificador adaptativo de patrones (SOM - LVQ) conduce a una agrupación más precisa.

Palabras Clave: Clusterización, Data Mining, Lógica Difusa, Redes Neuronales Auto-organizativas, Fuzzy C-Means.

1.- Introducción

Hoy en día se destina gran cantidad de recursos a la adquisición, almacenamiento, procesado y análisis de la información. Sin embargo, el conocimiento más valioso suele estar oculto entre los datos recogidos. El término Data Mining (Minería de Datos) refleja bien la idea de adquirir un conocimiento valioso en medio de grandes cantidades de información. Pero, hoy en día la preocupación principal ya no es sólo la adquisición del conocimiento, sino la delimitación de su alcance y su validez; necesitamos asignar un grado de certeza al conocimiento, “saber en qué medida conocemos algo”[1].

El término *descubrimiento de conocimiento en bases de datos (KDD- knowledge discovery in databases)* empezó a utilizarse en 1989 para referirse al amplio proceso de búsqueda de conocimiento en bases de datos, y para enfatizar la aplicación a “alto nivel” de métodos específicos de minería de datos [2],[3]. Se puede decir que la minería de datos es la parte fundamental, en la que más esfuerzos se han realizado, como puede verse en [4], [5], [6]. Shalvi, utiliza el mapa auto-organizativo de Kohonen (SOM) para clasificar un conjunto específico de datos médicos, con la finalidad de obtener conocimiento para el apoyo en las decisiones médicas [4]._Konig [5], combina dos métodos para la visualización y análisis de datos, el mapa auto-organizativo de Kohonen (SOM) [3] y el mapeo no lineal de Sammon [7], esta arquitectura ofrece ventajas para el análisis de nuevos datos de entrada, sin embargo la clasificación aún sigue siendo rígida. Quintana [6], combina la utilización del algoritmo de agrupamiento fuzzy c-means y el mapa auto-organizativo de Kohonen (SOM), en el cual se aprovecha este algoritmo difuso para definir los grados de pertenencia de un objeto a cierto número de grupos y luego realizar una clasificación más real mediante el mapa auto-organizativo. No obstante, podría aprovecharse mejor el modelo SOM empleando el modelo LVQ para optimizar el resultado obtenido en la clasificación.

En el presente trabajo, se presenta una arquitectura basada en la combinación de algoritmos de clusterización difusa y el modelo de red neuronal Kohonen. Dicho modelo, presenta dos variantes:

1. Modelo de red neuronal SOM (Self-Organizing Maps).- Con el fin de extraer las características más representativas de los datos difusos.
2. Modelo de red neuronal LVQ (Learning Vector Quantizer) con el objetivo de afinar la clasificación de los resultados obtenidos del mapa característico.

La implementación del sistema fue desarrollado en Visual C++ 6.0, aprovechando la potencia del mismo en el desarrollo de sistemas orientados a objetos.

El artículo esta estructurado de la siguiente manera: En la sección 2, se presenta conceptos previos sobre el que se basa el trabajo de investigación. Luego en la sección 3, se describe detalladamente cada uno de los componentes de la arquitectura del sistema de data mining. En la sección 4, se explica brevemente conceptos de segmentación de mercado. En la sección 5, se muestran los resultados computacionales aplicados a un caso de estudio. Finalmente, se presentan las conclusiones del trabajo.

2.- Conceptos Previos

En la presente sección, se explica brevemente algunos conceptos básicos relacionados a redes neuronales y lógica difusa.

2.1.- Redes Neuronales

Las redes neuronales artificiales son modelos matemáticos inspirados en sistemas biológicos, adaptados y simulados en computadoras convencionales [8]. Existen varios modelos que son aplicados en la solución de diversos problemas. A continuación, explicaremos brevemente el Modelo de Kohonen.

2.1.1.- Modelo de Kohonen

Teuvo Kohonen presentó en 1982 un sistema con un comportamiento semejante a la autoorganización de las neuronas en determinadas zonas del cerebro. Es decir, plantea un modelo de red neuronal con capacidad para formar *mapas de características* de manera similar a como ocurre en el cerebro [8].

