"Herramienta software para implementar minería de datos: clusterización utilizando lógica difusa"

SANABRIA GARZÓN J.
Ingeniería de Sistemas
(Recibido: Abril 4 de 2004 - Aceptado: Mayo 31 de 2004)

R E S U M E N

La minería de datos se ha convertido en un área de investigación y desarrollo, desde la cual se proponen técnicas que apuntan a encontrar el conocimiento oculto en grandes colecciones de datos. Estos datos contienen información valiosa, que puede ser usada para mejorar la competitividad de las instituciones dueñas de la información.

La información por descubrir puede tener muchas formas, entre ellas reglas asociativas o grupos de conjuntos denominados (Cluster), si a esto se le suma la capacidad que tiene la lógica difusa de romper con el principio del tercero excluido y permitir la pertenencia de un elemento a varios Cluster, se tiene una metodología útil a la hora de clasificar en grupos el contenido de las bases de datos.

En el presente artículo se presenta la implementación del algoritmo denominado C-Means para la agrupación de datos en conjuntos difusos, como técnica de minería de datos, esta técnica se implementó en el programa SM2D 1.2 Beta (Software Minería Datos Difusa), y se presenta como ejemplo el analisis del rendimiento académico de la asignatura fisiología vegetal.

Palabras Claves: Bases de Datos (BD), Conjuntos Difusos, Cluster, C-Means, Minería de Datos.

A B S T R A C T

The data mining has become an investigation and development area, in which are intending technicals that point to find the hidden information in a huge data. These data contain valuable information that can be used to improve the competitiveness of institutions owners of these data.

The information to discover can have many forms, among them associative rules or groups of denominated sets (Cluster), if to this we add the capacity that has the Fuzzy Logic of breaking up with the third excluded principle and to allow the relevancy from an element to several Cluster, we have a quite useful methodology when classifying in sets the content of the databases.

In this paper is shown the implementation of an algorithm denominated C-Means for the grouping of data in fuzzy sets, this it has been implemented with the development of a denominated program (Software Mining Data Fuzzy) SM2D 1.21.0 Beta.

Key words: Data Bases (BD), Fuzzy Sets, Cluster, C-Means, Data Mining.

INTRODUCCIÓN

Considerando que el conocimiento puede ser visto como una abstracción a un nivel de información encima de los datos, existe la necesidad de áreas de estudio dentro de la computación que traten este asunto como el llamado Aprendizaje de Máquina, el surgimiento de la minería de datos es una forma de conseguir la información oculta que presentan los datos, la mayoría de las veces almacenados en grandes bases de datos, denominadas bodegas de datos.

La lógica difusa es una rama de la inteligencia artificial que permite analizar la información del mundo real en una escala entre falso y verdadero. Los matemáticos dedicados a la lógica definieron un concepto clave: "Todo es cuestión de grado", los sistemas difusos son una nueva alternativa a las nociones de pertenencia y Lógica clásicos [3].

El presente trabajo se centra en la utilización de un algoritmo en el desarrollo de una tarea clásica de minería de datos como es la de agrupamiento, saliendo de las tendencias estadísticas y manuales con la que se ha estado haciendo, el principal objetivo de este es utilizar un algoritmo fuzzy C-means para ayudar a solucionar el problema de asignación estática de patrones a una clase específica, esto es muy común en aplicaciones reales donde no se puede modelar el mundo solamente con una agrupación estática, y se necesita también manejar información borrosa, para mejorar el análisis e interpretación de la información encontrada, para la generación de conocimiento y apoyo a la toma de decisiones.

