Modelos Evolutivos de Extracción de Conocimiento en Aplicaciones Médicas: Enfermedad de Parkinson y Urgencias Psiquiátricas

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

[1]Dpto. Informática. Universidad de Jaén.
{jjaguile,mjjesus,pglez}@ujaen.es
[2]Dpto. Ciencias de la Computación e I.A. Universidad de Granada
herrera@decsai.ugr.es
[3]Hospital Ramón y Cajal, Madrid.
mnavioa@hrc.insalud.es

Resumen. En este trabajo se describen propuestas de algoritmos evolutivos que permiten extraer información relevante sobre dos problemas médicos: el diagnóstico precoz de la enfermedad de Parkinson y la determinación de patrones de llegada en función de la franja horaria en un servicio de urgencias psiquiátricas.

1 Introducción

La minería de datos consiste en la extracción automática de conocimiento de alto nivel de un conjunto de datos reales [11]. Se incluye dentro de un área más amplia, el descubrimiento de conocimiento, en la que intervienen además 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 tareas las siguientes [32]: *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 [11][15] 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) [22][18] 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 en la Sección 2 AGs aplicados a la tarea de clasificación para el estudio de la enfermedad de Parkinson y en la Sección 3 procesos evolutivos para el descubrimiento de reglas de asociación en el problema de urgencias psiquiátricas. Finalmente, en la Sección 4 se exponen las conclusiones.

2 Estudio de la Enfermedad de Parkinson

La enfermedad de Parkinson se diagnostica por la presencia de síntomas de parkinsonismo progresivo: manifestaciones cardinales motoras de bradicinesia o acinesia, temblor en reposo, rigidez y alteración de los reflejos posturales, entre otras.

El estudio de esta enfermedad se ha planteado desde dos perspectivas distintas:

Problema 1: Diagnóstico precoz de la Enfermedad de Parkinson. Actualmente no existe ningún marcador que permita diagnosticar la enfermedad de forma previa a la aparición de los síntomas motores parkinsonianos. Se está estudiando la posibilidad de un diagnóstico en base al análisis de rasgos de personalidad del paciente, pero se desconoce el conjunto de características más relevantes.

Problema 2: Separabilidad entre Parkinson familiar y Parkinson esporádico. Algunos estudios analizan grupos de enfermos con algún familiar afectado con la enfermedad con pacientes en los que la enfermedad se manifiesta de forma esporádica. Se desconoce si se trata de la misma entidad nosológica o existen diferencias entre ambos grupos que permitan considerarlas como entidades distintas. Es importante, para el avance en el tratamiento de esta enfermedad, determinar un conjunto de características que diferencien entre ambas subclases.

Disponemos de una base de datos procedente del departamento de Neurología del Hospital Clínico San Cecilio de Granada. Para el problema 1 se recoge información relativa a 76 rasgos de personalidad en 96 pacientes (enfermos de parkinson o no) y para el problema 2 se almacena información sobre 87 características de 64 pacientes [30]. En el Apéndice 1 se muestra una tabla con la descripción de las características estudiadas.

Para resolver los problemas considerados es necesario diseñar un sistema de clasificación. Existen distintos enfoques y propuestas sobre modelos de aprendizaje de este tipo de sistemas [36] y en este trabajo se utiliza la regla del vecino más cercano [8] que determina la clase de un ejemplo no clasificado en base a la información proporcionada por el ejemplo con valores más próximos en las características observadas.

En la Tabla 1 se muestran los porcentajes de predicción obtenidos con el clasificador 1-NN y la técnica de estimación de error Leaving One Out [36] para ambos problemas. De esta tabla se extraen las siguientes conclusiones:

- Para el problema 1 se obtienen porcentajes de clasificación correcta altos, pero es necesario determinar las variables más significativas para el diagnóstico precoz.
- En el problema 2 no se puede distinguir adecuadamente, con el conjunto completo de variables, entre enfermos con antecedentes familiares de Parkinson y enfermos sin antecedentes. Es necesario seleccionar las variables más relevantes y determinar los pacientes a considerar en el proceso de clasificación.

Tabla 1. Porcentajes de clasificación correcta con 1-NN y todas las variables.