Este modelo tiene dos variantes, denominadas LVQ (Learning Vector Quantization) y SOM (Self-Organizing Map). Ambas se basan en el principio de formación de mapas topológicos para establecer características comunes entre las informaciones (vectores) de entrada a la red, aunque difieren en las dimensiones de éstos, siendo de una sola dimensión en el caso de LVQ, y bidimensional, e incluso tridimensional, en la SOM [3],[8].

La arquitectura de la LVQ del modelo de Kohonen es como sigue:

- Cada una de las m neuronas de entrada se conecta a las l neuronas de la capa de salida a través de conexiones hacia delante (feedforward).
- Entre las neuronas de la capa de salida, puede decirse que existen conexiones laterales de inhibición (peso negativo) implícitas, pues aunque no estén conectadas, cada una de estas neuronas va a tener cierta influencia sobre sus vecinas. La influencia que una neurona ejerce sobre las demás es función de la distancia entre ellas, siendo muy pequeña cuando están muy alejadas.

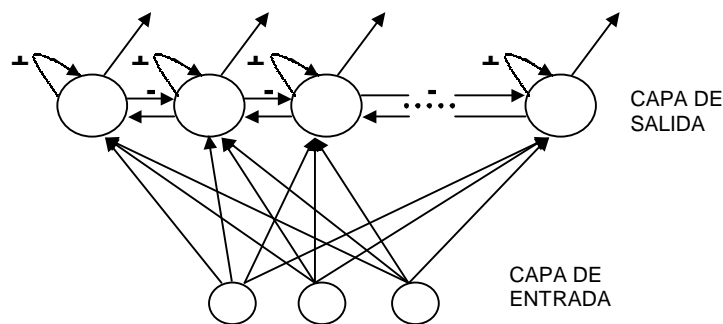


Fig. 1 Arquitectura LVQ

Por otro lado, la versión del modelo denominado SOM es como sigue:

- Trata de establecer una correspondencia entre los datos de entrada y un espacio bidimensional de salida, creando mapas topológicos de dos dimensiones.
- Poseen una arquitectura de dos capas (una capa de entrada y una capa de salida bidimensional), flujo de información unidireccional (feedforward).
- La interacción lateral entre las neuronas de la capa de salida sigue existiendo, aunque ahora hay que entender la distancia como una zona bidimensional que existe alrededor de cada neurona.

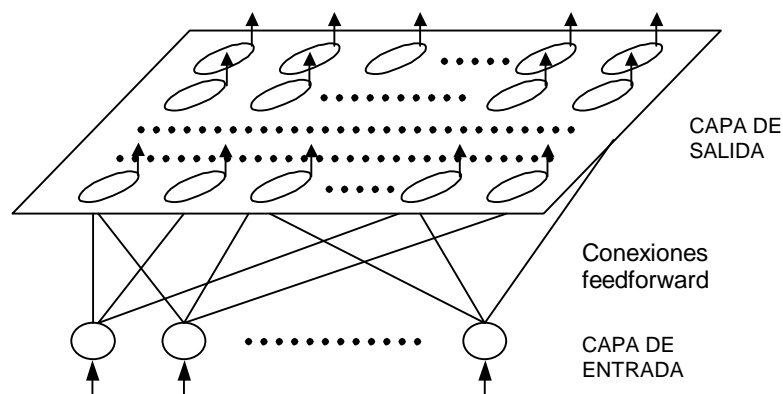


Fig. 2. Arquitectura SOM

2.2.- Fuzzy C-Means

La idea básica de la teoría de conjuntos difusos (*fuzzy set theory*) es que a diferencia de la teoría clásica de conjuntos un elemento puede ser miembro de un conjunto con un grado de pertenencia que normalmente es un número entre 0 y 1 [9]. La noción de un conjunto difuso se define de la siguiente manera:

Sea X un conjunto convencional de elementos. Un conjunto difuso se define como A^\wedge sobre X de tal manera que:

$$A^\wedge = \{(x, \mu_{A^\wedge}(x)) | x \in X\}$$

La expresión $\mu_{A^\wedge}(x)$ se denomina grado de pertenencia del elemento x al conjunto difuso A^\wedge , llamándose μ_{A^\wedge} la función de pertenencia, que normalmente asume valores entre 0 y 1.

