Extracción Evolutiva de Reglas de Asociación en un Servicio de Urgencias Psiquiátricas

J. J. Aguilera, M.J. Del Jesus, P. González[1], F. Herrera[2], M. Navío, J. Sáinz[3]

En este trabajo se presenta una propuesta evolutiva para un proceso de extracción de reglas de asociación en un servicio de urgencias psiquiátricas. Por las características del problema, una base de datos en la que la mayoría de las variables son discretas, con ausencia de gran cantidad de valores en la mayoría de ellas, la ejecución del algoritmo determina un conjunto de reglas muy precisas, con un alto grado de confidencia, pero poco generales, es decir, con poca completitud. Un análisis de los resultados obtenidos nos permite determinar cuatro líneas de trabajo futuro en el contexto de los algoritmos evolutivos: la utilización de un modelo iterativo con penalización para extraer un conjunto de reglas con cardinalidad variable; el desarrollo de un algoritmo genético multiobjetivo que refleje de una forma más adecuada los criterios de calidad de las reglas; el diseño de un algoritmo memético con un proceso de búsqueda local que incremente la generalidad de las reglas; y la modificación de las propuestas anteriores con un esquema de codificación que permita la obtención de reglas DNF.

Palabras Clave: Minería de Datos, Reglas de Asociación, Algoritmos Genéticos, Modelos de Nichos, Algoritmos Genéticos Multiobjetivo, Algoritmos Meméticos.

I. INTRODUCCIÓN

La minería de datos consiste en la extracción automática de conocimiento de alto nivel de un conjunto de datos reales [6]. Se incluye dentro de un área más amplia, el descubrimiento de conocimiento, en la que intervienen métodos de pre-procesamiento que facilitan la aplicación del algoritmo de minería de datos y métodos de post-procesamiento que refinan y mejoran el conocimiento extraído.

El pre-procesamiento de datos incluye entre otras las siguientes tareas [21]: integración de datos, cuando los datos provienen de distintas fuentes; limpieza de datos, que detecta y corrige errores y valores perdidos; discretización, que prepara los datos para algoritmos incapaces de trabajar con datos continuos; y selección de atributos, que en ocasiones está integrada en el propio algoritmo de minería de datos.

La etapa de post-procesamiento tiene como objetivo el incremento de comprensibilidad e interés del conocimiento extraído.

Dentro de los procesos de minería de datos en la bibliografía especializada [6][11] se especifican distintos conjuntos de tareas, consideradas como tipos particulares de problemas resueltos por algoritmos de minería de datos:

- *Clasificación*, cuyo objetivo es predecir el valor para un atributo objetivo especificado por el usuario en base a valores de otros atributos predictivos.
- *Modelado de dependencias*, que se puede considerar una generalización de la tarea de clasificación ya que intenta predecir el valor de varios atributos.
- *Agrupamiento*, una forma de aprendizaje no supervisado en la que el algoritmo de minería de datos debe determinar las clases dividiendo el conjunto de ejemplos en grupos.
- *Descubrimiento de reglas de asociación*, en la que se obtiene conocimiento interesante para los usuarios en forma de reglas de asociación que reflejan relaciones entre los atributos presentes en los datos.

Los Algoritmos Genéticos (AGs) [16][13] son técnicas de búsqueda con operaciones basadas en la genética natural que han mostrado tener capacidad de búsqueda robusta en espacios complejos. Este es el motivo por el que constituyen un enfoque válido para resolver algunos de los problemas mencionados anteriormente que están presentes en los procesos de extracción de conocimiento.

En este trabajo presentamos un algoritmo evolutivo de extracción de reglas difusas de asociación aplicado al proceso de extracción de conocimiento en un servicio de urgencias psiquiátricas. Para ello en la Sección II describimos las características del problema y, en la Sección III se expone la propuesta evolutiva. El análisis de los resultados obtenidos permite determinar líneas de trabajo detalladas en la sección IV. Por último, en la Sección V se describen las conclusiones del estudio realizado.

