Inducción evolutiva multiobjetivo de reglas de descripción de subgrupos en un problema de marketing

María José del Jesus
Dpto. de Informática
Universidad de Jaén
23071 Jaén
mjjesus@ujaen.es

Pedro González
Dpto. de Informática
Universidad de Jaén
23071 Jaén
pglez@ujaen.es

Francisco Herrera
Dpto. de Ciencias de la Computación
e Inteligencia Artificial
Universidad de Granada
18071 Granada
herrera@decsai.ugr.es

Resumen

En este trabajo se presenta una propuesta para la extracción de conocimiento en un problema del área de marketing, el estudio de la influencia que tienen las variables de planificación de un certamen ferial sobre el nivel de consecución de los objetivos planteados previamente para el mismo. En este problema el objetivo es extraer reglas que describan subgrupos aportando información relevante sobre los mismos.

El algoritmo evolutivo de inducción de reglas de descripción de subgrupos (*subgroup discovery*) propuesto sigue un enfoque multiobjetivo para optimizar de forma adecuada los distintos indicadores de calidad utilizados en este tipo de problemas. Los resultados obtenidos muestran la adecuación del enfoque multiobjetivo para la obtención de un conjunto de reglas de cardinalidad variable que describe el conocimiento extraído con un nivel alto de confianza, soporte e interés.

1. Introducción

En el área de marketing, y en concreto en la planificación de ferias de muestras, es importante extraer conclusiones de información de ferias anteriores para determinar la relación entre las variables de planificación ferial y el éxito del stand. Para este problema es adecuado un algoritmo de inducción de reglas de descripción de subgrupos.

En el área de descubrimiento de subgrupos (*subgroup discovery*) [25] cualquier algoritmo de inducción de reglas debe optimizar simultáneamente distintos objetivos como el nivel de interés o novedad, el soporte o nivel de generalidad y la precisión del conocimiento extraído, entre otros. La combinación de dichos objetivos en una única medida de calidad suele llevar a una compensación entre los mismos que hace que no se optimicen de forma adecuada los objetivos particulares.

Los Algoritmos Genéticos (AGs) [18][14] tienen un carácter de búsqueda global que hace que sean especialmente adecuados para resolver distintos problemas presentes en cualquier proceso de descubrimiento de conocimiento [11]. En particular, los AGs multiobjetivo son adecuados para resolver problemas en los que intervienen distintos objetivos. En la bibliografía se pueden encontrar distintas propuestas evolutivas para optimización multiobjetivo [5][2]. Recientemente los AGs multiobjetivo se han empleado para extracción de conocimiento en minería de datos [16] [21].

En este trabajo se presenta una propuesta para la inducción de reglas de descripción de subgrupos basada en un AG multiobjetivo, que aúna el método de razonamiento aproximado de los sistemas difusos con las capacidades de aprendizaje de los AGs.

El artículo se organiza de la siguiente forma: la sección 2 describe el problema de marketing y en la sección 3 se realiza una revisión de AGs para la inducción de reglas de descripción de subgrupos. En la sección 4 se describen brevemente los AGs multiobjetivo y en la sección 5 se detalla el algoritmo propuesto. Finalmente, las secciones 6 y 7 muestran la experimentación realizada y las conclusiones obtenidas.

Actas del IV Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados, MAEB2005, pp.661-669
ISBN: Vol I:84-9732-440-4 Vol II: 84-9732-467-6

2. Extracción de conocimiento en certámenes feriales

Las empresas consideran los certámenes feriales un instrumento que facilita la consecución de objetivos comerciales. Uno de los principales inconvenientes de este tipo de certámenes es la elevada inversión que suponen. A esta inversión a veces se une una falta de planificación que enfatiza la sensación de que las ferias no son más que un "gasto" que las compañías han de afrontar por motivos varios (tradición, exigencia de los clientes, no dar la sensación de que las cosas van mal, etc.) [27]. Es conveniente, por tanto, la extracción automática de información sobre las variables implicadas que permita obtener datos desconocidos, determinantes en parte de la eficacia de los stands de un certamen.

En la Bienal de Máquina-Herramienta celebrada en Bilbao en Marzo de 2002, se recogió información sobre 104 variables de 228 expositores. De este conjunto de variables, 7 son continuas y el resto nominales, resultado de una discretización experta. Además, para cada uno de los expositores, en base a distintos criterios de marketing, se determinó la eficacia global de dicho stand en Eficacia Alta, Media o Baja en función del nivel de consecución de los objetivos planteados para el certamen.

