Una Metodología de KDD basada en Sistemas Difusos Genéticos Multiobjetivo para Modelado Causal en Marketing*

Jorge Casillas, Óscar Delgado
Dpto. Ciencias de la Computación e I.A.
Univ. de Granada
18071 Granada
casillas@decsai.ugr.es

Francisco J. Martínez-López
Dpto. Comercialización e Invest. Mercados
Univ. de Granada
18071 Granada
fjmlopez@ugr.es

Resumen

El trabajo presenta un problema novedoso basado en modelado causal en marketing y su resolución mediante descubrimiento de conocimiento, tal como se muestra con una aplicación real. Las características del problema (con datos inciertos y conocimiento experto disponible) y la optimización multiobjetivo propuesta hace a los sistemas difusos genéticos una buena herramienta para abordarlo.

1. Introducción

El trabajo presenta una metodología de *extracción de conocimiento en bases de datos* (KDD) a partir de modelos causales utilizados tradicionalmente en marketing. Se propone un método de *inducción descriptiva* [4] para descubrir reglas individuales que determinen patrones de especial interés en los datos. Para ello se consideran reglas difusas de asociación [2], pero con variables del antecedente y consecuente previamente fijadas. La extracción se realiza mediante sistemas difusos genéticos. Como es natural, llegado a este punto surgen dos cuestiones importantes: ¿por qué reglas difusas? y ¿por qué algoritmos genéticos (AGs)? Es decir, por qué usar estas herramientas de representación y aprendizaje frente a otras ampliamente usadas en KDD.

El uso de reglas difusas (en lugar de reglas intervalares, árboles de decisión, etc.) viene justificado principalmente por el tipo de dato manejado (ver sección 3.1). En nuestro caso, cada variable se compone de un conjunto de parámetros (ítemes) que aportan información parcial para describirla. Esto añade incertidumbre en los datos que puede tratarse adecuadamente con reglas difusas. Además, es posible expresar el conocimiento experto disponible en semánticas lingüísticas. Por último, las reglas difusas obtenidas presentan una gran legibilidad y ofrecen transiciones graduales de las relaciones, aspectos importantes en KDD.

Respecto al uso de AGs para inducir estas reglas difusas en lugar de otras técnicas de aprendizaje automático, se debe a los siguientes aspectos. Por un lado, dado que la calidad de las diferentes reglas difusas se valoran mediante objetivos contradictorios (tales como soporte y confianza), realizamos optimización multiobjetivo para tratarlos adecuadamente. Este es actualmente uno de los aspectos más prometedores y una de las señas de identidad de los AGs, donde se están destacando por su alto rendimiento frente a otras técnicas. Además, para conseguir mayor compacidad y, por tanto, interpretabilidad, consideramos una representación flexible de reglas difusas que puede manejarse cómodamente con AGs.

El trabajo se organiza de la siguiente forma. La sección 2 describe brevemente el problema abordado. La sección 3 introduce la metodología de KDD propuesta. La sección 4 muestra algunos resultados experimentales. Por último, la sección 5 presenta algunas conclusiones y trabajos futuros.

* Trabajo soportado en parte por el Proyecto TIC2002-04036-C05-01 del MCyT y fondos FEDER

Actas del IV Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados, MAEB2005, pp.681-688
ISBN: Vol I:84-9732-440-4 Vol II: 84-9732-467-6 **© 2005 Los autores, Thomson**

2. Modelado Causal en Marketing

Los académicos y profesionales de marketing han destacado la necesidad existente por conocer y explicar los patrones de comportamiento del consumidor de una manera cada vez más eficiente. Esto se debe principalmente a que las empresas centradas en los mercados de consumo operan en entornos altamente competitivos en los que se precisa que sus procesos de decisión sean lo más acertados posible. En este sentido, los modelos de comportamiento del consumidor se consideran como un caso específico de Sistemas de Apoyo a la Gestión y a lo largo del tiempo han demostrado ser un recurso fundamental para el desarrollo de la ciencia del marketing [6], así como instrumento para el soporte de las decisiones empresariales.