[1] Departamento de Informática. Escuela Politécnica Superior. Universidad de Jaén. {jaguile,mjjesus,pglez}@ujaen.es

[2] Departamento de Ciencias de la Computación e Inteligencia Artificial. E.T.S.I. Informática. Universidad de Granada. herrera@decsai.ugr.es

[3] Hospital Ramón y Cajal. Madrid. mnavio@hrc.insalud.es

II. Extracción de Conocimiento en el Problema de Urgencias Psiquiátricas

El objetivo del problema de urgencias psiquiátricas presentado en este trabajo es obtener información sobre ritmos horarios en la llegada al servicio de urgencias psiquiátricas del Hospital Ramón y Cajal de Madrid. Para ello se ha recogido información sobre 72 variables de tipo sociodemográfico, antecedentes personales, tratamientos previos, tipo de demanda, diagnóstico recibido e intervención realizada, en una muestra de 925 pacientes. En el Apéndice se muestra una breve descripción de las variables consideradas.

El algoritmo de minería de datos que resuelva este problema debe extraer un conjunto de reglas de asociación que, dado un atributo especial (la variable franja horaria) determinen las características que definen a los pacientes ingresados en dicha franja horaria. Por la importancia de la comprensibilidad de los resultados obtenidos y la existencia de variables continuas se ha elegido como herramienta de representación del conocimiento las reglas difusas [26].

III. Algoritmos Evolutivos de Extracción de Reglas de Asociación

Los AGs son adecuados para el descubrimiento de reglas ya que realizan una búsqueda global que utiliza la interacción entre variables de forma más adecuada que los algoritmos voraces utilizados frecuentemente en minería de datos. En la bibliografía especializada se pueden encontrar múltiples métodos genéticos para el descubrimiento de distintos tipos de reglas de clasificación, tanto desde el enfoque Michigan [15][12] en el que un cromosoma codifica una única regla y la población al completo un conjunto de reglas, como desde el enfoque Pittsburgh [5][18] en el que un conjunto de reglas se codifica en un único individuo.

En el problema de urgencias psiquiátricas el objetivo es extraer información significativa representada como reglas de asociación, ya que el objetivo no es tanto la precisión en la predicción –como ocurre en reglas de clasificación- como la comprensibilidad e interés de la información extraída [10]. Los AGs se han utilizado como herramienta para la extracción de reglas de asociación de distinto tipo que optimizan distintos criterios de precisión, comprensibilidad e interés [20][4][7][24].

Para resolver el problema presentado hemos diseñado una primera propuesta de algoritmo de minería de datos: un AG de extracción de reglas difusas de asociación que intenta optimizar la precisión, generalidad e interés de las reglas difusas de asociación. Los dos primeros objetivos son criterios conocidos y aplicados con frecuencia tanto el campo de las reglas de asociación como en el de clasificación. El criterio de interés es más específico de las reglas de asociación y se puede determinar mediante

- un enfoque subjetivo, que considera el conocimiento del usuario sobre el dominio de aplicación y,
- un enfoque objetivo, que a diferencia del anterior emplea una medida de calidad de reglas independiente del usuario y del dominio de aplicación.

Nuestra propuesta incluye una medida de interés objetiva (descrita a lo largo de esta sección) e incorpora conocimiento del usuario y del dominio en la definición previa del conjunto de términos lingüísticos para las variables continuas.

El AG tiene como objetivo descubrir una regla de asociación para un objetivo prefijado por lo que tendrá que ejecutarse tantas veces como valores distintos tenga el atributo objetivo y obtendrá una regla para cada uno de ellos. Se incorpora además una etapa de post-procesamiento que optimiza la regla obtenida con el objetivo de incrementar la completitud manteniendo el grado de confidencia de la misma.
En la siguiente sección se describe el proceso evolutivo a través de sus componentes.

A. Componentes del Algoritmo Genético

1) Esquema de codificación.

Cada cromosoma codifica sólo el antecedente de la regla ya que el consecuente es fijo durante la ejecución del AG. En un cromosoma de longitud fija se representa, utilizando codificación entera, el antecedente de una regla con longitud variable. Para ello se añade al conjunto de valores válidos de cualquier variable un valor especial indicador de la ausencia de dicha variable en la regla. En [24] se describe una codificación alternativa para representar el antecedente de una regla permitiendo la ausencia de algunas de las variables.