El objetivo del proceso de extracción de conocimiento para este problema es determinar la aportación que las distintas variables de planificación ferial ejercen sobre los resultados obtenidos por el expositor.

3. Algoritmos genéticos para inducción de reglas de descripción de subgrupos

En este trabajo se propone un AG de inducción de reglas que describen subgrupos, tarea que se encuadra dentro del área de minería de datos. Por ello, en esta sección se describe brevemente la tarea de descubrimiento de subgrupos y las tendencias generales en AG de inducción de reglas.

3.1. Descubrimiento de subgrupos

Dentro del área de minería de datos el descubrimiento de subgrupos es un tipo de inducción descriptiva que ha recibido últimamente mucha atención por parte de los investigadores. Se trata de una forma de aprendizaje inductivo supervisado de descripciones de subgrupos, en la que, dado un conjunto de datos y teniendo una propiedad de interés para el usuario, se intentan localizar subgrupos que sean de mayor interés para el usuario.

El descubrimiento de subgrupos [34][12] tiene como objetivo descubrir propiedades características de subgrupos construyendo reglas sencillas (con pocas variables), altamente significativas y con completitud alta (que cubran muchas instancias de la clase objetivo).

El concepto fue formulado inicialmente por Klosgen en su algoritmo de aprendizaje EXPLORA [24] y por Wrobel en el algoritmo MIDOS [34]. Tanto éstos como otros algoritmos posteriores como SD [12] o CN2-SD [26] son adaptaciones de modelos de extracción de reglas de clasificación para la tarea del descubrimiento de subgrupos. Actualmente se muestra interés en el desarrollo de enfoques de descubrimiento de subgrupos partiendo de algoritmos de extracción de reglas de asociación [25][1][[23].

3.2. Inducción evolutiva de reglas

Se han desarrollado múltiples propuestas de AGs para la extracción de reglas de distintos tipos, de clasificación, asociación o dependencias funcionales. Para el problema al que nos enfrentamos el objetivo es generar reglas en las que en el consecuente aparezca un único atributo (y establecido a priori), por lo que en esta sección haremos referencia a AGs para la extracción de reglas de clasificación y asociación.

El aspecto más determinante de cualquier AG de inducción de reglas es el esquema de codificación utilizado. En este aspecto, las distintas propuestas en la bibliografía especializada se agrupan en torno a dos enfoques [4]:

- El enfoque "Cromosoma = Regla", en el que cada individuo codifica una única regla.
- El enfoque "Cromosoma = Base de Reglas", o enfoque *Pittsburgh*, en el que cada individuo codifica un conjunto de reglas. GIL [22] y GA-MINER [9] son ejemplos de AGs de este tipo.

A su vez, dentro del enfoque "Cromosoma = Regla" existen tres propuestas genéricas:

- El enfoque *Michigan* en el que cada individuo codifica una única regla pero la solución final será la población final o un subconjunto de la misma. En este caso es necesario evaluar el comportamiento del conjunto de reglas al completo y la aportación de la regla individual al mismo, definiendo un componente de refuerzo. El algoritmo XCS [33] es un ejemplo de este enfoque.
- El enfoque IRL (*Iterative Rule Learning*) en el que cada cromosoma representa una regla, pero la solución del AG es el mejor individuo y la solución global está formada por los mejores individuos de una serie de ejecuciones sucesivas. SLAVE [15] es una propuesta con este modelo.
- El enfoque "cooperative-competitive", en el que la población completa o un subconjunto de ella codifica la base de reglas. REGAL [13] es un ejemplo de AG con este tipo de representación.

En procesos de descubrimiento de reglas de descripción de subgrupos es más adecuado el enfoque "Cromosoma = Regla" ya que el objetivo es encontrar un conjunto reducido de reglas en las que la calidad de cada regla se evalúa de forma independiente a la del resto. Este es el enfoque de codificación utilizado en esta propuesta evolutiva.