CONJUNTOS CLÁSICOS

Se toman algunos aspectos de la teoría de conjuntos convencionales (Conjuntos Concretos), y a partir de allí se hace una extensión a los conjuntos difusos:

Un conjunto concreto se define como una colección de elementos que existen dentro de un Universo. Así, sí el universo consta de los números enteros no negativos menores que 10:

$U=\{0,1,2,3,4,5,6,7,8,9\}$

Entonces podemos definir algunos conjuntos como, por ejemplo:

$A=\{0,2,4,6,8\}$
$B=\{1,3,5,7,9\}$

Con estas definiciones se establece que cada uno de los elementos del Universo pertenecen o no a un determinado conjunto. Por lo tanto, cada conjunto puede definirse completamente por una función de pertenencia, que opera sobre los elementos del Universo, y que le asigna un valor de 1 si el elemento pertenece al conjunto, y de 0 si no pertenece. (Figura 1)

Tomando un conjunto C que está compuesto por los números pares definidos dentro del universo U, su función de pertenencia $u_c(x)$ sería de la siguiente forma:

$$u_c(0)=0,\ u_c(1)=0,\ u_c(2)=1,\ u_c(3)=0,\ u_c(4)=1,$$
$$u_c(5)=0,\ u_c(6)=1,\ u_c(7)=0,\ u_c(8)=1,\ u_c(9)=0$$

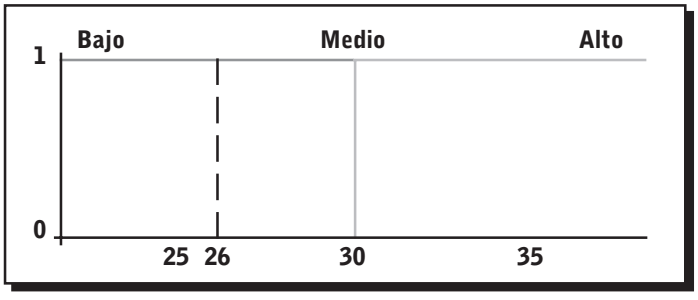

Figura 1: Ejemplo de Conjuntos Clásicos.

CONJUNTOS DIFUSOS (Fuzzy Sets)

Un Conjunto Difuso se define de forma similar, con una diferencia conceptual importante: un elemento puede pertenecer parcialmente a un conjunto. De esta forma, un conjunto difuso D definido sobre el mismo universo U puede ser el siguiente:

D={20%/1,50%/4,100%/7}

Esta definición significa que el elemento 1 pertenece en un 20% al conjunto D (y por tanto pertenece en un 80% al complemento de D), en tanto que el elemento 4 pertenece en un 50%, y el elemento 7 en un 100%.

En forma alternativa, se dice que la función de pertenencia $u_D(x)$ del conjunto D es la siguiente:

$u_D(0)=0.0$, $u_D(1)=0.2$, $u_D(2)=0.0$, $u_D(3)=0.0$, $u_D(4)=0.5$, $u_D(5)=0.0$, $u_D(6)=0.0$, $u_D(7)=1.0$, $u_D(8)=0.0$, $u_D(9)=0.0$

Algunas de las diferencias entre los Conjuntos Concretos y los Conjuntos Difusos son las siguientes:

- La función de pertenencia asociada a los Conjuntos Concretos sólo puede tener dos valores: 1 ó 0, mientras que en los conjuntos difusos puede tener cualquier valor entre 0 y 1.

- Un elemento puede pertenecer (parcialmente) a un conjunto difuso y simultáneamente pertenecer (parcialmente) al complemento de dicho conjunto. Lo anterior no es posible en los conjuntos concretos, ya que constituiría una violación al principio del tercero excluido.

- Las fronteras de un conjunto concreto son exactas, en tanto que las de un conjunto difuso son, precisamente, difusas, ya que existen elementos en las fronteras mismas, y estos elementos están a la vez dentro y fuera del conjunto.

Ejemplo:

Supóngase que se desea clasificar a los miembros de un equipo de fútbol según su estatura en tres conjuntos, Bajos, Medianos y Altos.