La Figura 1 muestra la codificación de la regla:

SI *la variable_0 toma el valor 2* Y *la variable_2 es alta* Y … Y *la variable_ k toma el valor 1*
ENTONCES clase C

V0	V1	V2		Vk
2	4	2	…	1

Figura 1: Esquema de codificación de una regla

En este ejemplo la variable 1 no aparecería en la regla ya que tiene tres valores posibles (1,2 y 3) y en

el gen correspondiente aparece el valor 4 que indica la ausencia de dicha variable en la regla. La variable 2 es una variable continua tratada como una variable lingüística con cinco posibles valores (muy_bajo, bajo, medio, alto y muy_alto). Los conjuntos difusos correspondientes a los términos lingüísticos vienen definidos por una partición difusa con funciones de pertenencia triangulares como la que se muestra en la Figura 2.

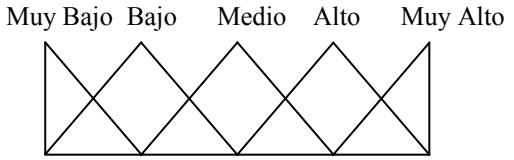

Figura 2: Ejemplo de partición difusa para una variable continua

2) Función de evaluación

La función de evaluación combina, según la siguiente expresión, tres factores: la confidencia, la completitud y el grado de interés de la regla.

$$fitness(c) = \frac{\omega_1 \cdot Completitud(c) + \omega_2 \cdot Interés(c) + \omega_3 \cdot Confidencia(c)}{\omega_1 + \omega_2 + \omega_3}$$

Cada uno de estos criterios se calcula de la siguiente forma:

1. Confidencia.

La confidencia de una regla es un factor que determina la precisión de la misma 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. Para el cálculo de este factor utilizamos una expresión modificada de la definición de precisión aportada por Quinlan en [22] que se utiliza frecuentemente en la generación de reglas difusas de clasificación [2][3][14]: SPAC/SPA, donde SPAC es la suma del grado de pertenencia de los ejemplos de la clase a la zona determinada por el antecedente y SPA representa la suma del grado de pertenencia de todos los ejemplos (independientemente de la clase a la que pertenezcan) a la misma zona. Para calcular estos grados de pertenencia se utilizan funciones de pertenencia triangulares y la t-norma mínimo.

2. Completitud.

La completitud de una regla es una medida del grado de cobertura que la regla ofrece a los ejemplos de la clase. Se calcula como el cociente NECA/NEC, donde NECA es el número de ejemplos de la clase que pertenecen al antecedente y NEC es el número total de ejemplos de la clase. El cómputo de este factor es común para reglas difusas y nítidas, y la expresión mencionada se ha utilizado en la evaluación de reglas de asociación dentro del campo de la medicina [23].

3. Interés.

Como se ha mencionado, el interés de una regla de asociación se puede determinar de forma objetiva (guiada por los datos) o subjetiva (guiada por el usuario). En la bibliografía especializada se pueden encontrar propuestas en ambos sentidos, dependiendo del problema específico al que se aplique el algoritmo de minería de datos y no se puede afirmar nada determinante respecto a las ventajas de uno u otro enfoque. No obstante, parece evidente que en la práctica es adecuado utilizar ambos enfoques: los criterios objetivos como medidas de filtro para seleccionar reglas potencialmente interesantes y los criterios subjetivos para que el usuario final determine reglas realmente interesantes [9].

En nuestra propuesta se sigue este enfoque y en el AG de minería de datos el grado de interés se evalúa de forma objetiva. Para ello utilizamos el criterio de interés proporcionado en [20] en un proceso de modelado de dependencias que considera que el nivel de interés de una regla viene determinado por dos términos, uno referido al antecedente y otro al consecuente, de la siguiente forma:

$$Interés = \frac{Interés\_Antecedente + Interés\_Consecuente}{2}$$

El grado de interés del antecedente se basa en una medida de información y viene dado por la siguiente expresión:

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

Donde *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 término del denominador se introduce para normalizar el valor global.