4. Algoritmos genéticos multiobjetivo

Como se ha comentado anteriormente, en el área de descubrimiento de subgrupos cualquier algoritmo de inducción de reglas debe optimizar simultáneamente distintos objetivos. La forma más adecuada de abordarlos es mediante algoritmos de optimización multiobjetivo en los que se busca un conjunto de soluciones alternativas (reglas en este caso) optimales en el sentido de que ninguna otra solución dentro del espacio de búsqueda sea superior a ella en todos los objetivos considerados. El experto utilizará el conjunto de reglas de salida para seleccionar todas o un subconjunto de ellas para la descripción de los subgrupos en función de la información de preferencia particular del problema.

Formalmente un problema de optimización multiobjetivo se puede definir de la siguiente forma:

$$\min/\max \vec{y} = f(\vec{x}) = f_1(\vec{x}), f_2(\vec{x}), \ldots, f_n(\vec{x}))$$

donde $\vec{x} = (x_1,x_2,...x_m)$ es el vector de decisión e $\vec{y} = (y_1,y_2,...,y_n)$ es el vector objetivo (una tupla con n objetivos). El objetivo de cualquier algoritmo de optimización multiobjetivo es encontrar todos los vectores de decisión para los cuales los correspondientes vectores objetivo no se puedan mejorar en una dimensión sin degradar otra, a lo que se denomina el conjunto Pareto optimal.

En las dos últimas décadas se ha desarrollado un interés creciente en el uso de AGs para optimización multiobjetivo. Existen múltiples propuestas de AGs multiobjetivo [5][2] que se pueden agrupar en torno a tres enfoques:

- Métodos de agregación que combinan los objetivos en una función escalar. Tienen como inconveniente la posible compensación entre objetivos, el conocimiento profundo sobre el problema que requieren y que, en general, no proporcionan una familia de soluciones. VOW-GA [17] y RW-GA [20] son ejemplos de algoritmos de este tipo.
- Métodos basados en población, en los que la búsqueda se guía en diferentes direcciones para generar poblaciones de soluciones no dominadas. Dentro de este enfoque se encuadran, entre otros, los algoritmos MOGA [8], NPGA [19], NSGA [32] y NSGA II [6].
- Métodos que utilizan elitismo, manteniendo una población elite con soluciones no dominadas que intervienen de distinta forma en la evolución según las propuestas. Dentro de este enfoque se incluyen los algoritmos SPEA [35], SPEA2 [36] y μ-λ MEA [30]. Nuestra propuesta está basada en el algoritmo SPEA2.

5. Un enfoque evolutivo multiobjetivo para la obtención de reglas difusas que describen subgrupos

En este apartado se describe un AG multiobjetivo para la extracción de reglas que describen subgrupos. El algoritmo extrae reglas que representan en el antecedente una conjunción de variables y cuyo consecuente está prefijado. El objetivo de este proceso evolutivo es extraer para cada valor de la variable objetivo un número

variable de reglas diferentes que expresen información sobre los ejemplos del conjunto de partida. El algoritmo permite generar reglas difusas y/o nítidas, para problemas con variables continuas y/o nominales.

El AG multiobjetivo sigue el enfoque SPEA2 [36], y por tanto aplica los conceptos de elitismo en la selección de reglas (utilizando una población secundaria) y búsqueda de soluciones óptimas en el frente de Pareto (se ordena a los individuos de la población de acuerdo a si cada individuo es o no dominado usando el concepto de óptimo de Pareto).

Cualquier AG multiobjetivo debe diseñarse para lograr dos propósitos: lograr buenas aproximaciones al frente de Pareto y mantener la diversidad de las soluciones, con el objetivo de muestrear adecuadamente el espacio de soluciones y no converger a una solución única o a una sección acotada del frente. Para preservar la diversidad a nivel fenotípico el algoritmo utiliza una técnica de nichos que considera la cercanía en valores de los objetivos y un objetivo adicional sobre originalidad para potenciar reglas que aporten información sobre ejemplos no descritos por otras reglas de la población.

El Algoritmo 1 muestra el esquema de funcionamiento del modelo propuesto.

Inicialización: Generar la población inicial P_0 y crear una población elite vacía $P'_0 = Ø$.
Repetir
  Asignación de fitness: Calcular el fitness de los individuos de la población P_t y de la población elite P'_t de forma conjunta.
  Selección de entorno: Copiar todos los individuos no dominados de la población P_t y la población elite P'_t en la población elite P'_{t+1}. Si el tamaño de P'_{t+1} sobrepasa el número de elementos a guardar, reducir P'_{t+1} mediante el operador de truncado; en otro caso, si el tamaño de P'_{t+1} es inferior al número de elementos, rellenarlo con individuos dominados de P_t y de P'_t.
  Esquema de reproducción: Realizar selección por torneo binario con reemplazo sobre la población elite P'_{t+1} aplicando después operadores de cruce y mutación. El resultado es la población P_{t+1}.
Mientras no se verifique la condición de parada.
Devolver los individuos no dominados de la población elite P'_{t+1}.