Como ejemplo podría plantearse que se es Bajo si se tiene una estatura inferior a 160 cm. que se es Mediano, si la estatura es superior o igual a 160 cm. e inferior a 180 cm., y se es alto si la estatura es superior o igual a 180 cm., con lo que se lograría una clasificación en Conjuntos Concretos.

Sin embargo, ¿qué tan grande es la diferencia que existe entre dos jugadores del equipo, uno con estatura de 179.9 cm. y otro de 180.0 cm?

Ese milímetro de diferencia quizás no represente en la práctica algo significativo, y sin embargo los dos jugadores han quedado rotulados con etiquetas distintas: uno es Mediano y el otro es Alto. Si se optase por efectuar la misma clasificación con conjuntos difusos estos cambios abruptos se evitarían, debido a que las fronteras entre los conjuntos permitirían cambios graduales en la clasificación.

Un jugador con 163 cm. de altura tendría un valor de pertenencia al conjunto denominado Bajo (0.8) y un valor de pertenencia al conjunto denominado Mediano (0.2). (Figura 2).

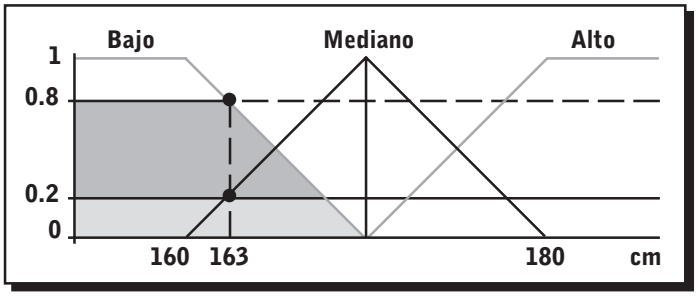

Figura 2: Ejemplo de Conjuntos Difusos

MINERÍA DE DATOS (Data Mining)

Definición de descubrimiento de conocimiento en bases de datos:

"*... es el proceso no trivial de identificar patrones en datos que sean válidos, novedosos, potencialmente útiles y, por último, comprensibles...* [6]"

Algunos autores definen la minería de datos como:

"*... un paso esencial en el proceso de descubrimiento de conocimiento en Bases de datos...* [7] "

"*... se refiere al acto de extraer patrones o modelos a partir de los datos...* [6]"

La minería de datos, consiste en la extracción de información oculta y predecible de grandes bases de datos, es una metodología que sirve para ayudar a las compañías e instituciones a concentrarse en la información más importante de sus bases de información.

Se ha convertido en una herramienta de toma de decisiones frente a las metodologías clásicas, es así como productos comerciales de Sistemas Manejadores de Bases de Datos (SMBD) de grandes compañías productoras de software ya implementan algoritmos de minería de datos y hasta permiten la implementación de propios.

Los algoritmos de minería de datos exploran las bases de datos en busca de patrones ocultos, encontrando información que un experto humano difícilmente encontraría, estableciendo relaciones y patrones de los cuales las empresas pueden obtener grandes beneficios.

Una idea general de lo que intenta la minería de datos es **describir** el contenido de las colecciones de información para **predecir** el comportamiento del sistema.

Algunas de las tareas que se pueden realizar aplicando algoritmos de minería de datos son:

- Regresión.
- Agrupamiento.
- Reglas Asociativas.
- Árboles de Decisión.
- Detección de Cambios.

En minería de datos se utilizan diversas técnicas para realizar tareas en grandes conjuntos de datos, este enfoque multidisciplinario combina áreas como la estadística, el aprendizaje de máquina, tecnologías difusas, redes neuronales, algoritmos genéticos y demás. [1]

Una representación frecuente de un proceso típico de descubrimiento de conocimiento en bases de datos, contempla los siguientes pasos [6]:

1. Desarrollar una comprensión del dominio de la aplicación.
2. Crear un conjunto de datos objetivo.
3. Limpieza y preprocesamiento de los datos.
4. Reducción y transformación de los datos.
5. Elegir la tarea de minería de datos.
6. Elegir los algoritmos de minería de datos.
7. Minería de datos.
8. Evaluar el resultado de la minería de datos.
9. Consolidar el conocimiento descubierto.