No obstante, los modelos actuales de comportamiento del consumidor no parecen cubrir todas las necesidades que supuestamente debería satisfacer un modelo que persiga asistir a la toma de decisiones de marketing. Así, en tanto que el principal problema que actualmente afrontan las empresas orientadas hacia los mercados de consumo no es tanto la disposición de información (datos), como la posesión de los niveles apropiados de conocimiento para tomar las decisiones adecuadas, la utilización de técnicas de extracción de conocimiento vanguardistas capaces de explotarla puede representar una ventaja competitiva.

En concreto, nuestro trabajo se centra en la mejora de las técnicas de estimación utilizadas. De este modo, mostramos los potenciales que presentan métodos de estimación basados en reglas difusas de cara a obtener modelos que sean más exhaustivos, complejos, flexibles, interactivos y que ofrezcan mucha más información cualitativa que las técnicas de estimación precedentes utilizadas en este campo [3, 6].

3. Modelado del Comportamiento del Consumidor mediante KDD basado en Reglas Difusas

Esta sección introduce el método propuesto de KDD para estimar el comportamiento del consumidor. Básicamente, consiste en preparar los datos y fijar el esquema que seguimos para representar el conocimiento del experto disponible. Una vez definidos estos aspectos, proponemos un método de aprendizaje automático basado en AGs para extraer patrones difusos relevantes.

3.1. Recogida de datos

El primer paso es la recolección de datos asociados con los elementos que definen el modelo teórico de comportamiento del consumidor propuesto; la mayoría de ellos usualmente son constructos o variables latentes. Así, como se ha realizado tradicionalmente en marketing, los datos se obtienen por medio de cuestionarios. Por tanto, en primer lugar, los diseñadores del modelo deben afrontar y desarrollar el proceso de medición de las variables que conforman el modelo de comportamiento.

En concreto, el caso más controvertido es el de la medición de las variables latentes. Se pueden sintetizar en dos las corrientes de medición para este tipo de variables, dependiendo de que defiendan que la medición de estos constructos puedan o no puedan medirse perfectamente por medio de variables observadas (indicadores). En otras palabras, si considera que existe o no existe una correspondencia unívoca entre un constructo y su medida. Ciertamente, aunque los diseñadores de modelos tendieron en un principio a utilizar lo que se conoció como la *filosofía de la definición operativa*, se postergó en pos de una aproximación más conveniente y razonable basada en la *filosofía de la interpretación parcial*. Esta aproximación distingue entre variables no observadas (constructos) y observadas (indicadores), siendo actualmente la predominante en el modelado de marketing. En esencia, considera múltiples indicadores —imperfectos cuando se consideran de forma individual, aunque fiables cuando se consideran de manera conjunta— asociados al constructo subyacente para obtener medidas válidas del mismo.

Por ejemplo, supongamos que disponemos de un sencillo modelo de medida presentado gráficamente en la figura 1. El modelo está compuesto de tres constructos o variables latentes, dos exógenas (*velocidad de interacción*

e *invasión de privacidad*) y una endógena (*actitud hacia Internet*). En cuanto a la escala de medida, supongamos, por un lado, que el primer y segundo constructo han sido medidos mediante diversas escalas de tipo Likert de nueve puntos que van desde 1: *absolutamente en desacuerdo* hasta 9: *absolutamente de acuerdo*. Por otro, se han empleado escalas de diferencial semántico con 9 puntos para el tercer constructo. Concretamente, en el cuadro 1 presentamos un ejemplo hipotético del conjunto de ítemes que podrían usarse para cada variable. Cada ejemplo o instancia del conjunto de datos sería la respuesta de un consumidor a estas cuestiones.