Algoritmo 1. Esquema del algoritmo propuesto

A continuación se describen el esquema de representación, los objetivos, la asignación de fitness, la selección de entorno y el esquema de reproducción.

5.1. Esquema de representación

Cada solución candidata se codifica mediante el enfoque "Cromosoma = Regla" representando en el cromosoma sólo el antecedente y asociando todos los individuos de la población con el mismo valor de la variable objetivo. Esta forma de codificar la variable objetivo obliga a realizar una ejecución del AG multiobjetivo para cada valor de la variable objetivo (clase), pero asegura la extracción de información sobre todas y cada una de las clases.

Algunas variables del problema son variables continuas tratadas como variables lingüísticas con etiquetas lingüísticas. Los conjuntos difusos correspondientes a los términos lingüísticos vienen definidos mediante información experta o por una partición difusa uniforme.

Toda la información relacionada con una regla se almacena en un cromosoma de longitud fija para el que se utiliza un modelo de representación entera (la i-ésima posición indica el valor adoptado por la i-ésima variable). El conjunto de posibles valores que pueden tomar las variables discretas es el indicado por el problema más un valor adicional que indica que la variable correspondiente no interviene en la regla.

5.2. Definición de los objetivos del algoritmo

En el proceso de descubrimiento de reglas se intentan conseguir reglas con capacidad predictiva alta, comprensibles e interesantes. En nuestra propuesta, se han definido cuatro objetivos:

- *Confianza*. Determina la precisión de la regla ya que refleja el grado con el que los ejemplos pertenecientes a la zona del espacio delimitado por el antecedente verifican la información indicada en el consecuente de la regla. Se calcula mediante una adaptación de la expresión de precisión de Quinlan [29] utilizada en la generación de reglas de clasificación difusas [3]: la suma del grado de pertenencia de los ejemplos de la clase a la zona determinada por el antecedente dividido entre la suma del grado de pertenencia de

todos los ejemplos (independientemente de su clase) a la misma zona.

- *Completitud*. Es una medida del grado de cobertura que la regla ofrece a los ejemplos de la clase. Se calcula como el cociente entre el número de ejemplos de la clase descritos por la regla y el número total de ejemplos de la clase.
- *Interés*. El grado de interés se determina en esta propuesta objetivamente mediante el criterio de interés aportado por Noda et al. [28] en un proceso de modelado de dependencias. En nuestra propuesta se utiliza sólo la parte referente al antecedente para el cálculo del interés, puesto que el consecuente está prefijado. La medida de información para el interés se define de la siguiente forma [10]:

$$Interes = 1 - \left(\frac{\sum_{i=1}^{n} Ganancia(A_i)}{n \cdot \log_2 (|dom(G_k)|)} \right)$$

donde n es el número de variables que aparecen en el antecedente de la regla, $Ganancia(A_i)$ es la ganancia de información del atributo A_i, y $|dom(G_k)|$ es la cardinalidad (el número de valores posibles) de la variable objetivo. Para calcular la ganancia de información en el caso de variables numéricas se realiza una discretización de la variable en tantos intervalos como etiquetas lingüísticas se consideren.

Las variables con una alta ganancia de información son adecuadas para predecir una clase cuando se consideran de forma individual. Sin embargo, si el usuario conoce cuales son las variables más predictivas para un dominio específico, las reglas que contienen estas variables son menos interesantes. De esta forma, el antecedente de una regla es más interesante si contiene atributos con una pequeña cantidad de información.

- *Soporte original*: Este objetivo es una medida del nivel de originalidad de la regla comparando con el resto de reglas. Se calcula sumando para cada ejemplo perteneciente al antecedente de la regla el factor *1/k*, siendo *k* el número de reglas de la población que también describen información sobre ese ejemplo. Esta medida favorece la diversidad en la población a nivel fenotípico.