Tal y como se discute en [9] las variables con un valor alto de ganancia de información son adecuadas para predecir una clase, cuando estas variables se consideran de forma individual. Pero, desde el punto de vista del interés de una regla, se entiende que el usuario ya conoce cuáles son las variables más predictivas para un dominio de aplicación concreto y por tanto las reglas que contienen dichas variables son menos interesantes para el mismo (por ser menos sorpresivas y aportar menos información).

Por eso se entiende que el antecedente de una regla es más interesante si contiene atributos con poca cantidad de información.

El cálculo del grado de interés del consecuente se basa en la idea de que cuanto mayor sea la frecuencia relativa del valor indicado en el consecuente dentro del conjunto de datos, menos interesante es. El objetivo, en este sentido, es obtener reglas con un consecuente no previsible, infrecuente. Para ello, en este trabajo hemos utilizado la siguiente expresión:

$$Interés\_Consecuente = (1 - \Pr(G_{kl}))^{1/\beta}$$

Donde *Pr(Gkl)* es la frecuencia relativa del valor G_{kl} de la variable objetivo (clase) y β es un parámetro especificado por el usuario que permite reducir la influencia del grado de interés del consecuente en el valor del interés global de la regla.

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.

3) Esquema de reproducción

El AG utiliza un modelo de reproducción de estado estacionario modificado [1] que sigue el esquema siguiente:

1. Se genera una población intermedia mediante asignación de probabilidades basada en ordenación lineal y en el esquema de selección de muestreo estocástico universal de Baker.
2. Se aplican los operadores de cruce y mutación a algunos individuos de esta población intermedia. El número de cromosomas a crear vendrá determinado por la probabilidad de cruce y mutación.
3. Los nuevos cromosomas creados sustituirán a los cromosomas peor adaptados de la población original.

De esta forma se sigue la filosofía de la reproducción estacionaria, ya que la población original sólo se modifica mediante la sustitución de los peores individuos por los individuos resultantes de la recombinación. La aplicación de los operadores de cruce y mutación a un porcentaje de cromosomas de una población intermedia, obtenida mediante selección de los individuos mejor adaptados según un esquema de ordenación lineal y muestreo estocástico universal, implica la generación de más de dos cromosomas nuevos introduce más diversidad en la nueva población.

4) Operadores de cruce y mutación

La recombinación se realiza a través del operador de cruce multipunto y un operador de mutación uniforme sesgado con el que la mitad de las mutaciones realizadas tienen el efecto de eliminar la variable correspondiente.

5) Etapa de post-procesamiento del AG: Algoritmo de búsqueda local

A este AG se ha añadido una etapa de post-procesamiento que mejora la regla obtenida mediante un proceso de ascensión de colinas. Este algoritmo de búsqueda local modifica la regla mientras se mantenga el grado de confidencia por encima de un nivel mínimo de confidencia y se incremente el grado de completitud. Para ello, mientras se verifiquen esas dos condiciones, en cada iteración se determina la variable tal que, al eliminarla aumenta en mayor grado la completitud de la regla resultante.

El proceso de búsqueda elimina variables del antecedente de la regla con el objetivo de conseguir reglas más generales, con un mayor nivel de completitud.

B. Experimentación y Análisis de Resultados

La experimentación ha permitido obtener un conjunto de reglas de asociación entre las que destacamos las siguientes:

SI *la edad* ES *alta* Y *no consume opiáceos* Y *bdz=1* Y *no ha tenido consulta previa* Y *existe retraso mental* Y *bdzs=1* Y *el tipo de alta es facultativa*
ENTONCES franja horaria 0
(Confidencia: 0.949; Completitud: 0.014)

SI *la edad es alta* Y *está jubilado* Y *tiene tratamiento psicofarmacológico* Y *la adhesión al tratamiento es buena* Y *tuvo consulta previa en el médico de cabecera* Y *el tipo de alta es facultativa*
ENTONCES franja horaria 1
(Confidencia: 1; Completitud: 0.016)

SI *tiene antecedentes médicos en neurología* Y *no consume alcohol* Y *no consume opiáceos* Y *tiene tratamiento psicofarmacológico* Y *isrs=0* y *no ha tenido otros tratamientos* Y *no tiene trastorno mental orgánico* Y *no tiene retraso mental* Y *no tiene gestos autolíticos* Y *el tipo de intervención que se hizo fue de ajuste*
ENTONCES franja horaria 2
(Confidencia: 1; Completitud: 0.032)

El método de extracción de conocimiento presentado tiene como objetivo determinar una regla de asociación para una clase específica que verifique un alto grado de confidencia, completitud e interés. Esto se refleja en una medida de calidad que combina los tres objetivos a través de pesos.