A base de estas definiciones elementales de un conjunto difuso, en el pasado se han desarrollado muchas teorías y aplicaciones en el área de lógica difusa, un resumen de las cuales se puede encontrar en [10]. Un método difuso, de especial interés para el presente trabajo, es el algoritmo de agrupación Fuzzy C-Means. Una presentación detallada del Fuzzy C-Means se encuentra en [11], y las aplicaciones de la agrupación difusa se describen, por ejemplo, en [12].

El algoritmo Fuzzy C-Means asigna un conjunto de objetos, caracterizados por sus respectivos valores de atributos, a un número determinado de clases. Como resultado del Fuzzy C-Means, cada objeto tiene un grado de pertenencia a cada clase, representada por su centro de clase. Básicamente, el algoritmo Fuzzy C-Means se realiza aplicando los siguientes cuatro pasos:

Paso 1: Inicialización

Utilizamos la siguiente notación:

- Número de clases a encontrarse: c
- Número de objetos a agrupar: J
- Vector de atributos del objeto j : x_j , $j = 1, \dots, J$
- Grado de pertenencia del objeto j a clase i : u_{ij} $i = 1, \dots, c$; $j = 1, \dots, J$.

Sea $A^{(0)}$ una matriz ($c \times J$) con el elemento u_{ij} en posición (i,j) , $i = 1, \dots, c$; $j = 1, \dots, J$.

Esta matriz se inicializa en forma aleatoria con la siguiente restricción:

$$\sum_{i=1}^c \mu_{ij} = 1, \quad \forall j = 1, \dots, J$$

Paso 2: Cálculo de Centros de Clase

Dados los valores de pertenencia u_{ij} los centros v_i de cada clase i están dados por:

$$v_i = \frac{\sum_{j=1}^J (\mu_{ij})^m x_j}{\sum_{j=1}^J (\mu_{ij})^m}, \quad \forall i = 1, \dots, c$$

El parámetro m , utilizado en la formula anterior, se llama difusor (*fuzzifier*) y determina el grado de difusión (*fuzziness*) para las clases encontradas ($1 < m < \infty$). Para m “cercano a 1” se calcula una solución con clases no-difusas (*crisp*); mientras mayor sea m más difusa se hace la solución.

Paso 3: Actualización de valores de pertenencia

Dados los centros de clase calculados en el paso 2, los valores de pertenencia u_{ij} son actualizados utilizando la siguiente fórmula:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}}, \quad \forall i = 1, \dots, c \quad \forall j = 1, \dots, J$$

El valor d_{ij} es la distancia entre el objeto j y el centro de clase i (v_i). En el cálculo de esta distancia se utilizan los centros de clase i (v_i), obtenidos en el paso 2.

Paso 4: El Criterio de Detención (*Stopping*)

Los pasos 2 y 3 se repiten en forma iterativa hasta cumplir con el siguiente criterio de detención:

$$\|A^{(t+1)} - A^{(t)}\| < \varepsilon$$

A es la matriz de los valores de pertenencia en la iteración t ; ε es un umbral a ser determinado por el usuario. Las siguientes dos condiciones deben cumplirse para asegurar la convergencia del algoritmo Fuzzy C-Means:

$$\sum_{i=1}^c \mu_{ij} = 1, \quad \forall j = 1, \dots, J$$

$$\mu_{ij} \in [0,1]; i = 1, \dots, c; j = 1, \dots, J$$

En términos de resultados, el algoritmo Fuzzy C-Means rinde centros de clase v_i para las c clases, así como los valores de pertenencia de cada objeto a cada clase u_{ij} .