5.3. Cálculo del fitness

Se realiza de la siguiente forma:

- Se calcula el valor de todos los objetivos para cada individuo de la población.
- Los valores que alcanzan cada uno de los individuos de la población (y de la población elite) se utilizan para calcular qué individuos dominan a qué otros de las poblaciones.
- Se calcula la fuerza (*strength*) de cada individuo como el número de individuos a los que domina.
- Se determina el fitness inicial (*raw fitness)* de cada individuo, como la suma de la fuerza de sus dominadores (tanto en la población como en la población elite).
- La asignación del fitness inicial aporta un mecanismo de nichos basado en el concepto de la dominancia de Pareto, pero puede fallar cuando muchos de los individuos son no dominados. Para evitarlo, se incluye información adicional sobre densidad para discriminar entre individuos con los mismos valores de fitness inicial. La técnica de estimación de densidad utilizada en SPEA2 es una adaptación del método del k-ésimo vecino más cercano [31], donde la densidad en un punto es función decreciente de la distancia al punto k-ésimo más cercano. En esta propuesta se toma la inversa de la distancia al vecino k-ésimo más cercano como estimación de densidad.
- El valor de fitness de cada individuo es la suma de su valor de fitness inicial y su densidad.

5.4. Selección de entorno

En este algoritmo se establece un tamaño fijo para la población elite, de forma que es necesario definir una función de truncado y otra de rellenado. La función de truncado permite eliminar soluciones no dominadas de la población elite si excede el tamaño definido. Para ello se utiliza un esquema de nichos definido en torno a la densidad medida según el k-ésimo vecino más cercano, en el que, en un proceso iterativo, en cada iteración se elimina de la población elite aquel individuo que está más cerca de otros respecto a los valores de los objetivos. La función de rellenado permite añadir elementos dominados

tanto de la población como de la población elite hasta completar el tamaño de la misma (ordenando los individuos según su valor de fitness).

5.5. Esquema de reproducción y operadores genéticos

Se utiliza el siguiente esquema de reproducción:

- Se une la población original con la población elite y se obtienen los elementos no dominados de la unión de ambas poblaciones.
- Se aplica un esquema de selección por torneo binario sobre los individuos no dominados.
- A la población resultante, se le aplica recombinación a través del operador de cruce en dos puntos y un operador de mutación uniforme sesgado con el que la mitad de las mutaciones realizadas tienen el efecto de eliminar la variable correspondiente, para incrementar la generalidad de las reglas.

6. Experimentación

Para analizar el comportamiento de la propuesta multiobjetivo sobre el problema de marketing planteado, se ha ejecutado sobre el mismo el algoritmo evolutivo de inducción de reglas de descripción de subgrupos AGI desarrollado por los autores [7], cuyas características generales se describen a continuación:

- Es un modelo iterativo que incluye un AG híbrido para la extracción de una regla de descripción de subgrupos de mismo tipo que las descritas en este trabajo.
- El proceso iterativo permite obtener nuevas reglas mientras las reglas generadas alcancen un nivel de confianza mínima y aporten información sobre áreas del espacio de búsqueda en que queden ejemplos sin describir por las reglas previamente generadas.
- El AG utiliza el esquema de codificación y operadores genéticos de la propuesta multiobjetivo.
- La función fitness es una combinación lineal ponderada de la confianza, completitud e interés, calculados con las expresiones indicadas en la sección 5.2
- A la regla obtenida por el AG se le aplica, en una etapa de post-procesamiento, un algoritmo proceso de optimización local en el que se van eliminando variables de la regla mientras se consigan reglas más generales (aumente la completitud) y se mantenga el nivel de confianza.

La experimentación, para ambos algoritmos, se ha realizado con 5 ejecuciones para cada una de las 3 clases del atributo objetivo (*eficacia baja*, *media* y *alta*), y con los siguientes parámetros comunes:

- Tamaño de la población: 100
- Número de evaluaciones: 10.000
- Probabilidad de cruce: 0,6
- Probabilidad de mutación: 0,01
- Etiquetas lingüísticas para las variables continuas: 3

Además, en el caso del AG multiobjetivo, el tamaño de la población elite es 25. El algoritmo AGI necesita un valor de confianza mínima bajo el cual no van a evolucionar reglas, que en esta experimentación se ha fijado al valor 0,6. En el AG multiobjetivo, la solución final estará formada por todas las soluciones del conjunto de soluciones no dominadas que superen el mismo umbral de confianza.

