Artículo 1.

PORTADA

J. Sanabria Garzón
herramienta software para implementar minería de datos: clusterización utilizando lógica difusa
Orinoquia, vol. 8, núm. 1, 2004, pp. 15 - 23.
Universidad de Los Llanos
Colombia

Orinoquia,
ISSN (Versión impresa): 0121-3709
orinoquia@hotmail.com
Universidad de Los Llanos
Colombia

Técnica de Lógica difusa: Minería de datos: Clusterización

Técnica difusa de minería de datos: Alternativa para la caracterización e interpretación del perfil del emprendedor potencial.

RESUMEN

"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)

RESUMEN

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.

Tecnicas de lógica difusa utilizada para el agrupamiento de los datos algoritmo C-Mean

ABSTRACT

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.

INTRODUCCION

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

MÉTODOS

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

Entonces podemos defin
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 com-

tenece. (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:

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

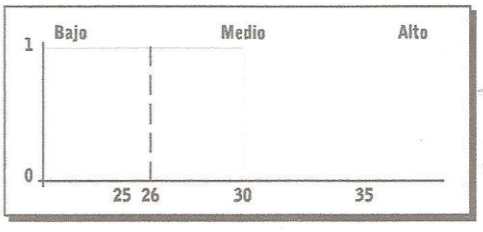

Figura 1: Ejemplo de Conjuntos Clásicos.

Clusterización o agrupación estática Minería de datos

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).

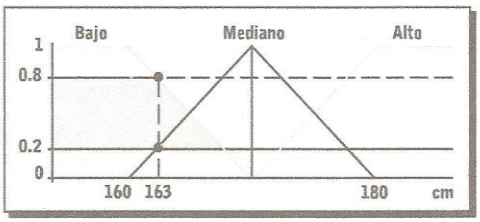

Figura 2: Ejemplo de Conjuntos Difusos

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.

Proceso típico de descubrimiento de datos

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.

objetivo

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:

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

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:

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

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:

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

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

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:

$$|| A(t+1) - A(t) || < \text{Umbral}$$

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.

RESULTADOS

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)

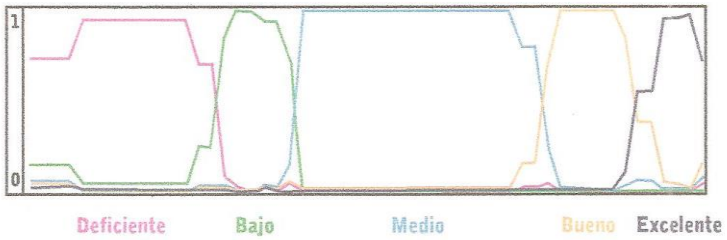

Figura 3. Representación gráfica de los conjuntos obtenidos.

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).

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

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.

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.

Artículo 2.

PORTADA

JIMENEZ RAMIREZ, CLAUDIA; ALVAREZ ZAPATA, HERNÁN

MINERÍA DE DATOS BASADA EN LÓGICA DIFUSA PARA LA INTERPRETACIÓN DE CONSULTAS VAGAS DEPENDIENTES DEL CONTEXTO LINGÜÍSTICO

Dyna, vol. 79, núm. 173, junio, 2012, pp. 75-84
Universidad Nacional de Colombia
Medellín, Colombia

INTRODUCCIÓN

MINERÍA DE DATOS BASADA EN LÓGICA DIFUSA PARA LA INTERPRETACIÓN DE CONSULTAS VAGAS DEPENDIENTES DEL CONTEXTO LINGÜÍSTICO

DATA MINING USING FUZZY LOGIC FOR THE TRANSLATION OF VAGUE QUERIES DEPENDING ON THE LINGUISTIC CONTEXT

CLAUDIA JIMENEZ RAMIREZ
Ph.D. Universidad Nacional de Colombia, Sede Medellín. csjimene@unal.edu.co
HERNÁN ALVAREZ ZAPATA
Ph.D. Universidad Nacional de Colombia, Sede Medellín. hdalvare@unal.edu.co

Recibido para revisar Febrero 3 de 2012, aceptado marzo 5 de 2012, versión final marzo 29 de 2012

RESUMEN: En este artículo se presenta un método propuesto para que un sistema flexible de consulta-respuesta a bases de datos pueda hallar, de manera autónoma y dinámicamente, la semántica de las condiciones vagas de las consultas, explorando los datos disponibles en la base de datos y usando lógica difusa. La máquina de inferencia del sistema, según el contexto lingüístico delimitado por cada consulta, elige un modelo de conjunto difuso entre los modelos predefinidos para diferentes patrones sintácticos con los que puede encajar el texto de una consulta vaga y considerando diferentes niveles de granularidad en la categorización de los objetos. Se estima el valor de los parámetros del conjunto difuso que representa una etiqueta lingüística, usando un método no supervisado y no paramétrico en el proceso de discriminación; evitando así, la intervención de expertos. Con esto se logra que los sistemas no sólo sean adaptables, sino confiables gracias a la validez de sus respuestas.

PALABRAS CLAVE: Sistemas flexible de consulta-respuesta, inferencia difusa, minería de datos

ABSTRACT: This paper presents a method in order to a flexible query-answering system can obtain, independently and dynamically, the semantics of imprecise words used as constraints in database queries by scanning the available data in the database and using fuzzy logic. The system's inference engine, according to the linguistic context defined by each query, chooses a fuzzy set model among the predefined theoretical models for different syntactic patterns with which can fit the text of a vague query and considering different levels of granularity in the fuzzy discrimination. It estimates the value of fuzzy set parameters by using a nonparametric method in the fuzzy discrimination, avoiding the intervention of experts. This achieves that the systems are not only adaptable, but reliable thanks to the validity of the responses.

