Influencia de la granularidad y las medidas de calidad en el modelo evolutivo SDIGA de Descubrimiento de Subgrupos

Pedro González [1] **Cristóbal J. Carmona** [1] **María José del Jesus** [1] **Francisco Herrera** [2]

[1] Depto. Informática, Universidad de Jaén, Jaén, 23071, España, {pglez,ccarmona,mjjesus}@ujaen.es
[2] Depto. CCIA, ETSIIT, Universidad de Granada, Granada, 18071, España, herrera@decsai.ugr.es

Resumen

En este trabajo presentamos un estudio preliminar sobre la influencia de distintos aspectos en el comportamiento del modelo evolutivo SDIGA de extracción de reglas difusas de descripción de subgrupos. El estudio se centra en la influencia de las medidas de calidad utilizadas por el algoritmo, el tipo de regla con el que se representa el conocimiento extraído y la granularidad seleccionada para las variables continuas. Se determina que los mejores resultados se obtienen utilizando como medidas de calidad para SDIGA la confianza difusa, el soporte nítido sobre los ejemplos de la clase y el interés, que el tipo de regla a utilizar se puede dejar a elección del experto, y que la granularidad más adecuada para el conjunto de bases de ejemplos estudiado es 3 ó 5 etiquetas.

Palabras Clave: Descubrimiento de Subgrupos, Minería de Datos, Reglas difusas, Algoritmos Genéticos.

1 INTRODUCCIÓN

La extracción de conocimiento se puede abordar desde dos perspectivas distintas: mediante un proceso de *inducción predictiva*, en el que se intenta obtener conocimiento para clasificación o predicción, o mediante un proceso *inducción descriptiva* cuyo objetivo fundamental es descubrir conocimiento de interés dentro de los datos. El descubrimiento de subgrupos [7][12] es un tipo de inducción descriptiva en el que, dado un conjunto de datos y una propiedad de los mismos que tenga interés para el usuario, encontrar (y describir) subgrupos relevantes respecto a esa propiedad. Un algoritmo de descubrimiento de subgrupos debe extraer reglas que representen de forma simbólica el conocimiento y que sean lo suficientemente sencillas y descriptivas como para ser utilizadas por el usuario final.

Así, la interpretabilidad del conocimiento extraído es uno de los aspectos fundamentales para cualquier algoritmo de descubrimiento de subgrupos. La lógica difusa tiene una gran afinidad con la forma en que representamos el conocimiento humano; esto hace que su uso en reglas con variables cuantitativas facilite la interpretabilidad de las mismas, su diseño, o la incorporación de conocimiento cualitativo sobre el problema [13] [4]. En la literatura especializada existen distintos modelos para la extracción de reglas de descripción de subgrupos [10] [6] [3], pero hasta donde conocemos solo los trabajos de Del Jesus et al. [5] utilizan reglas difusas como forma de representación del conocimiento.

En este estudio preliminar, nuestro objetivo es analizar la influencia de distintos aspectos sobre los resultados de los algoritmos de extracción de reglas difusas de descripción de subgrupos, como el tipo de reglas utilizadas, la granularidad, o las medidas de calidad utilizadas para la evaluación de las reglas obtenidas. Para ello se ha realizado una experimentación sobre distintos conjuntos de datos disponibles en el repositorio UCI[1] [2].

El trabajo se organiza de la siguiente forma: En la Sección 2 se describe el descubrimiento de subgrupos, y en la Sección 3 el algoritmo evolutivo SDIGA de inducción descriptiva de reglas difusas de descripción de subgrupos. El estudio experimental realizado se describe en la sección 4, y en la Sección 5 se muestran las conclusiones obtenidas.

2 DESCUBRIMIENTO DE SUBGRUPOS

Las técnicas de descubrimiento de subgrupos generan modelos basados en reglas cuya finalidad es descriptiva, empleando una perspectiva predictiva para obtenerlos [10]. Las reglas utilizadas en la tarea de descubrimiento

[1] www.ics.uci.edu/~mlearn/MLRepository.html

de subgrupos tienen la forma *Cond → Clase*, donde la propiedad de interés para el descubrimiento de subgrupos es el valor de la *Clase* que aparece en el consecuente de la regla [8], y el antecedente es una conjunción de variables (parejas atributo-valor) seleccionadas entre las variables del conjunto de datos.