En las Tablas 1, 2 y 3 se muestran los mejores resultados obtenidos con ambos algoritmos para todas las clases de la variable objetivo (*eficacia baja*, *media* y *alta*). En ellas se muestran el número de variables que intervienen en cada regla (columna etiquetada como N_V) y los valores para cada uno de los objetivos considerados (*Comp*. para la completitud, *Conf*. para la confianza, *Int*. para el interés y *S.O*. para el soporte original).

AG multiobjetivo					AGI			
N_V	*Comp*	*Conf*	*Int*	*S.O.*	N_V	*Comp*	*Conf*	*Int*
9	5,26	100,00	0,58	0,05	3	5,26	100,00	0,59
11	15,79	66,67	0,58	0,18	4	2,63	100,00	0,58
7	42,10	61,54	0,57	0,60	5	2,63	100,00	0,56
6	44,74	64,92	0,54	0,85				
11	21,05	100,00	0,56	0,24				
10	18,42	87,50	0,56	0,42				
8	5,26	76,92	0,59	0,05				
9	36,84	73,68	0,57	0,49				
11	34,21	87,16	0,56	0,41				
7	23,68	60,00	0,58	0,52				

Tabla 1. Resultados para *eficacia baja*.

Como se puede observar el AG multiobjetivo permite obtener conjuntos de reglas con mayor cardinalidad (mayor número de reglas) que el algoritmo AGI. Estas reglas, con valores adecuados de confianza, completitud e interés, describen más información sobre los tres subgrupos (*eficacia baja*, *media* y *alta*). Esto es debido a que el enfoque multiobjetivo nos permite obtener un conjunto de soluciones adecuadas según los distintos objetivos.

AG multiobjetivo					AGI			
N_V	*Comp*	*Conf*	*Int*	*S.O.*	N_V	*Comp*	*Conf*	*Int*
1	95,27	65,58	0,20	6,78	4	0,68	100,00	0,61
5	1,35	100,00	0,60	0,07	7	2,70	66,67	0,62
2	79,05	68,02	0,12	5,27	2	2,03	100,00	0,57
3	5,40	72,73	0,61	0,33	4	4,05	100,00	0,57
3	40,54	77,92	0,38	2,26	3	3,38	100,00	0,51
3	20,27	68,18	0,59	1,46	5	8,11	100,00	0,55
5	32,43	87,27	0,35	1,60	2	45,95	69,39	0,62
2	87,16	67,55	0,33	5,92				
4	64,86	71,57	0,32	3,81				
2	93,92	65,57	0,41	6,58				
3	54,73	72,73	0,39	2,98				
3	82,43	67,78	0,55	5,39				
3	35,81	82,81	0,19	1,89				
5	11,49	100,00	0,57	0,60				
2	90,54	66,34	0,60	6,35				
3	71,62	70,54	0,23	4,46				
1	98,65	64,89	0,62	7,08				
1	61,49	68,94	0,53	3,48				
5	44,59	72,53	0,58	2,07				
3	86,49	64,97	0,61	5,89				
4	23,65	83,33	0,51	0,97				
3	76,35	68,27	0,53	4,85				
4	36,49	80,60	0,43	1,91				
3	52,70	73,77	0,52	2,70				
4	50,00	73,27	0,57	2,38				

Tabla 2. Resultados para *eficacia media*.

AG multiobjetivo					AGI			
N_V	*Comp*	*Conf*	*Int*	*S.O.*	N_V	*Comp*	*Conf*	*Int*
5	2,38	100,00	0,61	0,08	4	7,14	75,00	0,58
9	28,57	93,75	0,55	0,37				
7	50,00	69,93	0,55	0,73				
9	23,81	71,43	0,56	0,34				
11	19,05	100,00	0,56	0,22				
11	9,52	76,92	0,56	0,11				
7	26,19	71,43	0,56	0,55				
6	57,14	61,24	0,55	0,91				

Tabla 3. Resultados para *eficacia alta*.

En el AG multiobjetivo la diversidad a nivel fenotípico en la población durante el proceso evolutivo (y por tanto en la solución final) se potencia a través de dos vías:

a) Mediante la inclusión de un nuevo objetivo que considera la aportación original (en cuanto a ejemplos cubiertos) de una regla. Esto permite obtener conjuntos de reglas con mayor soporte e incrementa las posibilidades de obtener un conjunto de reglas que describan información sobre todos los ejemplos, y no solo sobre la mayoría.
b) Mediante un esquema de nichos implementado en el operador de truncado, que, en caso de tener que reducir la población elite, elimina reglas con valores similares para los distintos objetivos.