La primera limitación de este enfoque es la dificultad en la determinación de valores adecuados para el conjunto de pesos. De hecho, en este problema, por la gran cantidad de variables –la mayor parte de ellas discretas-, la estructura de regla utilizada y la abundancia de valores perdidos, los mejores resultados se han conseguido con combinaciones de pesos en las que se da muy poca importancia al interés y se potencia la extracción de reglas con un elevado nivel de confidencia y completitud. Esto es así para evitar que el criterio de interés domine el proceso de búsqueda por la dificultad de encontrar valores elevados para la completitud. Aún en esta situación, las reglas obtenidas alcanzan un nivel adecuado de confidencia, pero representan conocimiento muy específico, válido para un reducido número de ejemplos.

La introducción de un proceso de búsqueda local que generaliza la regla (eliminando variables en el antecedente) siempre que se mantenga la confidencia por encima de un valor mínimo (en este trabajo, 0.9) y de forma independiente al criterio de interés, no consigue incrementar la completitud al nivel deseado.

Es necesario introducir mejoras en el proceso de extracción de conocimiento que permitan:

- Incrementar el grado de generalidad de las reglas (completitud) manteniendo un nivel de confidencia adecuado.
- Obtener un número variable de reglas para describir de forma adecuada cada una de las clases.

En la siguiente sección describimos las líneas de investigación en las que estamos trabajando para alcanzar ambos objetivos.

IV. Líneas de Trabajo

A. Proceso Iterativo de Extracción Evolutiva de un Conjunto de Reglas de Asociación

La propuesta descrita en la Sección III obtiene, para un objetivo prefijado, una regla en cada ejecución. La obtención de más de una regla para cada clase implica la ejecución del AG múltiples veces con distintas semillas. Es difícil, para un algoritmo de extracción de conocimiento, determinar a priori el número de reglas que describen adecuadamente una clase. Una propuesta realista debe considerar la obtención de un número variable de reglas (no determinado por el experto). Por otra parte, la ejecución repetida del AG no asegura la obtención de reglas diferentes ya que el conjunto de ejemplos no se modifica, lo que habitualmente lleva al AG a converger a un óptimo obtenido en ejecuciones anteriores.

Una vía de solución a ambos aspectos es la inclusión del AG de extracción de reglas de asociación en un proceso iterativo que penalice -una vez obtenida una regla- el conjunto de ejemplos representados por la misma para la generación de futuras reglas con el siguiente esquema:

```
INICIO
  Cto_Reglas ← Ø
  REPETIR
    Ejecutar el AG obteniendo la regla R
    CtoReglas ← CtoReglas + R
     Modificar el conjunto de ejemplos
  MIENTRAS consistencia(R) >= consistencia_min
FIN
```

Este proceso obtiene reglas difusas de asociación mientras las reglas generadas mantengan un determinado nivel de consistencia.

La completitud se potencia mediante un método de penalización incluido en la función de evaluación del AG en todas las ejecuciones salvo en la inicial. Con ella se penalizan aquellas reglas que cubren ejemplos pertenecientes a las zonas delimitadas por reglas obtenidas previamente. Es una penalización que no utiliza ninguna función de distancia ya que penaliza diferencias a nivel fenotípico.

Por otra parte, en la evolución del AG los criterios de calidad de cada regla candidata se calculan considerando como ejemplos positivos los correspondientes al conjunto de ejemplos no representados en otras reglas y como ejemplos negativos el conjunto completo. La operación indicada en el esquema como "Modificar el conjunto de ejemplos" elimina del conjunto de ejemplos aquellos cubiertos por la regla extraída. Como se ha mencionado, en el cálculo del número de ejemplos negativos para una regla se considera el conjunto total de ejemplos para evitar la obtención de reglas que aporten información contradictoria con ejemplos ya eliminados.