En la aplicación de segmentación de mercados por ejemplo, los clientes son los objetos considerados, que se describen mediante atributos tales como por ejemplo edad e ingreso. Aquí también se mostrará también cómo los parámetros de evaluación pueden ser usados para encontrar un número adecuado de clases.

3.- Arquitectura del sistema de minería de datos

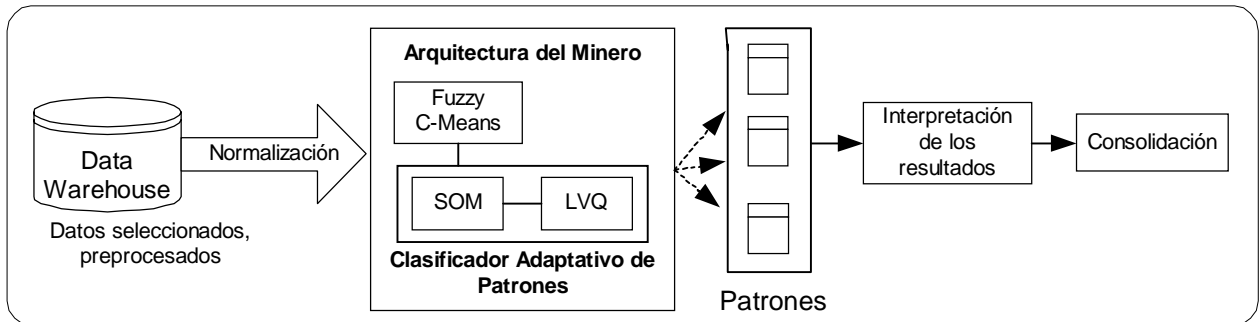


Figura 3. Esquema general del modelo propuesto

El presente esquema presenta algunos pasos del proceso de KDD, los cuales los podemos agrupar en tres etapas: pre-procesamiento (selección de datos objetivo y normalización), procesamiento (proceso de minado de datos) y el post-procesamiento (generación de patrones, interpretación de los resultados y consolidación) (vea Figura 3).

Seguidamente nos abocaremos a explicar la arquitectura del minero por ser el tema central del presente trabajo. En dicha arquitectura interactúan tres componentes: el primer componente está encargado de asignar el grado de pertenencia de un objeto j a cada una de las c clases definidas previamente. Esto se logra mediante el algoritmo Fuzzy C-Means.

El segundo componente es un mapa auto-organizativo (SOM) cuyos datos de entrada son el conjunto de datos difusos (grados de pertenencia), el objetivo es resumir o concentrar las características esenciales de los datos de entrada en un mapa bidimensional. El cual finalmente interactúa con la red neuronal LVQ para obtener una

clasificación final con mayor exactitud. La combinación del modelo SOM y del modelo LVQ, se constituye como un modelo híbrido denominado Clasificador Adaptativo de Patrones [13].

4.- Segmentación de mercados

“La segmentación de mercados es un proceso mediante el cual se identifica o se toma a un grupo de compradores homogéneos, es decir, se divide el mercado en varios submercados o segmentos de acuerdo a los diferentes deseos de compra y requerimiento de los consumidores”. [14]

Tal como se ve en la definición, la segmentación de mercado es primero que nada un proceso. Eso significa que la segmentación de mercado no es una actividad que se realiza una sola vez en la empresa y que acaba inmediatamente después, sino que es una actividad constante. Segundo, es importante remarcar que la segmentación consiste en identificar grupos y no en crearlos. Tercero, los segmentos se crean en función de las características de los consumidores y no en función de los productos que los satisfacen. [15]

Por lo tanto la segmentación de mercados se justifica en el hecho de que permite un mejor aprovechamiento de los recursos de la empresa y de la sociedad, a la vez que incrementa la satisfacción de los consumidores.