	Nº Variables	% Predicción
Problema 1	76	93.75
Problema 2	87	60.94

En las secciones 2.1, 2.2 y 2.3 describimos las propuestas evolutivas para los procesos de selección de características, de prototipos, y de características y prototipos conjunta. En la sección 2.4 se muestran los resultados obtenidos.

2.1 Algoritmos Evolutivos de Selección de Características

Un aspecto importante en el desarrollo de cualquier sistema de clasificación es la selección del conjunto de variables más informativas para el problema a resolver. Los algoritmos que proporcionan una solución a este problema de optimización con restricciones se denominan algoritmos de selección de características y su objetivo es encontrar un subconjunto del conjunto total de variables que haga posible que el proceso de aprendizaje inductivo genere un sistema de clasificación más sencillo y con mínimo error [29].

Los algoritmos de selección de características se agrupan en torno a dos enfoques:

- *Métodos de filtro* en los que la selección de variables se hace de forma independiente al algoritmo de aprendizaje, con medidas de separabilidad de clases.
- *Métodos de envoltura* en los que se utiliza el resultado del algoritmo de aprendizaje para determinar la calidad del subconjunto evaluado.

Los AGs se han utilizado frecuentemente en el diseño de algoritmos de selección de características en distintos campos [3][29][38] y, en concreto, en el campo del diagnóstico médico [21][24].

Para resolver los problemas planteados en la enfermedad de Parkinson hemos considerado dos AGs de selección de características de envoltura [5][30] que seleccionan conjuntos de variables con máxima predicción en el diagnóstico utilizando la regla de clasificación 1-NN.

2.1.1 Algoritmo Genético de Selección de Características de Envoltura con Codificación Binaria

Este AG utiliza codificación binaria para la selección del mínimo número de características que maximicen la función de adaptación siguiente:

$$f(c) = \frac{clasificación\_correcta}{100} - \alpha \cdot \frac{q}{N}$$

siendo *clasificación_correcta* la estimación del porcentaje de clasificación correcta obtenida mediante el 1-NN con el método Leaving One Out; *q* la cardinalidad del conjunto de variables representado por el cromosoma c, y α el peso asignado a la obtención de subconjuntos de variables con cardinalidad mínima.

El AG utiliza un modelo de reproducción de estado estacionario modificado [5] 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, que aplica los operadores de cruce y mutación a los mejores individuos, puesto que se utilizan los operadores de variación para un porcentaje de cromosomas de una población intermedia que selecciona los individuos mejor adaptados según un esquema de ordenación lineal y muestreo estocástico universal. La generación de más de dos cromosomas nuevos introduce más diversidad en la nueva población manteniendo las características del esquema de reproducción estacionario, ya que la nueva población sólo se diferencia de la anterior en estos cromosomas generados que sustituyen a los peor adaptados.

La recombinación de individuos se consigue con la aplicación de los operadores de cruce multipunto en dos puntos y el operador de mutación aleatoria simple.

2.1.2 Algoritmo Genético de Selección de Características de Envoltura con Codificación Entera

Este AG utiliza un esquema de codificación entera con cromosomas de longitud fija para obtener un subconjunto de características de un tamaño prefijado. De esta forma en un cromosoma de longitud H se codifica un subconjunto candidato de H variables, en el que el i-ésimo gen representa la i-ésima variable seleccionada.

La función que orienta la búsqueda es la estimación de la precisión alcanzable con las variables codificadas en el cromosoma mediante el uso del clasificador 1-NN y el método de estimación de error Leaving One Out.

El AG utiliza el esquema de reproducción descrito en la sección 2.1.1. el operador de cruce parcialmente complementario [29] y la mutación aleatoria simple.

El operador de cruce parcialmente complementario explota el espacio de búsqueda refinando las soluciones obtenidas de la siguiente manera: dados dos cromosomas de la población

$$C_v^t = (c_1, \ldots, c_j)$$
$$C_w^t = (c_1', \ldots, c_j')$$

se generan dos hijos

$$H_1 = (d_1, \ldots, d_k, h_{k+1}, \ldots, h_j)$$
$$H_2 = (d_1, \ldots, d_k, h_{k+1}', \ldots, h_j')$$

donde $d_1, \ldots, d_k$ son los genes comunes a los dos cromosomas seleccionados para ser cruzados y $h_{k+1}, \ldots, h_j$ y $h'_{k+1}, \ldots, h'_j$ son genes seleccionados aleatoriamente entre los no comunes. De esta forma, los hijos generados mantienen las variables comunes a sus padres y combinan de forma aleatoria el resto de la información que contienen. Son individuos válidos sin variables repetidas, por lo que no es necesario ningún proceso de reparación de cromosomas.