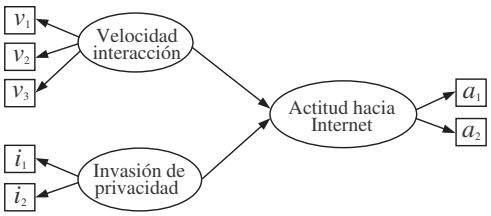

Figura 1: Ejemplo de un modelo de medida (estructural) simple

3.2. Procesado de Datos

En segundo lugar, es necesario adaptar los datos recogidos a un esquema fácilmente tratable por los métodos de aprendizaje difusos. Por tanto, nuestra aproximación metodológica debe ser consciente de las características particulares de los datos disponibles (con varios ítemes o indicadores para describir cierto elemento —variable latente— del modelo) cuando se adapten las variables observadas a un método de aprendizaje de reglas difusas. Una primera aproximación o solución intuitiva podría ser reducir directamente los ítemes asociados a un determinado constructo a un único valor: por ejemplo, la media aritmética.

El problema de estas aproximaciones es que los datos originales se deben transformar, por lo que se puede perder o tergiversar información relevante. Para solventar estos inconvenientes, proponemos un proceso más sofisticado que permite trabajar con el formato original de los datos: la *fuzzificación multi-ítem*. Al tratarse de un componente más del diseño del aprendizaje automático, los detalles del tratamiento de los ítemes se describen en la sección 3.4.2. En proceso consiste en hacer uso del operador T-conorma (máximo), tradicionalmente utilizado para desarrollar la unión de conjuntos difusos, para agregar la información parcial ofrecida por cada ítem durante el proceso de inferencia. Este proceso nos permite además trabajar con datos incompletos en caso de que el consumidor no responda a todos los ítemes de una determinada variable. Cuando esto sucede, se opera sólo con los ítemes disponibles. Dado el tipo de T-conorma utilizado, esto es equivalente a rellenar los ítemes en blanco con alguno de los ítemes disponibles para la correspondiente variable en la instancia considerada. Cuando no se dispone de ningún ítem para una determinada variable, se elimina la instancia en el preprocesado.

3.3. Representación e Inclusión del Conocimiento Experto

En esta etapa se deben abordar diversas cuestiones: el conjunto de variables (elementos del modelo) que se van a modelar, la transformación de las escalas de marketing utilizadas para la medición de las variables en semánticas difusas y la estructura de las reglas difusas (definición de las variables que constituyen el antecedente y consecuente de las reglas). Todas ellas se basan en la capacidad del experto para expresar su conocimiento en un formato interpretable por la lógica difusa.

3.3.1. Semánticas Difusas

Una vez que determinado tanto los constructos teóricos del modelo, como las variables observadas asociadas a cada uno de ellos (es decir, el modelo de medida), se debe realizar una transformación de las escalas originales de marketing a términos difusos. Con el objeto de simplificar el problema, en este trabajo nos centramos en las escalas de intervalo (es decir, escalas de tipo Likert y de diferencial semántico), por ser las más comúnmente utilizadas.

Cuadro 1: Ejemplo de un cuestionario asociado al modelo estimado de la figura 1

Velocidad de interacción	
V_1:	La interacción con las páginas Web es rápida y estimulante: $\{1, \ldots, 9\}$
V_2:	Internet es rápido: $\{1, \ldots, 9\}$
V_3:	Las páginas Web que visito usualmente se cargan lo suficientemente rápido: $\{1, \ldots, 9\}$
Invasión de privacidad	
I_1:	Cuando navego por la Web siento que mi privacidad está siendo invadida: $\{1, \ldots, 9\}$
I_2:	Las empresas en Internet no respetan la intimidad del visitante: $\{1, \ldots, 9\}$
Actitud hacia Internet	
A_1:	*Negativa* 1 2 3 4 5 6 7 8 9 *Positiva*
A_2:	*Desfavorable* 1 2 3 4 5 6 7 8 9 *Favorable*

La transformación se debe realizar considerando tres cuestiones fundamentales:

1. Se debe establecer el *número de términos lingüísticos* a utilizar en cada variable. Un número impar parece ser una buena aproximación si consideramos que en nuestro caso es útil expresar lingüísticamente un punto medio o de indiferencia. Teniendo en cuenta que las escalas de intervalo tradicionalmente utilizadas presentan entre 6 y 11 puntos, es suficiente tres o cinco términos lingüísticos (conjuntos difusos) para representar estos valores.