En el algoritmo AGI la obtención de un conjunto de reglas con suficiente diversidad se potencia mediante la inclusión del AG en un esquema iterativo que continúa mientras las reglas obtenidas describan información sobre nuevos ejemplos (nichos secuenciales a nivel fenotípico), pero la experimentación muestra que para este problema, son mejores los resultados obtenidos por el AG multiobjetivo.

El AG multiobjetivo elimina la compensación entre medidas de calidad y permite obtener conjuntos de reglas con un nivel elevado de confianza, completitud e interés. Es especialmente significativo el alto nivel de completitud obtenido en las distintas reglas con el AG multiobjetivo, incluso para las clases *eficacia alta* y *baja*, difíciles de describir en este problema.

En este aspecto, los resultados muestran que con el algoritmo AGI, se obtienen en ocasiones reglas con un mayor grado de confianza, especialmente para las clases *eficacia baja* y *media* (Tablas 1 y 2). No obstante, en AGI los altos valores en el objetivo relativo a la confianza sesgan la búsqueda y convierten la completitud en un objetivo difícil de alcanzar para las clases *eficacia baja* y *alta*.

Ambas propuestas permiten obtener conjuntos de reglas descriptivas por el uso de etiquetas lingüísticas para las variables continuas y por el

bajo número de variables implicadas (por debajo del 10% de las 104 variables). En este aspecto hay que destacar que las reglas obtenidas por el algoritmo AGI son más sencillas que las generadas por la propuesta multiobjetivo. La aplicación en AGI de un algoritmo de ascensión de colinas que optimiza cada una de las reglas obtenidas permite aumentar la simplicidad de las mismas.

7. Conclusiones

En este trabajo se describe la aplicación de un modelo evolutivo multiobjetivo para la inducción descriptiva de reglas difusas que describen subgrupos a un problema real de extracción de conocimiento en certámenes feriales.

Pese a las características del problema (elevado número de variables y valores perdidos, bajo número de ejemplos y pocas variables continuas) esta aproximación multiobjetivo al problema permite obtener conjuntos de reglas fáciles de interpretar, con un nivel alto de confianza, de soporte y completitud.

Como trabajo futuro nos planteamos el desarrollo de una propuesta para reglas en formato DNF y una medida de interés adecuada para este tipo de reglas.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología y los fondos FEDER bajo los proyectos TIC-2002-04036-C05-01 y TIC-2002-04036-C05-04, y las redes TIN2004-20061-E y TIN2004-21343-E.

Referencias

[1] Agrawal, R; Imielinski, T; Shrikant, R. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 207-216. Washington, D.C., 1993.

[2] Coello, C.A; Van Veldhuizen, D.A; Lamont, G.B. Evolutionary algorithms for solving multi-objective problems. Kluwer Academic Publishers, 2002.

[3] Cordón, O; del Jesus, M.J; Herrera, F. Genetic learning of fuzzy rule-based classification systems co-operating with fuzzy reasoning methods. International Journal of Intelligent Systems, 13 (10/11), pp. 1025-1053. 1998.

[4] Cordón, O; Herrera, F; Hoffmann, F; Magdalena, L. Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases. World Scientific, 2001.

[5] Deb, K. Multiobjective optimization using evolutionary algorithms. Wiley, 2001.

[6] Deb, K; Pratap, A; Agarwal, A; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2), pp. 182-197. 2002.

[7] Del Jesus, M.J; González, P; Herrera, F; Mesonero, M. Evolutionary inducción of descriptive fuzzy rules in a market problem. Proceedings of the First Workshop on Genetic Fuzzy Systems (GFS), pp. 57-63. Granada, 2005.

[8] Fonseca, C.M; Fleming, P.J. Genetic algorithms for multiobjective optimization: formulation, discussion and generalization. Proceedings of the Fifth International Conference on Genetic Algorithms (ICGA), pp. 416-423. San Mateo, CA, 1993.

[9] Flockhart, I.W; Radcliffe, N.J. GA-MINER: Parallel data mining with hierarchical genetic algorithms (Final Report by the University of Edimburgh, UK, EPCC-AIKMS-GA-Miner-Report 1.0), 1995.