El concepto de descubrimiento de subgrupos fue formulado inicialmente por Klösgen [7] y Wrobel [12], como "dado un conjunto de datos y una propiedad de esos datos que sea de interés para el usuario, buscar subgrupos que sean de mayor interés para el usuario". En este sentido, se dice que un subgrupo es interesante cuando tiene una distribución estadística inusual respecto a la propiedad en la que estamos interesados. El objetivo es descubrir propiedades características de subgrupos construyendo reglas individuales sencillas, altamente significativas y con soporte alto.

Uno de los aspectos más importantes a tener en cuenta en cualquier enfoque de inducción de reglas que describan subgrupos es la elección de las medidas de calidad a utilizar, tanto para seleccionar las reglas, como para valorar los resultados obtenidos en el proceso. A continuación se describen distintas medidas de calidad utilizadas en la tarea de descubrimiento de subgrupos.

- *Cobertura* [10]: mide el porcentaje de ejemplos cubiertos por una regla, y se define como:

$$Cob(R_i) = Cob(Cond_i \rightarrow Clase) = p(Cond_i) = \frac{n(Cond_i)}{N} \quad (1)$$

  donde $n(Cond_i)$ es el número de ejemplos que verifican la condición $Cond_i$, N es el número total de ejemplos y R_i denota la i-ésima regla. La cobertura media para un conjunto de reglas (*COB*) se calcula como la media de la cobertura de las reglas individuales.

- *Relevancia* [7]: se calcula en términos de su razón de verosimilitud (que mide la diferencia entre la distribución de probabilidad de la clase en el conjunto de ejemplos de entrenamiento cubiertos por la regla y la distribución de probabilidad de la clase en el conjunto de todos los ejemplos de entrenamiento), a través de la siguiente expresión:

$$Rel(R_i) = 2 \cdot \sum_j n(Clase_j.Cond_i) \cdot \log \frac{n(Clase_j.Cond_i)}{n(Clase_j) \cdot p(Cond_i)} \quad (2)$$

  donde para cada clase *j*, $n(Clase_j.Cond_i)$ es el número de instancias de la clase en el conjunto que cumplen el antecedente de la regla, $n(Clase_j)$ es el número de instancias de la clase, y $p(Cond_i)$ se utiliza como factor normalizador. La relevancia media del conjunto de reglas (*REL*) se calcula como la media de la relevancia de las reglas individuales.

- *Atipicidad* [9]: se mide como la precisión relativa ponderada de una regla (*WRAcc*, *weighted relative accuracy*), con la siguiente expresión:

$$Ati(R_i) = \frac{n(Cond_i)}{N} \cdot \left(\frac{n(Clase.Cond_i)}{n(Cond_i)} - \frac{n(Clase)}{N} \right) \quad (3)$$

  La atipicidad del conjunto de reglas (*ATI*) se mide como la media de la atipicidad de las reglas individuales.

- *Precisión predictiva*: mide la calidad del conjunto de reglas y se define como el porcentaje de instancias correctamente clasificadas. Para un problema binario de clasificación, la precisión se calcula como:

$$PREC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

  PREC mide la precisión del conjunto de reglas tanto sobre los ejemplos positivos como los negativos, mientras que la precisión de una regla (definida como $Prec(Cond \rightarrow Clase) = p(Clase|Cond)$) mide la precisión de una única regla sólo sobre los ejemplos positivos.

- *Soporte (o completitud)* [10]: se define como la frecuencia de ejemplos positivos cubiertos, de la forma:

$$Sop-N(R_i) = \frac{n(Clase.Cond_i)}{N} \quad (5)$$

  donde $n(Clase.Cond_i)$ es el número de ejemplos cubiertos por $Cond_i$ pertenecientes a *Clase*.
  También se suele calcular el soporte como el número de ejemplos de la clase cubiertos por la regla entre el número de ejemplos de la clase:

$$Sop_c\text{–}N(R_i) = \frac{n(Clase.Cond_i)}{n(Clase)} \quad (6)$$

  El soporte global del conjunto de reglas se calcula considerando la tasa de aciertos positivos para la unión de subgrupos:

$$SOP-N = \frac{1}{N} \sum_{Clase_j} n(Clase_j \cdot \bigvee_{Cond_i \rightarrow Clase_j} Cond_i) \quad (7)$$

  La expresión de la sumatoria corresponde al número de ejemplos distintos que verifican al menos una de las reglas del conjunto, por lo que los ejemplos que verifican varias reglas sólo se cuentan una vez.