CLUSTERIZACIÓN: Minería de Datos Difusa

El propósito de la agrupación de datos *(clustering)*, es la de segmentar la información de acuerdo con unos criterios definidos de similitud, de cumplimiento de características o patrones, de esta manera se generan conjuntos denominados Cluster, estos por lo general son de tipo clásico, dentro de los objetivos de este trabajo está el de generar Cluster de tipo difuso, que interpreten de mejor manera el mundo real, además que el análisis de respuesta con la interpretación de la agrupación apunte a la elaboración de estrategias para el mejoramiento del sistema.

La tarea de segmentación de datos en grupos autodefinidos cuyos rangos y medias son hallados automáticamente por la aplicación, se basan en la dispersión difusa de los mismos datos utilizando un método de agrupación difusa, de especial interés para el presente trabajo, es el algoritmo de agrupación (Fuzzy C-Means) [2], existen diversas aplicaciones de agrupación difusa [9].

Este algoritmo asigna un conjunto de datos, caracterizados por sus

respectivos valores de atributos, a un número determinado de conjuntos. Como resultado cada dato tiene un grado de pertenencia a cada conjunto, representada por su centro de conjunto, básicamente el algoritmo se realiza aplicando los siguientes cuatro pasos:

1) Inicialización.
2) Cálculo de centros de conjunto.
3) Actualización de valores de pertenencia.
4) El criterio de detención.

En la aplicación desarrollada se ha realizado la segmentación de datos utilizando la llamada agrupación difusa (fuzzy clustering), y selección automática de atributos, para aumentar las tasas de respuesta.

Además del cambio de utilización del algoritmo, se ignora el criterio de detención y se opta por el manual, siendo el usuario final de la aplicación quien aplica el criterio de detección.

Paso 1: Inicialización.

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

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

Eq. (1)

Donde:

- c: es el número de conjuntos a encontrarse.
- J: es el número de datos a agrupar.
- μ_{ij}; i = 1, ..., c; j = 1, ...,J: es el grado de pertenencia del dato j al conjunto i:

Paso 2: Cálculo de Centros de Conjunto.

Dados los valores de pertenencia μ_{ij} los centros V_i de cada conjunto i están dados por:

$$V_i = \frac{\sum_{j=1}^{J} (\mu_{ij})^m X_j}{\sum_{j=1}^{J} (\mu_{ij})^m}, \forall i = 1,...,c$$

Eq. (2)

Donde:
- X_j; j = 1, ...,J: es el vector de atributos del dato j:
- m: se denomina difusor (fuzzifier) y determina el grado de difusión (fuzziness) para los conjuntos encontrados (1<m<?) para m "cercano a 1" se calcula una solución con conjuntos clásicos.

Paso 3: Actualización de valores de pertenencia.

Dados los centros de conjunto calculados en el paso 2, los valores de pertenencia m(i,j) son actualizados utilizando la siguiente fórmula:

$$\mu_{ij} = \frac{1}{\sum_{k-1}^{c} \left[\frac{d_{ij}}{d_{kj}} \right]^{\frac{2}{m-1}}},$$

$$\forall i = 1,...,c; \forall j = 1,...,J$$

Eq. (3)

Donde:
- d_{ij}: es la distancia entre el dato j y el centro del conjunto i V_i.
- En el cálculo de esta distancia se utilizan los centros del conjunto i V_i obtenidos en el paso 2.

Paso 4: El Criterio de Detención

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

|| A(t+1) - A(t) || < Umbral

Eq. (4)

Donde:
- A es la matriz de los valores de pertenencia en la iteración t.