B. Algoritmo Genético Multiobjetivo

El método de extracción de conocimiento descrito en la Sección III se ha diseñado para determinar reglas con un alto nivel de confidencia, completitud e interés. Éste es un problema multiobjetivo para el que no tiene porqué existir una única solución optimal, sino un conjunto de soluciones no dominadas que forma la frontera del pareto.

Los algoritmos evolutivos que resuelven problemas multiobjetivo se agrupan en torno a dos enfoques:

1. Modelos evolutivos que utilizan pesos para agregar los objetivos
2. Modelos evolutivos que generan poblaciones de soluciones no dominadas.

En la propuesta mencionada, el problema multiobjetivo se aborda mediante una combinación con pesos de los tres criterios de calidad que determina un único punto de equilibrio en la frontera del pareto. El tratamiento conjunto de múltiples objetivos funciona bien en problemas con muchos ejemplos, clases claramente separables y pocas variables, pero tiene dificultades en problemas como el que se presenta. Para este caso sería más adecuado un modelo evolutivo del segundo tipo, en particular un modelo basado en técnicas de pareto, entre los que destacan MOGA [8], NPGA [17], NSGA [25] y SPEA [27], entre otros.

Nosotros seguiremos el enfoque del modelo elitista SPEA [27] en el cual, para cubrir de un modo adecuado el frente del pareto, se manteniene una población externa de soluciones no dominadas encontradas a lo largo del proceso de búsqueda. El fitness de un individuo se determinará a partir de las soluciones almacenadas externamente en base al concepto de dominancia de pareto. Además, el modelo incluye técnicas de clustering para reducir el número de soluciones no dominadas almacenadas de forma que no se destruya la característica de equilibrio de la frontera. Sigue el siguiente esquema:

```
INICIO
  REPETIR
    Generar la población inicial P y el conjunto P' vacío
    Copiar las soluciones no dominadas de P en P'
    Quitar de P' las soluciones no dominadas por otras
    SI |P'| > N'
    ENTONCES
      Reducir P' a tamaño N' mediante clustering
    Calcular el fitness de los individuos de P y P'
    Seleccionar N' individuos a partir de P+P'
    Aplicar cruce y mutación
  MIENTRAS no se alcance el nº máximo de iteraciones
FIN
```

C. *Algoritmo Memético*

En problemas de búsqueda en espacios combinatoriales complejos la hibridación de AGs con una técnica de búsqueda local que, con bajo coste computacional mejore la calidad de los individuos de la población, puede mejorar la evolución de la población al completo. Al algoritmo resultante de esta hibridación se le conoce con el nombre de Algoritmo Memético [19] y se inspira en modelos de adaptación dentro de sistemas naturales que combinan la adaptación evolutiva de poblaciones de individuos con el aprendizaje individual de cada individuo en su tiempo de vida.

En nuestro problema, la inclusión del proceso de búsqueda local descrito, en la evolución representada por el AG puede contribuir a mejorar de forma significativa la completitud de las reglas a obtener. El algoritmo de búsqueda local se aplicará en cada generación a los individuos nuevos que se vayan generando. Los individuos optimizados de esta forma, sustituirán a los cromosomas de los que provienen si son mejores que ellos (si tienen un mayor grado de completitud.

D. *Uso de Reglas DNF*

Las reglas difusas utilizadas representan en el antecedente una conjunción de proposiciones difusas o nítidas individuales. En un problema como el de urgencias psiquiátricas, en el que la mayoría de las variables son discretas y el número de valores perdidos es muy elevado, puede ser adecuado adaptar el proceso para extraer reglas en forma normal disyuntiva, reglas DNF. Para ello se utilizará un esquema de codificación binaria con tantos genes por variable como valores posibles existan para la misma:

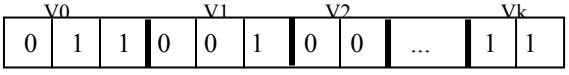

Figura 3: Codificación binaria de una regla DNF

El cromosoma descrito en la Figura 3 representa la siguiente regla:

SI *la variable 0 toma el valor 2 O el valor 3* Y *la variable 1 toma el valor 3* Y …
ENTONCES *clase C*

En dicha regla no intervienen la variable 2 (por no seleccionarse ninguno de sus valores posibles), ni la variable k (por seleccionarse todos los valores y convertirse de esta forma en una variable irrelevante).