KEYWORDS: flexible question answering systems, fuzzy inference, data mining

1. INTRODUCCIÓN

Puesto que en la interacción humano-máquina los términos vagos deben precisarse para obtener respuestas de esta última, en la presente propuesta se aborda el problema de la representación y manejo de la vaguedad dependiente del contexto lingüístico, como una estrategia clave para aproximar los lenguajes artificiales de los sistemas de consulta-respuesta al lenguaje natural.

Por la imprecisión de ciertas palabras, es corriente que la persona que consulta una base de datos no sepa si un objeto cualquiera se puede considerar "costoso", "reciente" o "pesado", entre muchas otras categorías no claramente diferenciables. Debido a esto, desde hace buen tiempo han surgido varias propuestas para flexibilizar el lenguaje estándar de consulta a bases de datos, el SQL (Structured Query Language) usando lógica difusa para admitir palabras vagas o imprecisas. Sin embargo, en esas propuestas, la forma y la ubicación espacial de los conjuntos difusos que representan las palabras vagas son definidas por expertos y proporcionadas, de antemano, a un sistema interactivo de consulta-respuesta. Esto hace que los modelos propuestos no puedan ser generales, ni objetivos y que además demanden mucho trabajo de actualización para preservar su validez; pues lo

que hoy se puede considerar "costoso", por ejemplo, posiblemente no lo sea en un futuro cercano.

Además de la variabilidad del significado de los términos vagos por el paso del tiempo, en las propuestas previas no se considera que muchos de ellos sean relativos o dependientes del contexto que delimita cada consulta. Un vehículo, por ejemplo, puede ser muy costoso en ciertos lugares, mientras que en otros no. Debido a esto, un sistema que no considere el contexto, no podrá capturar adecuadamente el significado de los términos vagos para ofrecer al usuario respuestas confiables.

En [1] se afirma que la principal dificultad para resolver en los sistemas de bases de datos basados en lógica difusa es la subjetividad en la representación de los conceptos vagos y su dependencia del contexto. Esta dificultad la motivación para abordar el problema de estimar, en forma dinámica y autónoma, la semántica de las condiciones vagas de una consulta en sistemas de bases de datos. Proponemos un método automático no supervisado que explora de los datos disponibles del contexto delimitado por la consulta, en la propia base de datos y realiza algunas mediciones para estimar los parámetros de los conjuntos difusos que representan las etiquetas lingüísticas detectadas. Se selecciona un modelo de conjunto difuso genérico aplicable en un caso, según el patrón sintáctico de la consulta y el número de categorías o clases que se deban considerar. Este modelo se particulariza o instancia gracias al proceso de minería de los datos realizada dinámica y automáticamente por el propio sistema. De esta forma, la máquina de inferencia emula a un experto calificado que puede descubrir nuevo conocimiento basándose en los datos disponibles. Por su autonomía y su capacidad de razonamiento puede considerarse un agente inteligente como se le denomina en la ingeniería del conocimiento.

El resto del presente artículo está estructurado de la siguiente manera. En la segunda sección, se hará una breve descripción de los conceptos fundamentales en los que se basa la propuesta: la minería de datos y una de sus tareas: la discriminación de objetos, basada en la lógica difusa. En la sección 3, se presenta el método propuesto para abordar el problema de capturar el significado de los términos vagos simples especificados en las condiciones de filtrado de las consultas, considerando diferentes niveles de granularidad. En esta sección, se presentan ejemplos de consultas vagas usando la extensión propuesta en el lenguaje estándar de bases de datos SQL, para que se aprecie cómo sintácticamente se preserva su proximidad con el lenguaje natural. Luego, en la sección 4, se presentan los modelos teóricos para hallar la semántica de combinaciones lineales de expresiones vagas, para finalizar con las conclusiones.

2. CONCEPTOS BÁSICOS

2.1. Minería de datos

La minería de los dat
vaguedad es el proces
de las palabras vagas o i
razonamiento conocido tam
"precisiation", y considerado un pre
para la representación y automatización del lenguaje natural puesto que permite la evolución de los Sistemas de Recuperación de Información a Sistemas Flexibles de Consulta-Respuesta [2].

La minería de datos es un proceso de exploración de los datos disponibles en las bases de datos, de forma automática o semiautomática, con el objetivo de encontrar patrones, tendencias o reglas que expliquen el comportamiento de cierto fenómeno en un determinado contexto. Encontrar patrones significa extraer información que permita establecer propiedades de o entre conjuntos de objetos [3].

Por su lado, la discriminación se considera el acto de separar o formar grupos, según algunos criterios o propiedades de los objetos con el objetivo de reconocer diferencias y similitudes entre los grupos y poder describirlas en forma gráfica o algebraica para lograr un mejor entendimiento de un determinado entorno [4]. Por esto, se puede decir que la discriminación y el reconocimiento de patrones son procesos equivalentes.

En el reconocimiento de patrones, algunas técnicas se consideran *no supervisadas* porque al sistema no se le ofrece una catalogación *a priori* de los patrones que se deben identificar para formar los grupos o clases pues no siempre se tiene un conjunto de ejemplares ya clasificados que permitan definir las clases, con base en sus propiedades. Una de las técnicas no supervisadas

MARCO TEORICO

más conocida es la técnica *c*-medias (k-means, en inglés), con todas sus variaciones, basadas en una medida de distancia como la euclidiana. Infortunadamente, este tipo de técnicas, además de ser costosas en términos computacionales, se pueden quedar en mínimos locales en el proceso iterativo de optimización de los centroides o no llegan a una solución [5].