En el algoritmo C-Means el umbral ha de ser determinado por el usuario, pero en la aplicación desarrollada es omitido para permitir el número máximo de iteraciones posibles logrando con esto un alto grado de solución difusa.

APLICACIÓN PRÁCTICA

La metodología planteada puede utilizarse para analizar la información en bases de datos de cualquier tipo de entidad (empresa, institución, universidad, etc.). Se ha tomado como sistema de muestra y pruebas el sistema de información del rendimiento académico de los estudiantes de la Universidad de los Llanos (Villavicencio, Meta, Colombia), en la cual se consigna la información personal de cada estudiante de las diferentes carreras, además de sus respectivas notas definitivas en cada una de las materias que componen el plan académico.

Para presentar en este artículo se ha decidido dividir el problema en subproblemas de menor tamaño para facilitar el entendimiento del mismo. La implementación de la aplicación en el sistema general de ejemplo es similar.

Se desea clasificar en cinco grupos (Excelente, Bueno, Medio, Malo, Deficiente) el rendimiento académico de los estudiantes de Ingeniería Agronómica de la Universidad de los Llanos que han cursado la materia "Fisiología Vegetal", la cual hace parte del sexto semestre del plan de estudios del programa.

El filtro SQL (Structured Query Language), para el ejemplo es: "SELECT nota FROM tagronomia WHERE codmateria = 10611 ORDER BY nota;"

corresponde a las notas de los estudiantes de Ingeniería Agrónomica en la materia "Fisiología Vegetal" con código 10651, hasta el primer periodo académico del año 2003:

Especificación de parámetros

- Número de Clusters: 5.
- Parámetro de difusidad: 2.
- Número de Datos: Es auto establecido cuando se cargan los datos pero puede ser modificado, para el ejemplo el número es de 1112 registros.
- Número de Iteraciones: 100.

Análisis de Resultados

A continuación se presenta un ejemplo detallado de los resultados obtenidos

- La Clusterización se realizó sobre 5 conjuntos (Conjunto 1,..., Conjunto 5)

  Ejemplo: Los 5 conjuntos pueden considerarse como rendimientos de tipo (Excelente, Bueno, Medio, Malo, Deficiente) según las notas obtenidas por los estudiantes y el centro de cada conjunto.

- Cada uno de los Conjuntos está centrado sobre un valor obtenido automáticamente por la aplicación. (Tabla 1)

	Conjunto 1	**Conjunto 2**	**Conjunto 3**	**Conjunto 4**	**Conjunto 5**
Centro	3.44	1.86	3.02	4.05	2.44

Tabla 1. Centros de conjuntos obtenidos

- De acuerdo con el centro del grupo se le asigna una etiqueta. (Tabla 2)

	Conjunto 1	**Conjunto 2**	**Conjunto 3**	**Conjunto 4**	**Conjunto 5**
Etiqueta	**Bueno**	**Deficiente**	**Medio**	**Excelente**	**Malo**

Tabla 2. Etiquetas de conjuntos

- Cada dato analizado tiene un valor de pertenencia a cada uno de los conjuntos obtenidos. (Tabla 3)
- Ejemplo: Estudiantes con las siguientes notas pertenecerían respectivamente a cada conjunto así:

	Conjunto 1	**Conjunto 2**	**Conjunto 3**	**Conjunto 4**	**Conjunto 5**
	Bueno	**Deficiente**	**Medio**	**Excelente**	**Malo**
	3.44	1.86	3.02	4.05	2.44
1.5	0.027	0.798	0.044	0.016	0.115
3.0	0.004	0.001	0.992	0.001	0.002
4.7	0.169	0.033	0.096	0.648	0.053

Tabla 3. Algunos Resultados

- El estudiante con nota 1.5 tiene mayor grado de pertenencia al conjunto denominado "Deficiente = 0.798" y el grado mayor de pertenencia con respecto a los demás conjuntos está en "Medio = 0.044", entonces se considera que tuvo un rendimiento "Deficiente" con una muy leve tendencia a "Medio" en Fisiología Vegetal.