2. También se debe definir el *tipo de función de pertenencia* que definirá el comportamiento de la variable difusa. En nuestro caso, consideramos más apropiado utilizar funciones lineales (trapezoidales o triangulares) en lugar de no lineales (gaussianas) para caracterizar los conjuntos difusos ya que facilita la interpretación de la escala de medida.

3. Finalmente, se deben establecer las *formas de las funciones de pertenencia*. En este respecto, proponemos imponer algunas restricciones con el objeto de asegurar una buena interpretación. Los valores extremos del intervalo deben tener un grado de pertenencia 1 a las etiquetas extremas de la variable lingüística. El valor medio del intervalo debe tener un grado de pertenencia 1 a la etiqueta media. Además, consideramos semántica difusa *fuerte* de Ruspini (donde la suma de los grados de pertenencia de cada valor al conjunto de términos lingüísticos es 1) para asegurar una buena interpretación. Por último, con el objeto de evitar el sesgo estadístico de cada término lingüístico, forzamos a que todos tengan el mismo grado de cobertura. De este modo, consideramos definiciones de las funciones de pertenencia tales que, dado el conjunto $S = \{min, \ldots, max\}$ que determina una variable medida mediante escala de intervalo, cumplan la siguiente condición:

$$\sum_{k \in S} \mu_{A_i}(k) = \frac{max - min}{l}, \quad \forall A_i \in \mathcal{A},$$

siendo l el número de términos lingüísticos y $\mathcal{A} = \{A_1, \ldots, A_l\}$ el conjunto de éstos.

En resumen, la figura 2 muestra un ejemplo de transformación de una escala de intervalo de nueve puntos (usualmente empleada en marketing para medir las variables observadas asociadas a los constructos) en una semántica difusa con tres términos lingüísticos.

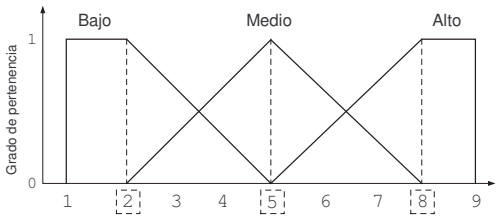

Figura 2: Transformación de una escala de medida de nueve puntos en una semántica difusa

3.3.2. Variables Lingüísticas del Antecedente y Consecuente

Además, una vez que el experto de marketing haya definido el modelo estructural, se utilizan reglas difusas para relacionar las variables del

antecedente con las del consecuente. Naturalmente, las hipótesis contenidas en el modelo se pueden utilizar para definir directamente las estructuras SI-ENTONCES a partir de las dependencias existentes entre las variables. De esta manera, consideramos como consecuente de la regla difusa el constructo endógeno y como antecedente los constructos relacionados con él. Por ejemplo, partiendo del modelo de medida de la figura 1, la estructura de la regla difusa tendrá la siguiente forma:

SI *Velocidad Interac.* es A_1 y *Invasión Privac.* es A_2 **ENTONCES** *Actitud Internet* es B.