- *Confianza*: definida como el porcentaje de ejemplos positivos cubiertos; es decir, el número de ejemplos que cumplen el antecedente y el consecuente de la

regla dividido entre el número de ejemplos que cumplen el antecedente [1]:

$$Conf-N(R_i) = \frac{n(Clase.Cond_i)}{n(Cond_i)} \quad (8)$$

La confianza del conjunto de reglas (*CONF-N*) se calcula como la media de la confianza de las reglas individuales.

3 ALGORITMO EVOLUTIVO SDIGA DE INDUCCIÓN DE REGLAS DIFUSAS DE DESCUBRIMIENTO DE SUBGRUPOS

El algoritmo SDIGA (Subgroup Discovery Iterative Genetic Algorithm) es un modelo evolutivo para la extracción de reglas difusas para la tarea de descubrimiento de subgrupos. Este algoritmo está descrito en detalle en [5], presentándose aquí de forma breve sus principales características.

En la tarea de descubrimiento de subgrupos hay un conjunto de variables descriptivas y una sola variable objetivo que describe los subgrupos. Como el objetivo es obtener un conjunto de reglas que describan subgrupos para todos los valores de la variable objetivo, el algoritmo genético (AG) de esta propuesta descubre reglas difusas donde el consecuente está prefijado a uno de los posibles valores de esta variable objetivo.

El modelo puede utilizar dos tipos de reglas: reglas *canónicas* en las que en el antecedente de las reglas sólo hay conjunciones de condiciones formadas por pares variable/valor, y reglas en *forma normal disyuntiva* (*DNF, Disjunctive Normal Form*) que admiten disyunciones en los valores de cada variable. Este tipo de reglas se pueden expresar como:

$$R_1 : \text{SI } (X_1 = LL_1^1 \text{ O } LL_1^3) \text{ Y } (X_7 = LL_7^1) \text{ ENTONCES } \quad Clase_j$$

donde LL_n^l es la etiqueta lingüística *l* de la variable *n.*

Cada solución candidata se codifica de acuerdo con el enfoque *"Cromosoma = Regla"*, representando sólo el antecedente de la regla (ya que todos los individuos de la población están asociados con el mismo valor de la variable objetivo). La información relativa a cada regla está almacenada 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 núcleo de SDIGA es un AG que utiliza una etapa de post-procesamiento basada en una búsqueda local (un procedimiento de ascensión de colinas). El AG híbrido extrae una regla difusa sencilla e interpretable optimizando el soporte y la confianza. La etapa de post-procesamiento se aplica para incrementar la generalidad de las reglas extraídas.

Este AG híbrido se incluye en un proceso iterativo para la extracción de un conjunto de reglas que describen diferentes zonas (no necesariamente disjuntas) del espacio de búsqueda. Esto se consigue marcando las instancias cubiertas por la regla obtenida, de forma que se evita que una regla obtenida después cubra exactamente los mismos ejemplos que otra previamente extraída. De esta forma se obtienen reglas difusas diferentes, aunque no se excluye la posibilidad de que estén solapadas.

La función de evaluación del AG combina, según la siguiente expresión, tres factores: la confianza, el soporte y el grado de interés de la regla:

$$fitness(c) = \frac{\omega_1 \cdot Confianza(c) + \omega_2 \cdot Soporte(c) + \omega_3 \cdot Interés(c)}{\omega_1 + \omega_2 + \omega_3} \quad (9)$$

Estas medidas se definen de la siguiente forma:

- *Confianza*: se puede calcular mediante una expresión nítida, como en (8), o de forma difusa, como 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 la clase a la que pertenezcan) a la misma zona:

$$Conf-D(R_i) = \frac{\sum_{E^k \in E \,/\, E^k \in Class_j} APC(E^k, R_i)}{\sum_{E^k \in E} APC(E^k, R_i)} \quad (10)$$

  donde *APC (Antecedent Part Compatibility)* es el grado de compatibilidad entre un ejemplo y el antecedente de una regla difusa, calculado según la expresión:

$$APC(E^k, R_i) = T(TC(\mu_{LL_1^1}(e_1^k),...,\mu_{LL_1^{l_1}}(e_1^k)),..., TC(\mu_{LL_{n_v}^1}(e_{n_v}^k),...,\mu_{LL_{n_v}^{l_{n_v}}}(e_{n_v}^k))) \quad (11)$$

  donde $LL_{n_v}^{l_{n_v}}$ es la etiqueta lingüística número l_{n_v} de la variable n_v, $\mu_{LL_{n_v}^{l_{n_v}}}(e_{n_v}^k)$ es el grado de pertenencia del valor de la variable n_v para el ejemplo E^s al conjunto difuso correspondiente a la etiqueta lingüística l_{n_v} para esta variable (n_v,), y *T* y *TC* son la t-norma y la t-conorma utilizadas para representar la unión e intersección difusa, respectivamente.
  La confianza difusa del conjunto de reglas (*CONF-D*) se calcula como la media de la confianza difusa de las reglas individuales.
- *Soporte*: se pude utilizar una expresión nítida, como en (5,6), o difusa, teniendo en cuenta el grado de pertenencia de los ejemplos de la clase a la zona determinada por el antecedente:

$$Sop\text{–}D(R_i) = \frac{\sum_{E^k \in E / E^k \in Class_j} APC(E^k, R_i)}{N} \quad (12)$$

donde *N* es el número total de ejemplos.
El soporte difuso del conjunto de reglas (*SOP-D*) se calcula como la media de las reglas individuales.

- *Interés*: el grado de interés se determina en esta propuesta objetivamente mediante el criterio de interés aportado por Noda y otros [11] en un proceso de modelado de dependencias. En la propuesta se utiliza sólo la parte referente al antecedente para el cálculo del interés, puesto que el consecuente está prefijado, según la expresión:

$$Int(R_i) = 1 - \left(\frac{\sum_{i=1}^{n} Ganancia(A_i)}{n \cdot \log_2 \left(|dom(G_k)| \right)} \right) \quad (13)$$

donde *Ganancia* es la ganancia de información, n es el número de variables que aparecen en el antecedente de la regla y |*dom*(*Gk*)| es la cardinalidad de la variable objetivo (el número de valores posibles para la variable considerada como clase).

El objetivo global de la función de evaluación es orientar la búsqueda hacia reglas que maximicen la precisión y la medida de interés, minimizando el número de ejemplos negativos y no cubiertos.

El AG utiliza un modelo de reproducción de estado estacionario modificado que intenta obtener un equilibrio entre convergencia y diversidad. La recombinación se lleva a cabo mediante un operador de cruce en dos puntos y un operador de mutación aleatorio sesgado que potencia la diversidad de la población.

4 ESTUDIO EXPERIMENTAL

En esta sección se describe la metodología utilizada para el desarrollo de la experimentación de este estudio. Se muestran y analizan los resultados obtenidos para cada uno de los aspectos estudiados.

4.1. Características de la experimentación

Para analizar el comportamiento de SDIGA, se va a realizar una experimentación con distintas bases de ejemplos sintéticas disponibles en el repositorio UCI [2]. Las características principales de estos conjuntos de datos se muestran en la Tabla 1, en la que para cada conjunto de datos, N_{Var} representa el número de variables, $N_{Var\text{-}D}$ el número de variables discretas, $N_{Var\text{-}C}$ el número de variables continuas, N_{Cl} el número de clases, y N_{Ej} el número de ejemplos del conjunto.

Respecto a los aspectos a analizar, en primer lugar trataremos las medidas de calidad utilizadas en la función de evaluación, un aspecto muy importante en los algoritmos de descubrimiento de subgrupos. Distintos autores han utilizado distintas medidas objetivas de calidad, pero no hay un consenso sobre cuál es la mejor medida a utilizar.
Un segundo aspecto a estudiar es la influencia del tipo de reglas difusas a utilizar para representar el conocimiento.

Por último, la calidad de los resultados obtenidos por el algoritmo depende de la idoneidad de la partición difusa utilizada, tanto en tipo de conjuntos difusos, como en número, como en definición de los mismos, y en el caso de una partición uniforme (por ausencia de conocimiento experto), de la granularidad considerada.