4.1.- Proceso de Segmentación

Para segmentar un mercado se requiere seguir un proceso relativamente simple que consiste en delimitar nuestra área de mercado, identificar las variables de segmentación que corresponden a las necesidades que nuestros productos satisfacen, realizar la segmentación de los mercados a base de estas variables y luego hacer un resumen de las características generales de cada segmento.

Proceso de segmentación[15]:

- Delimitación del área de mercado;
- Identificación de las variables de segmentación;
- Segmentación en función de las variables identificadas;
- Identificación de las características de cada segmento.

Las variables de segmentación más comúnmente utilizadas en el mercado individual son las siguientes[15]: Demográficas, Socioeconómicas, Psicoográficas, Por tipo de uso, Estilos de vida.

Con la finalidad de demostrar mediante resultados el desempeño de la arquitectura propuesta, se tomará un caso de estudio sobre segmentación de mercados, que se explica a continuación.

5.- Caso de Estudio

Este estudio tiene la finalidad de ayudar a conocer hacia qué clientes se va a dirigir una determinada presentación del producto (palomitas de maíz). Se trabajó con un total de 823 registros, donde cada registro representa los atributos del cliente (edad, sexo, condición, frecuencia de compra, consumo, lugar de compra, presentación del producto). Del conjunto de características se consideraron tres variables (características): el lugar de compra, la frecuencia de compra y el consumo del producto.

	Lugar de Compra	Frecuencia de Compra	Consumo por persona
GRUPO 1	Autoservicios y Tiendas-Bodegas	Mensual	1695 gr
GRUPO 2	Tiendas-Bodegas	Esporádica	702,7 gr
GRUPO 3	Mercado	Mensual	1283 gr

Durante esta etapa se emplea la combinación del algoritmo Fuzzy C-Means y el Clasificador Adaptativo de Patrones para procesar las características del cliente. Se realizaron pruebas para encontrar los grupos más representativos, con la ayuda del experto, mediante la interpretación de los resultados obtenidos.

Por lo tanto se definieron tres grupos que corresponden a los siguientes tipos de presentación :

- En el **Grupo 1** se encuentran clientes que compran generalmente en Autoservicios y en segundo lugar en Tiendas y Bodegas, realizándolo mensualmente. Dadas estas características, concluye el experto que la presentación dirigida a esta clase va ser un producto envasado en **Taper**.
- En el **Grupo 2** se encuentran clientes que compran en Tiendas y Bodegas, realizándolo esporádicamente. Dadas estas características, concluye el experto que la presentación dirigida a esta clase va ser un producto envasado en **Caja**.
- En el **Grupo 3** se encuentran clientes que compran en Mercados, realizándolo mensualmente. Dadas estas características, concluye el experto que la presentación dirigida a esta clase va ser un producto envasado en **Bolsa**.

Dichos grupos obtenidos se pueden apreciar en el mapa auto-organizativo generados por el sistema. Este mapa bidimensional (de 14*14) está conformado por 196 neuronas.

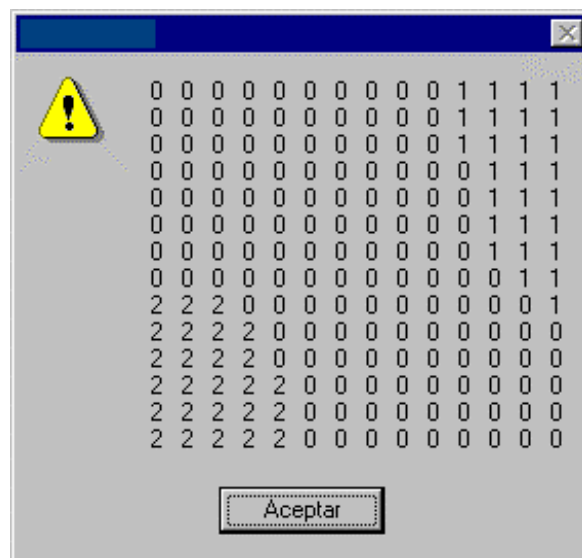


Fig. 4. Mapa bidimensional, que muestra los tres grupos generados por el sistema

Durante el procesamiento de los datos, cabe resaltar el resultado intermedio obtenido por el algoritmo Fuzzy C-Means el cual presenta una tabla de datos que indica el grado de pertenencia de cada cliente a cada uno de los grupos encontrados.