Existen otras técnicas estadísticas no supervisadas que se basan en la densidad de los datos, en lugar de las distancias, y que han sido empleadas en diferentes disciplinas. Como ejemplo, en sicología han servido para catalogar a una persona adulta como "subnormal", "normal" o "superdotada", con base en el cociente intelectual y bajo el supuesto de normalidad en la distribución del coeficiente. Una partición basada en la distribución normal tiene la ventaja de ofrecer un mecanismo de discriminación muy sencillo que depende sólo de dos parámetros: la media y la desviación estándar. Sin embargo, no siempre es el modelo apropiado para representar una colección de datos, pues pueden existir sesgos o asimetrías en la distribución y tratar de representarla únicamente con dos parámetros no sólo resulta insuficiente, sino inexacto porque los estimadores comunes de los parámetros, no son medidas robustas a los valores extremos.

Para superar la dificultad de representación con el modelo probabilístico normal, se han propuesto los modelos llamados no paramétricos. Este término no quiere decir que tales modelos carecen de parámetros, sino que el número y la naturaleza de los mismos pueden ser flexibles y no preestablecidos de antemano [6]. Esto porque los datos observados son los que determinan la forma y ubicación del modelo de distribución. Razón por la cual también se les denomina modelos libres o independientes de la distribución de los datos (distribution free models, en inglés). Dentro de esta categoría se encuentran los modelos basados en percentiles y los histogramas.

Formalmente, un percentil P_q es un punto del dominio de una variable, bajo el cual se encuentra un porcentaje q de los valores de una distribución de datos. Particularmente, se ha comprobado que una forma muy efectiva para la descripción de la distribución de los datos es la estadística de resumen de los cinco números compuesta por el valor mínimo P_0, el valor máximo P_{100} y los tres cuartiles de la distribución. A partir de esta estadística se suele construir el diagrama de caja y bigotes (box and whisker plot o abreviadamente *boxplot*, en inglés), uno de los modelos gráficos más informativos [8]. En la Figura 1, se puede observar que este diagrama parte los datos en tres clases: la primera clase representada con el 25% de los datos con valores más bajos, la segunda clase con el 50% de los valores más comunes (la caja) y la tercera clase con los valores más altos. En esta figura, se compara con el histograma de frecuencias para que se aprecie la buena representación de una colección de datos, pero en una forma más condensada, sin necesidad de tantos parámetros.

densidad condensado.

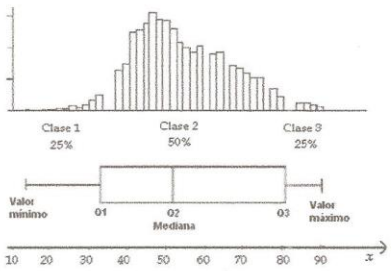

Figura 1. Diagrama de Caja y Bigotes

2.2. Discriminación y clasificación basada en lógica difusa

La discriminación basada en lógica difusa, como las demás técnicas basadas en esta lógica, se han propuesto para flexibilizar las técnicas de minería de datos tradicionales. En particular, las técnicas de agrupamiento (clustering, en inglés), al basarse en la lógica de Boole (cierto, falso), generan particiones matemáticas donde cada objeto pertenece a uno y sólo un grupo [9]. En cambio, en un modelo basado en lógica difusa, cada objeto puede pertenecer a varias clases rotuladas con una etiqueta lingüística, usando para ello, funciones de pertenencia. En la lógica difusa, cada clase o categoría E_i se representa mediante un conjunto difuso de pares ordenados:

$$E_i = \{(x, \mu_{E_i}(x)) \mid x \in U\} \quad (1)$$

En (1), U es el universo del discurso o dominio de la variable x y la medida $\mu_{E_i}(x)$ es el grado de pertenencia de x al conjunto con etiqueta E_i. Este

conjunto usualmente se define por medio de funciones predefinidas, parecidas a las funciones densidad probabilísticas, llamadas la trapezoidal, la gamma o la campana generalizada, entre muchas otras [10]. Los modelos de las funciones de pertenencia suelen ser elegidas con la ayuda de expertos. Sin embargo, independientemente de los modelos de los conjuntos difusos elegidos, la discriminación obtenida, debe considerarse una partición difusa que cumple las propiedades siguientes:

1. $E_i \neq \phi \; \forall i \in \mathrm{I}$

2. $\bigcup_{i \in I} E_i = U$

La primera propiedad especifica que no se pueden generar o definir conjuntos difusos vacíos en un marco de cognición y la segunda, demanda la cobertura total del dominio.

La computación granular, es llamada en lógica difusa, al paradigma de la representación y el manejo del concepto "gránulo de información", definido como aquel que surge de la derivación de conocimiento a partir de los datos [10]. Significa que un gránulo emerge de los datos como consecuencia de su resumen o condensación y el aspecto clave es la interpretación de los gránulos o grupos que se logra cuando se les puede fijar una etiqueta lingüística. Puesto que esto se quiere garantizar, una estrategia para asegurar que los gránulos o los conjuntos difusos sean interpretables, consiste en la definición de una sólida estructura lógica de razonamiento para realizar la discriminación difusa, utilizando una serie de restricciones que deben cumplirse.