Tabla 1: Propiedades de los conjuntos de datos del repositorio UCI utilizados

NOMBRE	N_{Var}	$N_{Var\text{-}D}$	$N_{Var\text{-}C}$	N_{Cl}	N_{Ej}
Australian	14	8	6	2	690
Breast-w	9	9	0	2	699
Bridges	7	4	3	2	102
Diabetes	8	0	8	2	768
Echo	6	1	5	2	131
German	20	13	7	2	1000
Heart	13	6	7	2	270
Hepatitis	19	13	6	2	155
Hypothyroid	25	18	7	2	3163
Ionosphere	34	0	34	2	351
Iris	4	0	4	3	150
Tic-tac-toe	9	9	0	2	958
Vote	16	16	0	2	435
Balance	4	0	4	3	625
Car	6	6	0	4	1728
Glass	9	0	9	6	214
Wine	13	0	13	3	178

Por lo tanto, se va a analizar el comportamiento del algoritmo utilizando distintos tipos de reglas, distintas medidas de calidad y distintas granularidades en las variables continuas. De esta forma, se ha realizado la experimentación para:

- 3 versiones de SDIGA: un esquema nítido (SDIGA-N), que utiliza las medidas de confianza nítida (*Conf-N*, como se define en (8)), soporte nítido (*Sop-N*, como se define en (5)) e interés (*Int*, como se define en (13)); un esquema difuso (SDIGA-D), que utiliza las medidas de confianza difusa (*Conf-D* (10)), soporte difuso (*Sop-D* (12)) e interés (*Int* (13)); y un esquema híbrido (SDIGA-H), que utiliza las medidas de confianza difusa (*Conf-D* (10)), soporte nítido sobre los ejemplos de la clase (Sop_c-*N* (6) e interés (*Int* (13)).
- 2 tipos de modelos de representación del conocimiento; es decir, 2 tipos de reglas: DNF y no DNF (o canónicas).
- Distinto número de etiquetas para las variables continuas: 3, 5 y 7 etiquetas.

Los experimentos se han llevado a cabo mediante validación cruzada con 10 particiones. Para cada base de ejemplos, y para cada algoritmo, por ser no determinista, se han realizado 5 ejecuciones y se muestran las medias de los resultados de estas 5 ejecuciones. Para estas experimentaciones, se ha utilizado un tamaño de la población de 100, 10.000 evaluaciones en cada ejecución del AG: 10.000, una probabilidad de mutación de 0,01, y como pesos para la función de adaptación: 0,4 para el soporte, 0,3 para la confianza y 0,1 para el interés.

Para analizar los resultados obtenidos, la Tabla 2 detalla los resultados promedios de los 17 conjuntos de datos estudiados. Esta tabla refleja las medias para las 5 ejecuciones de cada base de ejemplos, y muestra los resultados de las distintas medidas de calidad incluidas en la función de adaptación y del resto de medidas consideradas en la bibliografía especializada, cuyo contenido es el siguiente:

- El tipo de regla extraída (canónica o DNF).
- La versión del algoritmo SDIGA utilizado para la generación de reglas.
- El número de etiquetas para las variables continuas (N_{Et}), el número medio de reglas obtenidas (N_{Reg}) y el número medio de variables por regla (N_{Var}).
- Los valores medios de las medidas de calidad para los conjuntos de reglas: cobertura (*COB*), relevancia (*REL*), atipicidad (*ATIP*), precisión (*PREC*), soporte nítido sobre ejemplos de la clase (SOP_c-*N*), soporte nítido (*SOP-N*), soporte difuso (*SOP-D*), confianza difusa (*CNF-D*) y confianza nítida (*CNF-N*).

4.2. Análisis de resultados

La Tabla 2 muestra que para el conjunto de bases de ejemplos estudiado, los mejores resultados se obtienen en general (tanto para reglas canónicas como reglas DNF) con las medidas de calidad utilizadas en el esquema híbrido (SDIGA-H): confianza difusa (Conf-D), soporte nítido sobre los ejemplos de la clase (Sop_c-*N*) e interés (*Int*). De hecho, con el esquema evolutivo que utiliza esas tres medidas de calidad, no sólo se obtienen conjuntos de reglas con mejores valores en ellas, sino también para el resto de medidas.