Cliente	Grado de pertenencia al Grupo 1	Grado de pertenencia al Grupo 2	Grado de pertenencia al Grupo 3
7	0,6132	0,2084	0,1785
8	0,3182	0,3669	0,315
9	0,9365	0,0342	0,0293
10	0,6142	0,2078	0,178
11	0,0386	0,9283	0,0331
12	0,3354	0,3337	0,3309

Este tratamiento difuso nos permitirá conocer cuáles son las tendencias de cada cliente hacia cada presentación del producto. Por consiguiente, para la siguiente etapa utilizamos sólo el Clasificador Adaptativo de Patrones para conseguir el comportamiento de los clientes, aplicando combinaciones distintas de las tres presentaciones propuestas. Para ello agrupamos los grados de pertenencia obtenidos en siete grupos distintos para determinar qué presentación se va a dirigir a los clientes. Dicho número de grupos resulta de: $2^n - 1$, donde n representa el número de presentaciones del producto.

Mostramos en la tabla los resultados arrojados por el Clasificador Adaptativo de Patrones que trabajó en base a la tabla de grados de pertenencia (parte de ella presentada anteriormente).

	Grado de pertenencia al Grupo 1	Grado de pertenencia al Grupo 2	Grado de pertenencia al Grupo 3
Tendencia 1	0,138	0,743	0,118
Tendencia 2	--	--	--
Tendencia 3	0,041	0,041	0,919
Tendencia 4	0,185	0,186	0,629
Tendencia 5	0,338	0,327	0,334
Tendencia 6	0,593	0,219	0,188
Tendencia 7	0,936	0,034	0,029

Observando cada tendencia se concluye:

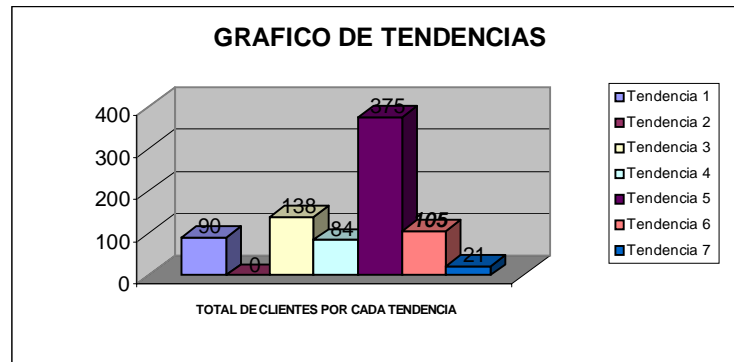
- Tendencia 1, más dirigido al Grupo 2, es decir, hacia la presentación en Caja;
- Tendencia 3, más dirigido al Grupo 3, es decir, hacia la presentación en Bolsa;
- Tendencia 4, más dirigido al Grupo 3 y equilibrado en los dos grupos restantes, con preferencia hacia la presentación en Bolsa;
- Tendencia 5, equilibrado en los tres grupos, posicionando las tres presentaciones;
- Tendencia 6, con mayor influencia en el Grupo 1, seguido en menor grado al Grupo 2, se dirige hacia la presentación en Taper y Caja.
- Tendencia 7, más dirigido al Grupo 1, es decir, hacia la presentación en Taper;
- El sistema muestra la inexistencia de clientes para la Tendencia 2.

De igual manera, estas tendencias se pueden apreciar en el mapa auto-organizativo generados por el sistema.



Fig. 5 Mapa bidimensional, que muestra los siete grupos generados por el sistema

Todos estos resultados son presentados y almacenados por el sistema, brindando información resumida y gráficos estadísticos en cuanto a la cantidad y comportamiento de los clientes en cada grupo. El siguiente gráfico muestra el número de clientes para cada una de las tendencias referidas a la presentación del producto.