Por limitaciones de espacio aquí no se presentan todas las restricciones que se consideraron en la presente investigación, pero es importante mencionar la convexidad de los conjuntos difusos para generar gránulos interpretables (que no siempre se obtienen con los métodos no supervisados basados en la distancia) y la restricción de complementariedad que demanda que la suma de los grados de pertenencia de cualquier elemento, a todas las clases, sea uno. Quiere decir que el grado de pertenencia total de cualquier elemento está repartido entre las clases y se considera importante pues garantiza que el grado de pertenencia de un elemento al conjunto universal sea uno [11]. Otra restricción considerada en este trabajo de investigación es que cualquier elemento ubicado en el área de solapamiento entre dos conjuntos tenga grados de pertenencia diferentes a los dos conjuntos; exceptuando el punto de cruce, cuyo grado de pertenencia a ambos conjuntos debe ser 0.5. Con ello, se busca la máxima especificidad en la discriminación de los objetos.

3. MINERÍA DE DATOS PARA LA CONCRECIÓN DE LA CONDICIONES VAGAS SIMPLES

Como se dijo ant... ...uestas previas para interpretar ... ...s de datos, los con... ...iciones vagas ... ...os. Sin embargo, ...a falta de generalidad de ... ...s y por su falta de adaptación a los distintos contextos lingüísticos que puedan delimitarse en las consultas. Por eso, con el fin de superar dichas limitaciones, se propone que la máquina, en forma automática y dinámica, realice un proceso de ajuste de modelos de los conjuntos difusos, a los datos existentes en la base de datos.

MÉTODO

En primer lugar, deberá realizar una exploración de la base de datos con el objeto de determinar el contexto, basándose en las condiciones concretas especificadas en la consulta. Luego, en la vista materializada obtenida, se realiza la minería de datos para la derivación de los modelos y las reglas de inferencia válidas en ese contexto. Todo ello considerando diferentes niveles de granularidad para ofrecer mayor flexibilidad en la categorización de los objetos.

Se considera que un adjetivo calificativo como "alto" es simple porque su significado depende de un sólo atributo. Para el caso de la representación de las etiquetas lingüísticas correspondientes a estos adjetivos simples, nos basamos en una partición matemática convencional a la cual se le realizó, luego, un proceso de difuminado, acorde con la lógica difusa.

En el proceso de discriminación definido, se supone la forma de la función de pertenencia de cada uno de los conjuntos que representan una etiqueta, pero se desconocen los valores de los parámetros. Estas formas, para definir los conjuntos difusos, varían según el número de clases o categorías que se deban considerar y la posición de la categoría difusa en cuestión.

Las formas son lineales, buscando simplicidad en los modelos: se eligió la función trapezoidal con parámetros (*a, b, c, d*) para definir un conjunto difuso, cuando la clase no sea una de los extremos, en el contexto considerado. Cuando sea una de éstas, se han elegido las formas semi-trapezoidales (conocidas también como *hombro izquierdo* y *hombro derecho*). Los parámetros de la función trapezoidal *a* y *d*, definen el soporte de la función de pertenencia (la base mayor del trapecio), y los parámetros *b* y *c* determinan núcleo de la misma.

Luego de seleccionar la forma de conjunto difuso apropiada para una etiqueta, se realiza la estimación de los parámetros del conjunto difuso que la representa. Aquí se propone una técnica basada en la densidad de los datos, usando estadísticas de posición relativas para lograr la adaptabilidad de la técnica a cada contexto e independencia sobre la distribución de los datos.

En el caso de una discriminación basada en dos categorías, se realizó una partición matemática usando la mediana para que cada clase contuviera el 50% de los datos. Se eligió esta medida en lugar de la media aritmética, por ser una medida robusta de la tendencia central [8]. Hecha esta partición convencional, se realizó un proceso de difuminado, basándonos en la estadística de resumen de los cinco números para definir su área de solapamiento como aquella con el 25% de los datos centrales, delimitada por los valores máximos de la primera clase ($x \geq P_{75}$) y los valores más pequeños de la segunda ($x \leq P_{25}$), como se ve en la Figura 2.

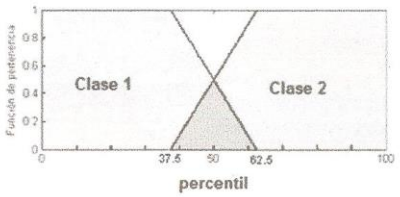

Figura 2. Partición difusa en dos clases

En el caso de una discriminación en tres clases, para determinar cuáles deberían ser los estimadores de los parámetros de las funciones que determinan los conjuntos difusos, se utilizó nuevamente la estadística de resumen de los cinco números sobre la clase intermedia. Por lo tanto, se consideró como núcleo de la función de pertenencia de esta clase, al 50% de sus datos centrales delimitados por sus dos cuartiles (Q_{N1} y Q_{N3}) como se muestra en la Figura 3. Dado que la clase intermedia contiene el 50% de los datos originales, los nuevos cuartiles cubren el 25% de los datos centrales de todos los datos. A partir de estos valores, se define la zona de solapamiento con las clases de los extremos.

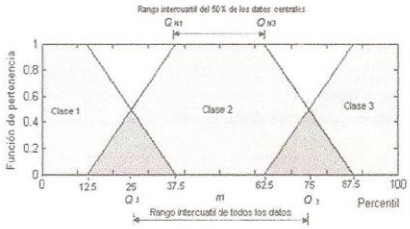

Figura 3. Partición difusa, en tres categorías

Además de discriminar considerando las anteriores categorías difusas, se consideraron otros niveles de granularidad más finos para mayor flexibilidad en la discriminación. Nuestra propuesta incluye hasta seis clases o categorías, como se muestra en la Tabla 1.