V. CONCLUSIONES

En este trabajo se describe un algoritmo de minería de datos con dos componentes: Un AG de extracción de reglas difusas de asociación y un algoritmo de ascensión de colinas que optimiza la regla obtenida.

La aplicación de este algoritmo a un problema de extracción de conocimiento en un servicio de urgencias psiquiátricas complejo por el elevado número de variables implicadas -la mayoría de ellas discretas- y por la presencia de valores perdidos en

gran parte de las mismas, determina un conjunto de reglas con alta confidencia pero baja completitud.

La extensión del algoritmo debe incrementar la completitud de las reglas obtenidas y obtener de forma automática un conjunto de reglas –con cardinalidad variable- que describa de forma adecuada una determinada clase. Para ello, proponemos líneas de trabajo que incluyen: la integración de la propuesta evolutiva en un modelo iterativo para extraer un conjunto de reglas con cardinalidad variable; el desarrollo de un algoritmo genético multiobjetivo que refleje de una forma más adecuada los criterios de calidad de las reglas; el diseño de un algoritmo memético con un proceso de búsqueda local que incremente la generalidad de las reglas; y la modificación de las propuestas anteriores con un esquema de codificación que permita la obtención de reglas DNF.

Nos planteamos como líneas futuras de trabajo el estudio y desarrollo de criterios de calidad de reglas, el análisis de la forma de combinación de los mismos, y el diseño de operadores específicos para extracción de reglas.

VI. 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.

REFERENCIAS

[1] Casillas, J., Cordón, O., del Jesus, M.J., Herrera, F. Genetic Feature Selection in a Fuzzy Rule-Based Classification System Learning Process for High-Dimensional Problems. Information Sciences 136. pp. 135-157. 2001

[2] Chi, Z., Yan, H., Handwritten Numeral Recognition Using Self-organizing Maps and Fuzzy Rules. Pattern Recognition 28 (1). pp. 59-66. 1995.

[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] Dhar, V., Chou, D., Provost, F., Discovering Interesting Patterns for Investment Decision Making with GLOWER-A Genetic Learner Overlaid with Entropy Reduction. Data Mining and Knowledge Discovery 4. pp. 251-280. 2000.

[5] De Jong, K.A., Spears, W.M., Gordon, D.F., Using Genetic Algorithms for Concept Learning. Machine Learning 13. pp. 161-188. 1993.

[6] Fayyad, U.M., Piatetsky-Shapiro, G, Smyth, P., From Data Mining to Knowledge Discovery: An Overview. En: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: *Advances in Knowledge Discovery & Data Mining.* AAAI/MIT. pp. 1-34. 1996.

[7] Fidelis, M.V., Lopes, H.S., Freitas, A.A., Discovering Comprensible Classification Rules with a Genetic Algorithm. Proc. Congress on Evolutionary Computation. pp. 805-810. 2000.

[8] Fonseca, C.M., Fleming, P.J., Genetic Algorithms for Multiobjective Optimization: Formulation, discussion and generalization. Proc. 5th. International Conference on Genetic Algorithms. pp. 416-123. 1993.

[9] Freitas, A.A., On Rule Interestingness Measures. Knowledge-Based Systems 12. pp. 309-315. 1999.

[10] Freitas, A.A., Understanding the Crucial Differences Between Classification and Discovery of Association Rules-A Position Paper. ACM SIGKDD Explorations 2 (1). pp. 65-69. 2000.

[11] Freitas, A.A., A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery. En: Ghosh, A., Tsutsi, S. (eds.): *Advances in Evolutionary Computation*. Springer-Verlag. 2002.

[12] Giordana, A., Neri, F., Search-Intensive Concept Induction, Evolutionary Computation 3 (4). pp. 375-416. 1995.

[13] Goldberg, D.E., *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. 1989.

[14] González, A., Pérez, R., Completeness and Consistency Conditions for Learning Fuzzy Rules. Fuzzy Sets and Systems 96 (1). pp. 37-51. 1998.