- El estudiante con nota 3.0 tuvo un rendimiento absolutamente "Medio = 0.992" con respecto a los conjuntos establecidos.

- El estudiante con nota 4.7 tuvo mayor pertenencia al conjunto "Excelente = 0.648"seguido de "Bueno = 0.169", lo cual indica que el rendimiento académico de este estudiante es "Excelente" con leve tendencia a "Bueno".

- Los demás datos se analizan de manera similar observando detalladamente el valor de pertenencia del dato a cada uno de los conjuntos.

Análisis Gráfico (Figura 3)

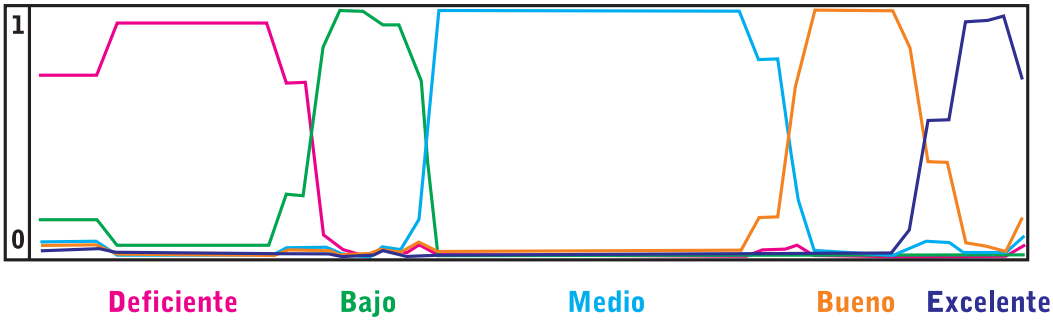

Figura 3. Representación gráfica de los conjuntos obtenidos.

OBSERVACIONES

La interpretación del gráfico obtenido puede ser de la siguiente manera:

- Como se observa en el gráfico, el conjunto más grande es el etiquetado como "medio" quiere decir que la tendencia de rendimiento de los estudiantes que han cursado Fisiología Vegetal es de tipo medio.
- Se observa que el conjunto "Deficiente" tiene un tamaño considerable lo que indica que el mal rendimiento de los estudiantes se hace presente, e indica que la mortalidad académica en el curso es bastante alta.
- El conjunto de menor tamaño es el denominado "Excelente" es decir la excelencia académica al cursar la materia es mínima.
- En cuanto a los conjuntos denominados "Bajo y Alto" se hacen presentes pero no con mayor trascendencia.

Los resultados arrojados por la aplicación se sometieron a un análisis riguroso para así determinar estrategias a seguir en el estudio del área.

Para este ejemplo la toma de decisiones para las estrategia de mejoramiento y seguimiento académico de los estudiantes es definida por el Consejo de Facultad y comité de cada programa, teniendo en cuenta las observaciones anteriormente nombradas.

Se identifican las áreas donde la mortalidad académica es alta, además de analizar los primeros semestres para determinar en qué áreas existe mayor mortalidad académica lo cual lleva a la deserción estudiantil al inicio de carrera.

Esto con el fin de ayudar al mejoramiento de la calidad de los programas ofrecidos por la Universidad de los Llanos.

CONCLUSIONES Y RECOMENDACIONES

La mezcla de áreas como minería de datos y lógica difusa permite obtener resultados más cercanos al pensamiento natural, es por ello que este trabajo es tan solo un primer paso en el estudio de un área muy grande por explorar, que intenta reiterar el trabajo realizado por el grupo de estudio CIULL (Computación Inteligente - Unillanos) del centro de investigaciones de la Facultad de Ciencias Básicas e Ingeniería de la Universidad de los Llanos (Villavicencio-Meta-Colombia).