Se observa que la versión SDIGA-N (con medidas nítidas) obtiene peores resultados y además genera siempre un mayor número de reglas. La utilización de las medidas de calidad de soporte y confianza difusas (SDIGA-D) aporta mejores resultados sólo en algunas bases de ejemplos con todas sus variables continuas y con variables objetivo con más de dos clases. No obstante, nuestra propuesta tiene como objetivo obtener reglas difusas, nítidas o mixtas en función del tipo de variables de cada problema, por lo que la combinación híbrida (soporte nítido sobre ejemplos de la clase y confianza difusa) es la más adecuada, obteniendo buenos resultados.

El análisis realizado sobre distintas medidas de calidad determina, bajo nuestro punto de vista, que en un proceso de descubrimiento de subgrupos es más adecuado el uso de medidas de:

- Exactitud en la descripción, como la confianza (en sus distintas definiciones) y la relevancia.

Tabla 2: Resultados por medidas de calidad

Regla	Versión	N_{Et}	N_{Reg}	N_{Var}	COB	REL	ATIP	PREC	SOP_c-N	SOP-N	SOP-D	CNF-D	CNF-N
Canónica	SDIGA-N	3	9,25	3,35	0,122	3,082	0,028	0,532	0,292	0,623	0,092	0,573	0,525
		5	10,48	3,28	0,120	2,614	0,022	0,516	0,241	0,610	0,089	0,529	0,473
		7	8,69	3,03	0,115	2,542	0,021	0,505	0,223	0,613	0,084	0,518	0,470
	SDIGA-D	3	5,66	3,12	**0,278**	4,581	**0,050**	0,550	0,455	0,720	**0,199**	0,527	0,489
		5	6,79	3,01	0,240	4,138	0,045	0,490	0,372	0,760	0,177	0,451	0,424
		7	6,99	**2,95**	0,243	4,050	0,041	0,477	0,371	**0,762**	0,172	0,436	0,413
	SDIGA-H	3	**5,29**	3,54	0,231	**4,958**	0,049	**0,601**	**0,530**	0,580	0,175	**0,627**	**0,583**
		5	6,57	3,23	0,237	4,803	0,045	0,563	0,438	0,647	0,182	0,547	0,512
		7	6,71	3,13	0,245	4,303	0,041	0,537	0,414	0,680	0,187	0,519	0,484
DNF	SDIGA-N	3	13,94	3,51	0,191	3,337	0,027	0,488	0,362	0,779	0,116	0,527	0,508
		5	14,36	4,14	0,191	3,610	0,029	0,549	0,367	0,735	0,121	0,594	0,558
		7	17,38	4,31	0,191	3,633	0,028	0,557	0,365	0,751	0,113	0,605	0,554
	SDIGA-D	3	5,83	**3,08**	0,231	5,993	0,046	0,416	0,407	0,837	0,138	0,399	0,380
		5	6,05	3,34	0,228	6,219	0,051	0,445	0,410	**0,841**	0,140	0,431	0,408
		7	6,04	3,51	0,243	6,377	0,048	0,461	0,437	0,840	0,144	0,448	0,421
	SDIGA-H	3	5,67	4,52	**0,253**	**7,275**	**0,055**	**0,611**	**0,667**	0,584	**0,188**	**0,622**	**0,601**
		5	5,71	5,09	0,236	7,242	0,053	0,601	0,616	0,560	0,170	0,613	0,591
		7	**5,62**	5,19	0,211	6,802	0,042	0,572	0,573	0,557	0,150	0,600	0,559

Tabla 3: Resultados por tipo de regla y granularidad para SDIGA-H

Tipo regla	N_{Et}	N_{Reg}	N_{Var}	COB	REL	ATIP	PREC	SOP_c-N	SOP-N	SOP-D	CNF-D	CNF-N
Canónica	3 etiq	**5,29**	3,54	0,231	4,958	0,049	0,601	0,530	0,580	0,175	**0,627**	0,583
	5 etiq	6,57	3,23	0,237	4,803	0,045	0,563	0,438	0,647	0,182	0,547	0,512
	7 etiq	6,71	**3,13**	0,245	4,303	0,041	0,537	0,414	**0,680**	0,187	0,519	0,484
DNF	3 etiq	5,67	4,52	**0,253**	**7,275**	**0,055**	**0,611**	**0,667**	0,584	**0,188**	0,622	**0,601**
	5 etiq	5,71	5,09	0,236	7,242	0,053	0,601	0,616	0,560	0,170	0,613	0,591
	7 etiq	5,62	5,19	0,211	6,802	0,042	0,572	0,573	0,557	0,150	0,600	0,559

- Generalidad de la descripción, pero considerando ejemplos bien descritos (ejemplos positivos), como el soporte.
- Novedad en la descripción, como el interés o la atipicidad.