Tabla 1. Modelos teóricos de etiquetas vagas

clases (k)	Modelos propuestos (j = posición de la clase dentro del marco)
2	Si $j = 1$ → Hombro izquierdo(P_0, $P_{37.5}$, $P_{62.5}$) Si $j = 2$ → Hombro derecho ($P_{37.5}$, $P_{62.5}$, P_{100})
3	Si $j = 1$ → Hombro izquierdo (P_0, $P_{12.5}$, $P_{37.5}$) Si $j = 2$ → Trapezoidal ($P_{12.5}$, $P_{37.5}$, $P_{62.5}$, $P_{87.5}$) Si $j = 3$ → Hombro derecho ($P_{62.5}$, $P_{87.5}$, P_{100})
4	Si $j = 1$ → Hombro izquierdo (P_0, $P_{15.6}$, $P_{28.1}$) Si $j = 2$ → Trapezoidal ($P_{15.6}$, $P_{28.1}$, $P_{43.8}$, $P_{56.3}$) Si $j = 3$ → Trapezoidal ($P_{43.8}$, $P_{56.3}$, $P_{71.9}$, $P_{84.4}$) Si $j = 4$ → Hombro derecho ($P_{71.9}$, $P_{84.4}$, P_{100})
5	Si $j = 1$ → Hombro izquierdo (P_0, P_5, P_{25}) Si $j = 2$ → Trapezoidal (P_5, P_{25}, P_{50}, P_{65}) Si $j = 3$ → Trapezoidal (P_{50}, P_{65}, P_{80}, P_{95}) Si $j = 4$ → Trapezoidal (P_{65}, P_{80}, P_{95}, P_{100}) Si $j = 5$ → Hombro derecho (P_{95}, P_{95}, P_{100})
6	Si $j = 1$ → Hombro izquierdo(P_0, P_{10}, P_{18}) Si $j = 2$ → Trapezoidal(P_{10}, P_{18}, P_{26}, P_{34}) Si $j = 3$ → Trapezoidal (P_{26}, P_{35}, P_{46}, P_{54}) Si $j = 4$ → Trapezoidal(P_{46}, P_{54}, P_{64}, P_{72}) Si $j = 5$ → Trapezoidal(P_{64}, P_{72}, P_{82}, P_{90}) Si $j = 6$ → Hombro derecho(P_{82}, P_{90}, P_{100})

3.1. Sintaxis de las condiciones vagas simples

La sintaxis de una consulta con condiciones vagas simples dependientes del contexto, demanda una extensión del lenguaje de consulta. Para ello, la presente propuesta, se basa en el lenguaje SQLf3 [12], entre otras extensiones, debido a su proximidad con el lenguaje natural. Por esto, la forma para la especificación de una condición vaga simple, en una consulta, es:

SELECT proyección
FROM relaciones
WHERE expresión IS E [j/k]
[WITH CALIBRATION {n| λ |n, λ }]

En este tipo de sentencia, *proyección* es la lista de propiedades que el usuario quiere visualizar de una relación restringida a aquellos ejemplares que puedan ser calificados con la etiqueta lingüística *E* cuya posición en el marco es la *j* considerando *k* categorías difusas. El valor *n* es llamado el calibrador cuantitativo, que permite restringir a un número máximo "*n*" de las mejores respuestas y el umbral o calibrador cualitativo, permite visualizar sólo las tuplas cuyo grado de satisfacción a las condiciones especificadas sea mayor a un nivel mínimo de tolerancia λ, en el encajamiento [13].

Los valores *j* y *k* sólo se especificarían si se usa una etiqueta lingüística genérica como "alto" o "bajo" que no haya sido guardada como parte del conjunto de términos lingüísticos asociados a una variable cuantitativa en los metadatos. Como ejemplo, en la sentencia siguiente se pide el nombre y la dirección de los hoteles en Madrid cuyo valor noche pueda considerarse "medio", considerando tres categorías:

SELECT nombre, dir FROM hoteles
WHERE precio IS "medio" 2/3
AND ciudad = "Madrid"

Para visualizar el método de concreción propuesto, se realizaron pruebas experimentales usando una base de datos de referencia sobre 398 autos, utilizada ampliamente en Minería de Datos y Aprendizaje de Máquinas [14]. En una de las pruebas, se hallaron los modelos de los conjuntos difusos de la potencia de los autos, medida en caballos de fuerza, considerando diferentes contextos y tres categorías en la discriminación. En Tabla 2 se puede observar cómo cambian los modelos de los autos cuando se restringen a los autos de ciertas marcas o a aquellos que se demoran 16 o más segundos para pasar de cero a 60 millas/hora. Esta variabilidad en los parámetros de los conjuntos difusos que representan cada etiqueta ratifica que el significado de los términos vagos depende, generalmente, del contexto considerado.

Tabla 2. Parámetros de las funciones de pertenencia en distintos contextos

Potencia del vehículo (HP)	"Baja"	"Alta"
Contexto	Parámetros	Parámetros
Todos (n=398)	(46 67 87)	(105 150 230)
Marca Ford (n=48)	(65 78 88)	(130 153 215)
Marca Chevrolet (n=46)	(52 72 100)	(125 150 220)
Aceleración ≥ 16 (n=166)	(46 61 72)	(88 105 193)

Con el ánimo de mostrar cómo la máquina puede ser adaptable, no sólo a los cambios en el tiempo o en el espacio, en la Figura 4 se presentan dos marcos de cognición diferentes, a los cuales se podría ajustar autónomamente. Allí se puede ver que cuando se consideren dos clases, un auto se es de potencia "baja", si se tiene entre 50 y 100 caballos de fuerza (hp). En cambio, esta categoría deja de cubrir los autos con una potencia entre 80 y 100 hp, cuando se consideran tres clases.