3.4. Proceso de Minería de Datos

Una vez fijadas las variables lingüísticas que representen adecuadamente la información, se debe emplear un proceso de aprendizaje para extraer automáticamente el conocimiento existente en los datos considerados. Este proceso es, sin lugar a dudas, el más importante desde el punto de vista del KDD. Como hemos mencionado en la sección 1, en este trabajo estamos interesados en realizar inducción descriptiva. Por tanto, empleamos un AG estilo Michigan para obtener reglas individuales relevantes. Consideraremos dos criterios de calidad, el soporte (grado de representatividad de la regla respecto al conjunto de datos) y la confianza (grado de precisión de la relación establecida en la regla). Resulta intuitivo comprobar que a mayor soporte, mayor dificultad para mantener altos grados de confianza. Para ello, proponemos el uso de un AG multiobjetivo debido a su buen comportamiento para trabajar con múltiples objetivos contradictorios. Los siguientes apartados describen los principales componentes del método propuesto.

3.4.1. Estructura de Regla Difusa

Para garantizar un alto grado de legibilidad hemos optado por una descripción más compacta de las relaciones difusas basada en la *forma normal disyuntiva* (DNF). Este tipo de regla difusa tiene la siguiente estructura:

SI X_1 es $\widetilde{A_1}$ y ... y X_n es $\widetilde{A_n}$ **ENTONCES** Y es B

donde cada variable de entrada X_i, $i \in \{1, \ldots, n\}$, toma valores de un conjunto de términos lingüísticos $\widetilde{A_i} = \{A_{i1} \vee \ldots \vee A_{il_i}\}$, cuyos miembros se unen mediante un operador de disyunción (en nuestro caso, la T-conorma de la *suma ponderada*, $min\{1, a+b\}$). Esta estructura es un soporte natural para permitir la ausencia de variables de entrada en cada regla simplemente haciendo que $\widetilde{A_i}$ sea el conjunto de términos lingüísticos completo.

3.4.2. Fuzzificación Multi-ítem

Con el objeto de considerar apropiadamente el conjunto de ítemes disponible para cada variable (discutido en la sección 3.2), proponemos una extensión del cálculo del grado de pertenencia que llamaremos *fuzzificación multi-ítem*. El proceso se basa en la unión de la información parcial aportada por cada ítem para describir su variable correspondiente. Dado X_i y Y medidas mediante los vectores de ítemes $\vec{x}_i = (x_1^{(i)}, \ldots, x_{h_i}^{(i)}, \ldots, x_{p_i}^{(i)})$ y $\vec{y} = (y_1, \ldots, y_t, \ldots, y_q)$, respectivamente, las proposiciones difusas "X_i es $\widetilde{A_i}$" y "Y es B" se interpretan respectivamente como sigue:

$$\mu_{\widetilde{A_i}}(\vec{x}_i) = \text{mín}\left\{1, \bigvee_{h_i=1}^{p_i} \sum_{A \in \widetilde{A_i}} \mu_A(x_{h_i}^{(i)})\right\},$$

$$\mu_B(\vec{y}) = \bigvee_{t=1}^{q} \mu_B(y_t),$$

con $\vee$ siendo una T-conorma (en este trabajo, el *máximo*).

3.4.3. Descubrimiento de Subgrupos

Para realizar el proceso de inducción descriptiva aplicaremos un método con ciertas similitudes a la técnica de descubrimiento de subgrupos, ampliamente utilizado en aprendizaje de reglas de clasificación [4] donde la propiedad de interés es la clase asociada a las variables del consecuente. Por tanto, se persigue agrupar el conjunto de datos en subgrupos diferenciados incluyendo en cada uno de ellos aquellos ejemplos representados por el consecuente y descubrir un conjunto de reglas representativo para cada subgrupo. Así, el enfoque más habitual consiste en ejecutar el algoritmo diseñado en cada subconjunto de datos que satisfagan la propiedad del consecuente fijado.

En lugar de esto, en el algoritmo que proponemos realizamos un descubrimiento de subgrupos *simultáneo*, en el cual se forman nichos de reglas difusas diferenciadas por el consecuente y se optimizan en paralelo para generar finalmente un conjunto de soluciones subóptimas en cada uno. Para realizar este proceso simultáneo, como se muestra en las siguientes secciones, variamos el concepto de dominancia multiobjetivo y hacemos que los operadores genéticos actúen únicamente en la parte del antecedente de las reglas.