Una vez determinada la mejor combinación de medidas, para el modelo SDIGA, la Tabla 3 muestra los resultados obtenidos solamente por las experimentaciones realizadas sobre el esquema híbrido SDIGA-H, para analizar el comportamiento del algoritmo en función del tipo de regla utilizada para expresar el conocimiento extraído. En esta tabla se puede observar que los resultados medios muestran una ligera superioridad de las reglas DNF respecto a las canónicas.

Por último, respecto al análisis de la granularidad de las variables continuas, se puede observar que se obtienen mejores resultados con 3 y 5 etiquetas. Cuando se consideran 7 etiquetas, este parece un nivel elevado de granularidad. La elección entre 3 y 5 puede depender de las características concretas del problema a resolver.

5 CONCLUSIONES

En este trabajo, se presenta un estudio sobre la influencia de distintos aspectos en los resultados obtenidos por el algoritmo SDIGA de inducción reglas difusas de descripción de subgrupos. Se trata de un estudio preliminar en el que se han utilizado distintas bases de ejemplos sintéticas habitualmente utilizadas en el análisis de los algoritmos de descubrimiento de subgrupos.

Este estudio ha determinado que la versión de SDIGA que utiliza como medidas de calidad el soporte nítido sobre ejemplos de la clase, la confianza difusa y el interés es la más adecuada. Por otro lado, las reglas DNF obtienen resultados ligeramente mejores que las reglas canónicas y ofrecen una estructura más flexible que permite expresar el conocimiento extraído de forma más descriptiva.

Como trabajos futuros, pretendemos ampliar este estudio con la inclusión en la función de evaluación de SDIGA de otras medidas de calidad como la relevancia o la atipicidad. Por otro lado, se pretende desarrollar una versión multiobjetivo del algoritmo y ampliar este estudio con sus resultados.

Agradecimientos

Este trabajo ha sido financiado por el MCYT a través de los proyectos TIN2005-08386-C05-01 y TIN2005-08386-C05-03.

Referencias

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. y Verkamo, I. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, Fayyad, U., et al., Editors. AAAI Press: Menlo Park. Pág. 307–328, 1996.

[2] Asuncion, A. y Newman, D.J. UCI machine learning repository (http://www.ics.uci.edu/~mlearn/MLRepository.html). Univ. California, Irvine, School of Information and Computer Sciences, 2007.

[3] Atzmueller, M. y Puppe, F. SD-Map - A fast algorithm for exhaustive subgroup discovery. *10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*. Springer, 2006.

[4] Au, W.H. y Chan, K.C.C. An effective algorithm for discovering fuzzy rules in relational databases. *IEEE International Conference on Fuzzy Systems (Fuzz IEEE'98)*. Pág. 1314-1319, 1998.

[5] del Jesus, M.J., González, P., Herrera, F. y Mesonero, M. Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A Case Study in Marketing. *IEEE Transactions on Fuzzy Systems*, 15(4). Pág. 578-592, 2007.

[6] Kavsek, B. y Lavrac, N., APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20. Pág. 543–583, 2006.

[7] Klösgen, W. Explora: A multipattern and multistrategy discovery assistant. *Advances in Knowledge Discovery and Data Mining*. Pág. 249-271, 1996.

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

[9] Lavrac, N., Flach, P. y Zupan, B. Rule evaluation measures: A unifying view. *Inductive Logic Programming*. Pág. 174-185, 1999.

[10] Lavrac, N., Kavsek, B., Flach, P. y Todorovski, L. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5. Pág 153-188, 2004.

[11] Noda, E., Freitas, A.A. y, Lopes, H.S., Discovering Interesting Prediction Rules with a Genetic Algorithm. *Congress on Evolutionary Computation*. Pág. 1322-1329, 1999.

[12] Wrobel, S. An algorithm for multi-relational discovery of subgroups. In Proc. of the First European Conference on Principles of Data Mining and Knowledge Discovery, pp. 78-87, 1997.

[13] Zadeh, L.A. The concept of a linguistic variable and its applications to approximate reasoning, Parts I, II, III. *Information Sciences*, 8-9. Pág. 199-249, 301-357, 43-80, 1975.