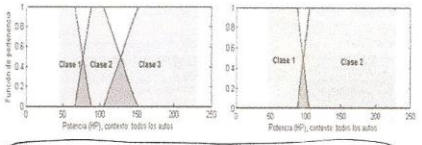

Figura 4. Marcos con distinto número de clases

3.2 Modificadores lingüísticos en las condiciones vagas simples

Un modificador lingüístico cambia los valores de verdad de una sentencia. Por ejemplo, un término como "joven" origina otros como "no muy joven" o "muy joven", gracias a la negación o al uso de adverbios de cantidad.

Convencionalmente, la compatibilidad de un objeto con una etiqueta lingüística modificada por un adverbio de cantidad o por la negación, se infiere, de manera deductiva, de la función de pertenencia definida para la etiqueta lingüística que le da origen.

La negación representa el complemento de una etiqueta lingüística. A pesar de que existen varias propuestas para hallar el valor de pertenencia al complemento de un conjunto difuso como la de Yaguer o la de Sugeno, se ve conveniente emplear la definición clásica. Dicha definición es una negación fuerte que cumple con la ley de involución definida para el álgebra de Boole [15]. Por esto, si se desea encontrar el grado de pertenencia de una persona al grupo de los "no jóvenes", por ejemplo, se obtendría la respuesta, así:

$$\mu_{\text{"no joven"}}(x) = \mu_{\text{"joven"}}{}^{C}(x) = 1 - \mu_{\text{"joven"}}(x)\ \forall x \in U$$

Por otro lado, un adverbio de cantidad como "muy" o "algo", se considera un operador que acentúa o relaja el significado de un adjetivo calificativo. Si la etiqueta E_i se caracteriza por una función de pertenencia $\mu Ei(x)$, entonces la función $\mu_{muy}\,Ei(x) = \mu^{k}Ei(x)$ se interpreta como una versión modificada del valor lingüístico original. Usando esta función exponencial se obtienen corrientemente las representaciones para los adverbios de cantidad vagos que dependiendo del exponente k se denomina función de dilatación o de concentración [15].

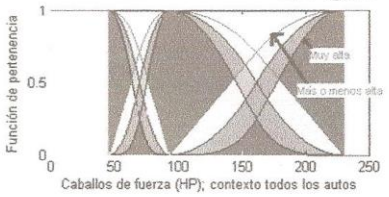

Figura 5. Modificadores de un conjunto difuso

En la Figura 5 se muestra un ejemplo de aplicación de la función de concentración $E_i{}^2$ y la función de dilatación $E_i{}^{1/2}$, para la potencia de los autos de la base de datos de referencia. A simple vista, se puede observar que las densidades de los subconjuntos enfatizados con el adverbio "muy" superan el 75% de la densidad del conjunto del cual fueron derivados, mostrando un efecto pobre del operador sobre las clases originales.

De forma análoga, el operador de dilatación no genera cambios significativos sobre el conjunto difuso del cual se origina. Los cambios serían aún más pequeños para las formas trapezoidales, pues el núcleo permanece inalterado [7].

Como lo señalan De Cock y Kerre [16], las funciones exponenciales que se han propuesto para la definición de los modificadores lingüísticos son sólo herramientas técnicas que conservan la propiedad de inclusión entre los subconjuntos difusos obtenidos, con el conjunto original etiquetado E, pero que carecen de significado como propiedad inherente. Por esto, hemos optado por otra estrategia que consiste en volver a realizar un proceso de discriminación sobre el conjunto que representa la etiqueta de interés.

Si la clase que necesita modificarse con el adverbio "muy" corresponde a la clase de los valores más pequeños y dado que el límite superior es el percentil $P_{37.5}$, entonces el soporte del subconjunto difuso "muy bajo(a)" o "muy pequeño(a)" debe ser menor o igual al percentil $P_{37.5}$ de ese conjunto, que equivale al percentil P_{14} de la distribución de todos los datos. Adicionalmente, el núcleo de la nueva función de pertenencia contiene al 12.5% de los datos menores del 37.5% de la clase con etiqueta "baja" o "pequeña". Por esto, el núcleo de la clase acentuada equivale al 4.7% de los valores más pequeños en todos los datos. Un razonamiento similar se aplica a la clase intermedia y a la clase de los valores "mayores" o "altos". En la Figura 6 se muestra la partición difusa propuesta para representar los conjuntos acentuados con el adverbio "muy".

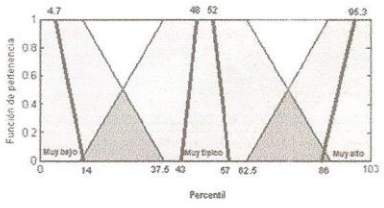

Figura 6. Clases acentuadas con el adverbio "muy"

Por su lado, el adverbio "extremadamente E", se puede interpretar como la acentuación del término "muy E". Entonces, se divide nuevamente el conjunto ya

acentuado, en tres subconjuntos difusos, y se realiza un procedimiento similar al recientemente descrito.

De acuerdo con lo anterior, los conjuntos difusos para encontrar la semántica de "muy" y "extremadamente", se pueden derivar de los datos del contexto (ver Tabla 3).