3.4.4. Esquema de Codificación

Cada individuo de la población representa una regla difusa. El esquema de codificación es binario para representar el antecedente y entero para el consecuente. Así, el alelo '1' en la parte del antecedente significará que el término lingüístico asociado al gen se usa en la variable correspondiente. En el consecuente se codificará directamente el índice del término lingüístico usado. Por tanto, el tamaño para codificar una regla difusa DNF será igual a la suma del número de términos lingüísticos empleado en cada variable de entrada más el número de variables de salida. Por ejemplo, suponiendo tres términos lingüísticos para cada variable, la regla [SI X_1 es Pequeño y X_2 es {Mediano o Grande} ENTONCES Y es Mediano] se codifica como [100 011|2].

3.4.5. Funciones Objetivo

Consideramos los dos criterios empleados con mayor frecuencia para valorar la calidad de las reglas de asociación [2]: soporte y confianza.

Soporte: Esta medida valora el grado de representatividad de la regla difusa entre el conjunto de datos analizados. Se calcula como el grado de cubrimiento medio de la regla a cada uno de estos datos. Como cubrimiento consideramos la conjunción entre los grados de pertenencia a las distintas variables, tanto del antecedente como del consecuente. La medida de soporte (para maximización) de la regla difusa $R : A \Rightarrow B$ queda definida como sigue:

$$Sop(R) = \frac{1}{N} \sum_{e=1}^{N} T\left(\mu_A\left(\mathbf{x}^{(e)}\right), \mu_B\left(\vec{y}^{(e)}\right)\right)$$

siendo N el tamaño del conjunto de datos, $\mathbf{x}^{(e)} = (\vec{x}_1^{(e)}, \ldots, \vec{x}_n^{(e)})$ y $\vec{y}^{(e)}$ la e-ésima instancia multi-ítem de entrada y salida respectivamente, T la T-norma del *producto*, y $\mu_A(\mathbf{x}^{(e)}) = \text{mín}_{i\in\{1,\ldots,n\}}\, \mu_{\widetilde{A}_i}(\vec{x}_i^{(e)})$ el grado de cubrimiento del antecedente de la regla R para este ejemplo (es decir, se considera la T-norma del *mínimo* para interpretar el conectivo "y" de la regla difusa). Recordemos que empleamos la fuzzificación multi-ítem descrita en la sección 3.4.2 para calcular $\mu_{\widetilde{A}_i}(\vec{x}_i^{(e)})$ y $\mu_B(\vec{y}^{(e)})$.

Confianza: Este segundo objetivo valora la fiabilidad de la relación entre antecedente y consecuente descrita por la regla difusa analizada. Se ha empleado una medida de confianza que evite la acumulación de pequeñas cardinalidades [2]. Se calcula (para maximización) del siguiente modo:

$$Con(R) = \frac{\sum_{e=1}^{N} T\left(\mu_A\left(\mathbf{x}^{(e)}\right), I\left(\mu_A\left(\mathbf{x}^{(e)}\right), \mu_B\left(\vec{y}^{(e)}\right)\right)\right)}{\sum_{e=1}^{N} \mu_A\left(\mathbf{x}^{(e)}\right)}$$

Se emplea la S-implicación de Dienes, $I(a,b) = max\{1-a, b\}$. Se considera nuevamente la T-norma del producto y la fuzzificación multi-ítem.

3.4.6. Esquema Evolutivo

Consideramos enfoque generacional con la estrategia de reemplazo elitista multiobjetivo de NSGA-II [1] y selección mediante torneo binario basado en la distancia de concentración (*crowding distance*) en el espacio de las funciones objetivo. Para realizar correctamente el descubrimiento de subgrupos simultáneo necesitamos redefinir el concepto de dominancia. Para ello, una solución (regla) dominará a otra cuando, además de al menos igualar en todos los objetivos y mejorar en uno de ellos, posee el mismo consecuente que la otra regla. De esta

forma, aquellas reglas con consecuentes distintos no se dominan entre sí, con lo cual provocamos que el algoritmo forme tantos nichos de búsqueda (conjuntos Pareto) como consecuentes distintos (subgrupos) se consideren.