[10] Freitas, A.A. On objective measures of rule surprisingness. Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-98). Lecture Notes in Artificial Intelligence, 1510, pp. 1-9. 1998.

[11] Freitas, A.A. Data mining and knowledge discovery with evolutionary algorithms. Springer, 2002.

[12] Gamberger, D; Lavrac, N. Expert-guided subgroup discovery: methodology and application. Artificial Intelligence Research, 17, pp. 501–27. 2002.

[13] Giordana, A; Neri, F. Search-intensive concept induction. Evolutionary Computation, 3 (4), pp. 375-416. 1995.

[14] Goldberg, D.E. Genetic algorithms in search, optimization and machine learning. Addison-Wesley, 1989.

[15] González, A; Pérez, R. SLAVE: a genetic learning system based on an iterative approach, IEEE Trans. Fuzzy Systems, 7(2), pp. 176-191. 1999.

[16] Ghosh, A; Nath, B. Multi-objective rule mining using genetic algorithms. Information Sciences, 163 (1-3), pp. 123-133. 2004.

[17] Hajela, P; Lin, C.Y; Genetic search strategies in multicriterion optimal design. Structural Optimization, 4, pp. 99-107. 1992.

[18] Holland, J.H. Adaptation in natural and artificial systems. University of Michigan Press, 1975.

[19] Horn, J; Nafpliotis, N. Multiobjective optimization using the niched pareto genetic algorithms (IlliGAL Report 93005, University of Illinois, Urbana, Champaign). 1993.

[20] Ishibuchi, H; Murata, T. A multiobjective genetic local search algorithm and its application to flowshop scheduling. IEEE Trans. System, Man and Cybernetics, 28 (3), pp. 392–403. 1998.

[21] Ishibuchi, H; Yamamoto, T. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. Fuzzy Sets and Systems, 141 (1), pp. 59-88. 2004.

[22] Janikow, C.Z. A knowledge-intensive genetic algorithm for supervised learning. Machine Learning, 13, pp. 189-228. 1993.

[23] Jovanoski, V; Lavrac, N. Classification rule learning with APRIORI-C. Proceedings of the Tenth Portuguese Conference on Artificial Intelligence (EPIA), pp. 44–51. Berlin, 2001.

[24] Klösgen, W. Explora: a multipattern and multistrategy discovery assistant. In Fayyad, V; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R (eds.) Advances in Knowledge Discovery and Data Mining, pp. 249-271. MIT Press, 1996.

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

[26] Lavrac, N; Flach, P; Kavsek, B; Todorovski, L. Adapting classification rule induction to subgroup discovery. Proceedings of the Second IEEE International Conference on Data Mining (ICDM), pp. 266–273. Maebashi City, 2002.

[27] Miller, S. Saque el máximo provecho de las ferias. Ediciones Urano, 2003.

[28] Noda, E; Freitas, A.A; Lopes, H.S. Discovering interesting prediction rules with a genetic algorithm. Proceedings of the Congress on Evolutionary Computation (CEC), pp. 1322-1329. Washington D.C., 1999.

[29] Quinlan, J.R. Generating production rules from decision trees. Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI), pp. 304-307. San Mateo, CA, 1987.

[30] Sarker, R; Liang, K.H; Newton, C. A new multiobjective evolutionary algorithm. European Journal of Operational Research, 140, pp. 12-23. 2002.

[31] Silverman, B.W. Density estimation for statistics and data analysis. Chapman and Hall, 1986.

[32] Srinivas, N; Debl, K. Multiobjetive optimization using nondominated sorting in genetic algorithms. Evolutionary Computation, 2, pp. 221-248. 1995.

[33] Wilson, S.W. Classifier fitness based on accuracy. Evolutionary Computation, 3(2), pp. 149-175. 1995.

[34] Wrobel S. An algorithm for multi-relational discovery of subgroups. Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD), pp. 78–87. Berlin, 1997.

[35] Zitzler, E; Thiele, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. IEEE Transactions on Evolutionary Computation, 3(4), pp. 257-217. 1997.

[36] Zitzker, E; Laumanns, M; Thiele. L. SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimisation. In Giannakoglou, K; Tsahalis, D; Periaux, F; Papailiou, K; Fogarty, T (eds.) Evolutionary methods for design, optimisation and control, pp. 95-100. CIMNE, 2002.