Como posibles trabajos futuros generados a partir de este se tienen:

- Implementar nuevas tareas de Minería de Datos, como Reglas Asociativas, Árboles de decisión.
- Implementar nuevos algoritmos de Minería de Datos.
- Utilizar los conjuntos obtenidos con la aplicación como valores de entrada en sistemas basados en lógica difusa Tipo Mamdani o Tipo Takani-Sugeno.
- Centros y Rangos de Conjuntos Difusos determinados por el usuario.
- Implementar nuevos algoritmos de Clusterización.
- Generación de resultados en lenguaje natural.
- Implementar sistemas híbridos con Redes Neuronales y Algoritmos Genéticos.
- Desarrollar compatibilidad para la entrada a Sistemas Basados en lógica Difusa en MatLab® (ToolBox Lógica Difusa)
- Implementar soporte para cualquier Sistema Manejador de Bases de Datos.

Análisis Futuros (Tabla 4, 5, 6):

Objetivo	Una vez realizada la agrupación determinar en qué sectores y bajo qué factores se tiene un mejor rendimiento en productos y/o servicios de diferente clase y en aquellos de bajo rendimiento determinar las causas.
Entidades	• Facultad de Ciencias Agropecuarias, Universidad de los Llanos (En ejecución, BD Cultivos Invernadero): • Instituto de Acuicultura de los Llanos (IALL) (En trámite, BD Producción Piscícola). • Federación Colombiana de Ganaderos (FEDEGAN). (BD Producción ganadera).

Tabla 4. Entidades para análisis Futuros

Objetivo	Determinar bajo qué factores y características algunas enfermedades se propagan en el municipio de Villavicencio (Meta, Colombia)
Entidades	• (En ejecución) Facultad de Ciencias de la Salud, Universidad de los Llanos.

Tabla 5. Entidades para análisis Futuros

Objetivo	Establecer estrategias de mercadeo para aumentar ingresos.
Entidades	• Grandes almacenes de cadena en la ciudad de Villavicencio (Meta, Colombia).

Tabla 6. Entidades para análisis Futuros

REFERENCIAS

1. ADRIAANS, P. Y ZANTINGE, D. 1996: Data Mining. Addison-Wesley, Harlow.

2. BEZDEK, J.C. 1981: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, Nueva York.

3. DELGADO, ALBERTO. 1998. "Inteligencia Artificial y Mini robots", Editorial Ecoe Ediciones.

4. DUARTE, OSCAR G. 1997. "UNFUZZY - Software para el análisis, diseño, simulación e implemen-tación de Sistemas de Lógica Difusa". M.Sc. Tesis. Universidad Nacional de Colombia, Facultad de Ingeniería, Maestría en Automatización Industrial.

5. DUARTE, OSCAR G. "Sistemas de Lógica Difusa - Fundamentos", Revista Ingeniería e Investigación No.43, Revista de Facultad de Ingeniería Universidad Nacional de Colombia.

6. FAYYAD, U. M. 1996: "Data Mining and Knowledge Discovery: Making Sense out of Data." IEEE Expert, Intelligent Systems & Their Applications, Octubre 1996, 20-25.

7. HAN J. & KAMBER M., 2000: Data Mining: Concepts and Techniques. 519 pags. Editorial Morgan Kaufmann Publishers. New York.

8. MARTIN MCNEILL, ELLEN THRO. 1994. "Fuzzy Logic, a Practical Aproach", Editorial AP Profesional.

9. MEIER, W., WEBER, R. Y ZIMMERMANN, H.-J. 1994: "Fuzzy Data Analysis – Methods and Industrial Applications." *Fuzzy Sets and Systems* 61, 19-28.

10. PÉREZ H., GUSTAVO. "Sistemas de Lógica Difusa - Notas de Clase", Universidad Nacional de Colombia.

11. TIMOTHY, J. ROSS. 1997. "Fuzzy Logic With Engineering Applications", Editorial McGraw-Hill.