3.4.7. Operadores Genéticos

La población inicial se construye definiendo tantos grupos (de igual tamaño) como consecuentes distintos haya. En cada uno de ellos, los cromosomas se generan fijando dicho consecuente y construyendo de forma aleatoria un antecedente simple donde cada variable de entrada se asocia con un término lingüístico. Los dos operadores de reproducción sólo actúan en la parte del antecedente de la regla. Esto hace que el tamaño de cada subgrupo en la población sea constante.

El operador de *cruce* selecciona dos puntos de corte (en la parte del antecedente) e intercambia la subcadena central. El operador de *mutación* selecciona aleatoriamente una variable del antecedente de la regla difusa codificada en el cromosoma y realiza alguna de las tres siguientes operaciones (escogida aleatoriamente entre las posibles): *expansión*, que cambia a '1' un gen escogido aleatoriamente entre aquellos con valor '0'; *contracción*, que cambia a '0' un gen escogido aleatoriamente entre aquellos con valor '1'; o *desplazamiento*, que cambia a '0' un gen escogido aleatoriamente entre aquellos que contengan el valor '1' y que cuenten con al menos un gen adyacente con valor '0', y cambia a '1' el gen inmediatamente anterior o posterior a él.

4. Resultados Experimentales e Interpretación

Para la experimentación hemos considerado un modelo estructural basado en el análisis del *estado flow* del consumidor en entornos informáticos interactivos. Los datos se han obtenido a partir de la encuesta usada en [5] para validar un modelo conceptual presentado previamente por los mismos autores. Para ilustrar el concepto de estado *flow* en Internet, podemos decir que se alcanza cuando el consumidor está tan intensamente concentrado en navegar por la Web que "nada más parece importarle" [5]. Los datos de entrenamiento constan de 1.154 ejemplos (respuestas de consumidores). Debido a las limitaciones de espacio de este trabajo, nos hemos centrado en analizar una relación específica entre las seis disponibles en el modelo estructural, con un total de 12 variables. En nuestro caso consideramos cuatro constructos (*velocidad de interacción*, *habilidad/control*, *desafío/estímulo* y *telepresencia/distorsión de tiempo*) como antecedentes primarios del estado *flow* del consumidor. En este sentido, se considera la hipótesis de que estos cuatro constructos están relacionados positivamente con el constructo central.

Hemos empleado la semántica difusa de la figura 2 para todas las variables. Se han realizado 10 ejecuciones del algoritmo con los siguientes valores de los parámetros: 100 generaciones, tamaño de la población 100, probabilidad de cruce 0,7 y probabilidad de mutación por cromosoma 0,1.

En la figura 3 se muestran los frentes Pareto obtenidos. A partir de ellos podemos observar que, en lo que respecta al valor que toma el consecuente *flow* en las reglas generadas, la salida más verosímil es "medio". En efecto, existe una clara supremacía de las reglas con esta salida sobre las dos restantes en lo que a soporte y confianza de las mismas se refiere. Este hecho se intensifica a medida que el soporte de las reglas aumenta, sin que exista una pérdida de fiabilidad ostensible. Por tanto, se puede deducir que el nivel del estado de *flow* más representativo de los consumidores cuando desarrollan procesos de navegación Web es moderado.