Tabla 3. Modelos teóricos para clases acentuadas

Etiqueta modificada con acentuador	Modelo teórico de la clase difusa
"Muy bajo(a)"	*Hombro izquierdo* (P_0 $P_{4.7}$ P_{14})
"Muy típico(a)"	*Trapezoidal* (P_{43} P_{48} P_{52} P_{57})
"Muy alto(a)"	*Hombro derecho* (P_{86} $P_{95.3}$ P_{100})
"Extremadamente bajo(a)"	*Hombro izquierdo* (P_0 P_1 P_2)
"Extremadamente común"	*Triangular* (P_{49} P_{50} P_{51})
"Extremadamente alto"	*Hombro derecho* (P_{98} P_{99} P_{100})

El adverbio "más o menos E", considerado sinónimo de "algo E", indica una interpretación más relajada del concepto vago que modifica. Es por esto que para definir el soporte del conjunto modificado se opta por incluir los elementos que superen a la mediana de la clase adyacente, cuando ésta sea menor o incluir los elementos que sean menores que la mediana de la clase adyacente cuando sea mayor a la considerada. Además, el núcleo de la función se amplía hasta cubrir los elementos que pertenezcan a la intersección con la clase vecina. Por esto, el conjunto que representa la relajación de un conjunto debido a la aplicación del modificador vago "más o menos" será hombro izquierdo, si se trata de la primera clase en el marco de cognición, una función trapezoidal si es intermedia y será hombro derecho si se trata de la clase con los valores más altos.

4. INTERPRETACIÓN DE COMPOSICIONES VAGAS

Un término vago complejo se deduce de una composición formada con el uso de las conectivas lógicas de la disyunción y la conjunción. El sistema de inferencia propuesto, en la interpretación de estos términos, usa el modelo FITA (acrónimo de *First Infer Then Aggregate*), que consiste en primero inferir los grados de pertenencia marginales o individuales a cada etiqueta lingüística especificada y luego agregarlos [18].

En lógica difusa, las operaciones de unión e intersección se determinan mediante las funciones de pertenencia.

$$\mu_{A\cup B}(x) = \oplus(\mu_A(x)\ \mu_B(x)\ \forall x \in U \quad (2)$$
$$\mu_{A\cap B}(x) = \otimes(\mu_A(x)\ \mu_B(x)\ \forall x \in U \quad (3)$$

En (2) y (3), los símbolos $\oplus$ y $\otimes$ representan los operadores de la disyunción y la conjunción, respectivamente. Se han propuesto varias alternativas para representar la semántica de los operadores de la conjunción y la disyunción que cumplen con las restricciones para las s-normas y t-conormas [18]. Los operadores más comúnmente usados para representar las conectivas en la lógica difusa son el valor mínimo y el máximo de los grados de pertenencia, respectivamente, pues son duales con respecto a la negación fuerte. Esto quiere decir que, en su álgebra son válidas las leyes de De Morgan por eso un operador de estos puede derivarse o deducirse del otro. Sin embargo, el uso del máximo para representar el grado de pertenencia a la disyunción de conjuntos difusos, hace que se incumpla la ley del medio excluido, dado que $\max(\mu_A(x)\ \neg\mu_A(x) < 1$ cuando el objeto x se encuentre en un área de solapamiento entre dos conjuntos difusos. Esto significa que se podría inferir que el grado de pertenencia de ciertos valores x al conjunto universal no sea 1, como se esperaría por la definición de este conjunto. Esto ocurre por no generar un conjunto convexo, como se aprecia en la Figura 7.

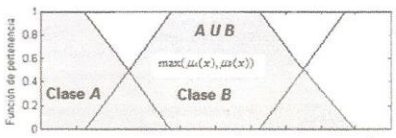

Figura 7. Conjunto difuso $A \cup B$ convencional

Por lo anterior, el operador elegido para representar la disyunción, en un marco de cognición, es la suma convencional. Este operador es una s-conorma, bajo la restricción de Ruspini que se mencionó antes. Adicionalmente, la suma es una función continua que no produce grandes cambios en el conjunto derivado, cuando los cambios son pequeños en alguno de los dos conjuntos que actúan como operandos, tal como lo exige la teoría difusa estándar [19]. Por lo tanto, en esta propuesta, el conjunto derivado de la disyunción, queda determinado por la ecuación siguiente.

$\mu_{A \cup B}(x) = \mu_A(x) + \mu_B(x) \forall x \in U \wedge A, B \in Marco$ (3) De acuerdo con (3), la representación gráfica de la unión de conjuntos es la que se presenta en la Figura 9, donde se ve claramente que el conjunto resultante es convexo, condición fundamental para que un conjunto difuso sea interpretable.

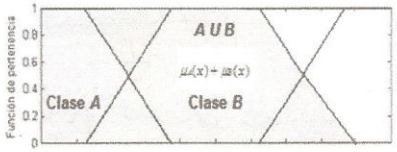

Figura 8. Unión de conjuntos usando la suma como

Infortuna... perador para rep... o éstos sean de distin... porque no se cumple la ley de la clausura. Debido a esto, en algunos casos, se podría concluir que el grado pertenencia global al conjunto difuso derivado por la conjunción de varias características sea mayor que 1. Por esto, han surgido propuestas distintas, que además consideran un peso diferente para cada etiqueta en la interpretación de las palabras vagas complejas. El operador LOWA (Linguistic Ordered Weighted Averaging) es un operador de este estilo que se basa en la operación OWA. Una operación OWA de la forma $\sum_{i=1}^{p} \beta_i \mu(x_i)$ es una función con pesos $\beta_i \in [0,1] \wedge \sum_{i=1}^{p} \beta_i = 1$, que cumple ciertas propiedades como las condiciones de borde y la monotonía. El atributo de mayor peso, en la interpretación de la etiqueta vaga compleja, va primero y los demás van en orden decreciente. Por esto, el grado de pertenencia global de un objeto a la unión de p etiquetas vagas simples, se define mediante la ecuación siguiente.

$$\mu_{E1 \cup E2 \ldots \cup Ep}(t_i) = \sum_{i=1}^{p} \frac{\mu_{Ei}(t_i)}{p} \quad \forall t_i \in R \quad (4)$$

Por otro lado, para la representación del operador de la conjunción, se considera al mínimo de los grados de pertenencia de los operandos como la norma triangular más apropiada, puesto que usando este operador se cubre toda la zona de solapamiento entre dos conjuntos. El producto de los grado de pertenencia, que es otro operador propuesto, no la cubre en su totalidad, como se muestra en la Figura 11.