[15] Greene, D.P., Smith, S.F., Competition-Based induction of decision models from examples. Machine Learning 13. pp. 229-257. 1993.

[16] Holland, J.H., *Adaptation in Natural and Artificial Systems.* University of Michigan Press. 1975.

[17] Horn, J., Nafpliotis, N., Multiobjective Optimization Using the Niched Pareto Genetic Algorithms. IlliGAL Report 93005, University of Illinois. 1993.

[18] Janickow, C.Z., A Knowledge-intensive Genetic Algorithm for Supervised Learning. Machine Learning 13. pp. 189-228. 1993.

[19] Moscato, P., On Evolution, Search, Optimisation, Genetic Algorithms and Martial Arts: Towards Memetics Algorithms. Technical Report Caltech Concurrent Computation Program, Report 826, California Institute of Technology, Pasadena, California, USA.1989.

[20] Noda, E., Freitas, A.A, Lopes, H.S., Discovering interesting prediction rules with a genetic algorithm. Proc. Congress on Evolutionary Computation. pp. 1322-1329. 1999.

[21] Pyle, D., *Data Preparation for Data Mining. Morgan Kaufmann*. 1999.

[22] Quinlan, J.R., *Generating production rules Machine Learning.* Morgan Kaufmann. 1987.

[23] Richards, G., Rayward-Smith, V.J., Sönksen, P.H., Carey, S., Weng, C., Data Mining for Indicators of Early Mortality in a Database of Clinical Records. Artificial Intelligence in Medicine 22. pp. 215-231. 2001.

[24] Romao,W., Freitas, A.A., Pacheco, R.C.S., A Genetic Algorithm for Discovering Interesting Fuzzy Prediction Rules: applications to science and technology data. Proc. Genetic and Evolutionary Computation Conf. 2002.

[25] Srinivas, N., Debl, K., Multiobjetive Optimization Using Nondominated Sorting in Genetic Algorithms. Evolutionary Computation 2. pp. 221-248. 1995.

[26] Zadeh, L.A., Fuzzy Sets. Information and Control 8. pp. 338-353. 1965.

[27] Zitzler, E., Thiele, L., Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. IEEE Transactions on Evolutionary Computation 3 (4) (1999) 257-271.

APÉNDICE

Nº	*Descripción*	*Nº*	*Descripción*
0	Derivación	36	Inicio clínica
1	Sexo	37	Consulta previa
2	Edad	38	Tiempo consulta
3	Educación	39	T. mental orgánico
4	Laboral	40	T. mental por sustancias
5	Conviven.	41	T. psicótico
6	Motivo consulta	42	T. afectivos
7	Antec. Médicos	43	T. Neuróticos
8	Antec. Psiquiátr.	44	T. Disfunc. Fisiol.
9	Consumo sustan.	45	T. Personalidad
10	Alcohol	46	Retraso mental
11	Cannabis	47	T. del desarrollo
12	Opiáceos	48	T. infantiles
13	Cocaina	49	T. alimentación
14	Otros	50	Gesto autolítico
15	Gestos autolesiones	51	Ef. Secundarios
16	Fumador	52	Psicopatología
17	Tratam. Prev.	53	Tratam. Urgente
18	Tratam. Psicofarma	54	Bdzs
19	Bdz	55	Neurolep. Clas.
20	Neurolep. Clas.	56	Neurolep. Atip.
21	Neurolep. Tric.	57	Antidep. Tric.
22	Antidep. Tric.	58	Isrs
23	Isrs	59	Isrna
24	Isrna	60	Otros antidep.
25	Otros antidep.	61	Litio
26	Litio	62	Eutimizantes
27	Eutimizante	63	Otros ttos.
28	Otros tratam.	64	Neurolep depot.
29	Nl. Depot.	65	Destino alta
30	Psicoterapia	66	Intervención
31	Adhesión	67	Ingreso volunta.
32	Ingr. Psiq. Prev.	68	Tipo alta
33	Ingr. Medicos Prev.	69	Ant. Fam. Psiq.
34	Análisis demand.	70	Grado parentesco
35	Acompañante	71	Ingresos psiquiatr.