Podemos comprender mejor el comportamiento del consumidor analizando algunas reglas difusas seleccionadas (marcadas con un símbolo negro en la figura 3) y recogidas junto a sus valores de soporte y confianza en el cuadro 2. Dentro del conjunto Pareto real, R_1 es la regla de mayor soporte con consecuente *bajo*, R_2 representa la regla con mayor confianza con consecuente *medio*, y R_3 es la regla de mayor soporte con confianza superior a 0,7. De forma sintética, de todos los antecedentes

Cuadro 2: Algunas reglas difusas obtenidas y sus correspondientes valores de soporte y confianza

	Veloc. Interac	*Habilidad/Control*	*Desafío/Estímulo*	*Telepres./Dist. Tiempo*	*Flow*	Sop	Con
R_1	bajo alto	medio		bajo	bajo	0,0104	0,7980
R_2	medio	bajo alto	alto	medio	medio	0,0120	0,7937
R_3	medio			medio alto	medio	0,3947	0,7051

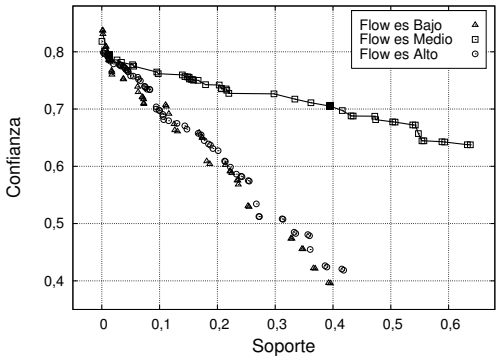

Figura 3: Frente Pareto de cada subgrupo y frente Pareto real (unido mediante una línea)

considerados destaca la influencia de la percepción de *telepresencia/distorsión del tiempo* (TP/DT) sobre el estado de *flow* del consumidor; obsérvese cómo su valor es determinante para la existencia de estados de *flow* reducidos (R_1) o moderados (R_2 y R_3). Asimismo, el resto de antecedentes parecen ejercer una influencia escasa o nula sobre el consecuente. Este hecho puede deberse también a que la influencia del elemento TP/DT eclipse la influencia del resto. En cualquier caso, se ratifica la idea principal que extraíamos del gráfico del frente Pareto, es decir, la no existencia de combinaciones de antecedentes que produzcan estados de *flow* elevados con una fiabilidad y representatividad significativas; obsérvese cómo incluso cuando el antecedente más influyente (TP/DT) toma valores elevados, el estado de *flow* del consumidor en el proceso de navegación tiende a permanecer moderado.

5. Conclusiones

Se ha presentado un problema de KDD novedoso —modelado causal en marketing— y su resolución mediante sistemas difusos genéticos. El problema presenta un tipo específico de dato con incertidumbre que justifica el uso de reglas difusas. Además, se realiza optimización multiobjetivo con el fin de obtener reglas con altos grados de soporte y confianza. La metodología de KDD propuesta se ha aplicado de forma satisfactoria en un problema real de análisis del comportamiento del consumidor en entornos informáticos interactivos.

Como trabajo futuro nos planteamos el uso de otras métricas y el aprendizaje no supervisado sin antecedente y consecuente previamente fijado por el experto.

Referencias

[1] K. Deb, A. Pratap, S. Agarwal, T. Meyarevian. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evolutionary Computation*, 6(2):182–197, 2002.

[2] D. Dubois, H. Prade, T. Sudkamp. On the representation, measurement, and discovery of fuzzy associations. *IEEE Trans. Fuzzy Systems*, 13(2):250–262, 2005.

[3] H. Gatignon. Commentary on P. Leeflang and D. Wittink's "Building models form marketing decisions: past, present and future". *Int. J. Research in Marketing*, 17:209–214, 2000.

[4] N. Lavrac, B. Cestnik, D. Gamberger, P. Flach. Decision support through subgroup discovery: three case studies and the lessons learned. *Machine Learning*, 57(1-2):115–143, 2004.

[5] Y. Novak, D. Hoffman, Y. Yung. Measuring the customer experience in online environments: a structural modelling approach. *Marketing Science*, 19(1):22–42, 2000.

[6] G. van Bruggen, B. Wierenga. Broadening the perspective on marketing decision models. *Int. J. Research in Marketing*, 17:159–168, 2000.