Primeramente definimos el concepto de mercado para posteriormente tratar la segmentación del mercado.

5.- Conclusiones

El emplear sólo el modelo de red neuronal de Kohonen implica obtener una agrupación “rígida”, es decir, que un determinado objeto va a pertenecer exclusivamente a un grupo específico.

Por otro lado, aprovechar el concepto de grado de pertenencia de la lógica difusa nos permite realizar agrupaciones flexibles, es decir, que un determinado objeto va a pertenecer a diferentes grupos con ciertos grados de pertenencia.

Con el Clasificador Adaptativo de Patrones se logra cierto incremento en la exactitud y performance para la generación automática de grupos.

En cuanto a la definición del número de neuronas en la capa de salida del modelo SOM resulta ser un proceso arbitrario por parte del usuario. Esto se debe, a que aún se sigue siendo trabajo de investigación para los científicos del área asegurar un número adecuado de neuronas para una determinada cantidad de datos a procesar.

Referencias Bibliográficas

- [1] GALINDO, José y ARANDA, M. Carmen. *Gestión de una Agencia de Viajes usando Bases de Datos Difusas y FSQL*. Dpto. Lenguajes y Ciencias de la Computación Universidad de Málaga.
- [2] U. M. FAYYAD, G. PIATETSKY-SHAPIO y P. SMYTH. *From Data Mining to Knowledge Discovery*. En *Advances in Knowledge Discovery and Data Mining*, Fayyad, Piatetsky-Shapiro, Smyth y Uthurusamy Eds., AAAI Press, Menlo Park, California, 1996.
- [3] KOHONEN, T. Analysis of Processes and Large Data Sets by a Self-Organizing Method. Intelligent Processing and Manufacturing of Materials, 1999. IPMM '99. Proceedings of the Second International Conference on , Volume: 1. Page(s): 27 -36 vol.1 1999
- [4] SHALVI, D., DECLARIS, N. “An Unsupervised Neural network approach to medical data mining techniques”. Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on , Volume: 1. Page(s): 171-176 vol.1, 1998.

- [5] KONIG, A. "*Interactive Visualization and Analysis of Hierarchical Neural Projections for Data Mining*". Neural Networks, IEEE Transactions on, Volume: 11 Issue: 3. Page(s): 615-624, 2000.
- [6] QUINTANA, M., "*Modelo Híbrido para los procesos de Data Mining en el apoyo a la Toma de Decisiones basados en Tecnologías Inteligentes Conexiónistas y Difusas*". SPCM Magazine, de la Sociedad Peruana de Computación. Jul. 2002.
- [7] SAMMON J.W., "*A nonlinear mapping for data structure analysis*". IEEE Trans. Comput., vol. C-18, pp. 401-409, 1969.
- [8] HILERA, J. y MARTINEZ, V., "*Redes Neuronales Artificiales*". 2000. Pág. 9.
- [9] ZADEH, L. A. "*Fuzzy Sets*". Information and Control 8. Págs. 338-353, 1994.
- [10] ZIMMERMANN, H. J. *Fuzzy Set Theory - and Its Applications* 3er edición. Kluwer Academic Publishers, Boston Dordrecht, Londres. 1996.
- [11] BEZDEK, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms* Plenum Press, Nueva York. 1981.
- [12] MEIER, W., WEBER, R. y ZIMMERMANN, H.-J. "*Fuzzy Data Analysis -Methods and Industrial Applications.*" *Fuzzy Sets and Systems* 61, págs. 19-28. 1994.
- [13] HAYKIN, Simon. "*Neural Networks*" *A comprehensive foundation*. EEUU, Ed. Prentice-Hall, Segunda Edición, 1999. Págs. 468-470.
- [14] KOTLER, Philip y ARMSTRONG, Gary. *Fundamentos de Mercadotecnia*. México, Prentice – Hall, Cuarta Edición, 1998.
- [15] ARELLANO, Rolando. *Marketing: Enfoque América Latina*. pág.489. 2000.