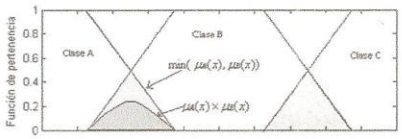

Figura 9. Operadores para representar la conjunción

El valor mínimo y la suma de los grados de pertenencia para representar las operaciones de la conjunción y de la disyunción, respectivamente, no cumplen la relación de dualidad con respecto al complemento. Sin embargo, la inexistencia de una relación de dualidad no es una limitación grave en la interpretación de una expresión vaga formada con conectivas lógicas en las condiciones de filtrado de una consulta.

CONCLUSIONES

Con la posibilida... considerando su ... propone, se pueden co... de consulta-respuesta más flexibles, sino más confiables y amigables con el usuario final.

La técnica no supervisada propuesta para la minería de los datos, permite obtener los modelos de los conjuntos difusos, dinámicamente, sin la necesidad de expertos e independientemente de la distribución de los datos contextuales. Esto muestra la generalidad de la técnica.

Por otro lado, con la posibilidad que se le otorga al sistema de inferencia para discriminar en un número variable de clases o conjuntos difusos, se aprecia otra flexibilidad de nuestra propuesta, no sólo por la posibilidad de adaptarse a diferentes contextos, sino por admitir diferentes niveles de granularidad en la categorización difusa.

REFERENCIAS BIBILOGRÁFICAS

[1] Galindo, J., Introduction and Trends in Fuzzy Logic and Fuzzy Databases. En Handbook of Research on Fuzzy

Information Processing in Databases. José Galindo. Idea Group Inc (IGI). 2008

[2] Zadeh, L., Chapter 9. From Search Engines to Question Answering Systems– The Problems of World Knowledge, Relevance, Deduction and Precisiation. *Capturing Intelligence*, Vol 1, pp. 163-210. 2006.

[3] Carrasco, J., Reconocimiento de patrones. Instituto Nacional de Astrofísica Óptica y Electrónica. 2010. Disponible en: http://ccc.inaoep.mx/~ariel/recpat.pdf. [Citado en marzo de 2012].

[4] Soto, C. y Jiménez, C., Aprendizaje Supervisado para la Discriminación y Clasificación Difusa. *Revista Dyna*. Vol 78 nro 169. 2011.

[5] Dyer, C., Machine Learning. Lecture Notes. Universidad de Wisconsin. Capítulos 18.1-18.3. Disponible en http://www.cs.wisc.edu/~dyer/cs540/notes/learning.html. [Citado en marzo de 2012].

[6] Johnson, R. and Wichern, D., Applied Multivariate Statistical Analysis. Prentice Hall. 6ta ed. EEUU. 2007.

[7] Jiménez, C., Razonamiento Aproximado y Adaptable en el Procesamiento de Consultas Vagas. Tesis doctoral, Universidad Nacional de Colombia, Medellín. 2008.

[8] Borgelt, C., Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Intelligent and Soft Computing, Vol 77, 611-618. 2010

[9] Rokach, L., Using Fuzzy Logic in Data Mining. *Data Mining and Knowledge Discovery Handbook*. Springerlink. 2012. http://www.springerlink.com. [Citado en marzo de 2012].

[10] Pedrycs, W., Granular Computing -The Emerging Paradigm. *Journal of Uncertain Systems*, pág.38-61. 2007. Disponible en: www.jus.org.uk. [Citado en marzo de 2012]

[11] Mencar, C., (2004). Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues. Tesis doctoral. Universidad de Bari. Italia. Disponible en: www.di.uniba.it/~mencar/download/research/tesi_mencar.pdf. [citado en marzo de 2012].

[12] Gonçalves, M. and Tineo, L., "A New Step towards Flexible XQuery". *Avances en Sistemas e Informática*. Vol 4(3). pp. 27-34. 2007

[13] Gonçalves, M., Rodriguez, C. y Tineo, L., Incorporando Consultas Difusas en el Desarrollo de software. *Avances en Sistemas e Informática*. Vol 6(3). pp. 87-101. 2009

[14] UCI. Machine Learning Repository. University of California, School of Information and Computer Science. Disponible en: http://archive.ics.uci.edu/ml/datasets/Iris. [citado en enero de 2012]: enero de 2012.

[15] Fodor, J., "Left-continuous t-norms in fuzzy logic: An overview". Acta Polytechnica Hungarica 1(2), ISSN 1785-8860. 2004 Disponible en: http://www.bmf.hu/journal/Fodor_2.pdf. [citado en enero de 2012].

[16] De Cock, M. and Kerre, E., Fuzzy modifiers based on fuzzy relations. Information Sciences. Lecture Notes in Artificial Intelligence 3214, pp. 779-785. 2004

[17] Cintula, P., Esteva, F., Gispert, J., Godo, L. and Noguera, C., Distinguished algebraic semantics for t-norm based fuzzy logics: methods and algebraic equivalencies, *Annals of Pure and Applied Logic* 160, pp. 53–81. 2009

[18] Cox, E., The fuzzy systems handbook: a practitioner's guide to building, using and maintaining fuzzy systems. Academic Press. Estados Unidos 1994

[19] Pradera, A., Trillas, E., Guadarrama, E. and Renedo., On Construction Imprecise Fuzzy Set Theories. *European Centre for Soft Computing*. 2006.