El operador de mutación aleatoria simple modifica aleatoriamente uno o más genes de un cromosoma eliminando la variable correspondiente y sustituyéndola por otra no presente en el cromosoma.

2.2 Algoritmos Evolutivos de Selección de Prototipos

En el diseño de un sistema de clasificación es conveniente, en ocasiones, seleccionar los ejemplos para el proceso de aprendizaje inductivo, bien por la elevada dimensionalidad del conjunto completo de ejemplos que hace inabordable el proceso de diseño o bien por la necesidad de determinar los ejemplos o patrones más significativos para el mismo. Los algoritmos que realizan este proceso de búsqueda de un subconjunto de ejemplos dentro del conjunto total se denominan algoritmos de selección de instancias o prototipos.

Entre los métodos de selección de prototipos destacan los métodos de edición del conjunto de referencia [37] que tienen como objetivo incrementar la eficiencia y

eficacia en la predicción del clasificador 1-NN mediante la selección de un conjunto reducido y representativo de prototipos. Este tipo de métodos son especialmente adecuados para el problema que nos ocupa.

En la bibliografía especializada se han utilizado AGs en distintas propuestas de algoritmos de selección de prototipos [26][27][2].

Para el proceso de selección de prototipos en los problemas considerados proponemos un AG que codifica en un cromosoma binario con longitud igual al número total de ejemplos, un subconjunto de prototipos seleccionados.

El AG orienta la búsqueda mediante una función de evaluación que combina, de forma análoga a la función de evaluación del AG binario de selección de características (Sección 2.1.1), el porcentaje de clasificación correcta obtenido con la regla 1-NN (considerando sólo vecinos del conjunto de prototipos que representa el cromosoma) y una penalización por la selección de un número elevado de prototipos.

Utiliza el mismo modelo de reproducción de estado estacionario modificado y operadores de cruce y mutación que el AG binario de selección de características.

2.3 Algoritmos Evolutivos de Selección de Características y Selección de Prototipos

En [16][28][23] se describen algoritmos que aúnan los dos enfoques de reducción de la dimensionalidad en el diseño de un sistema de clasificación, realizando una búsqueda simultánea de conjuntos de características y prototipos.

Para el análisis del problema del Parkinson presentamos una propuesta evolutiva que codifica en un cromosoma binario información a dos niveles [1]: un primer nivel en el que se codifica información relativa a las variables y un segundo nivel con la información correspondiente a los prototipos seleccionados.

La función de evaluación utilizada es una combinación del porcentaje de clasificación correcta alcanzado considerando el prototipo más cercano en base a las variables consideradas, junto con una penalización proporcional al número de variables y de prototipos seleccionados:

$$f(c) = \frac{clasificación \quad correcta}{100} - \alpha \cdot \left[\frac{q}{N} + \frac{p}{NP} \right]$$

Sigue el mismo esquema de reproducción y operadores de cruce y mutación que el AG binario de selección de características y de selección de prototipos.

2.4 Experimentación y Análisis de Resultados

Para estimar la capacidad de predicción de los sistemas de clasificación obtenidos se ha utilizado 10-validación cruzada [36] y cada algoritmo no determinístico se ha ejecutado 15 veces. En las Tablas 2 y 3 se muestra la media aritmética del mejor

resultado obtenido en cada partición, y la media aritmética de los valores medios de cada una de las particiones.

Los resultados se comparan con los proporcionados por el algoritmo de selección de características de filtro LVF [29] que busca subconjuntos de variables de cualquier cardinalidad sin inconsistencias y con los obtenidos por el algoritmo greedy filtro de Battiti [4] que selecciona conjuntos de variables de tamaño prefijado.

En la Tabla 2 se muestran los resultados relativos al problema 1. Como se puede observar el AG con codificación binaria determina un conjunto de variables con capacidad de predicción máxima pero sin reducir al máximo la cardinalidad. La medida de adaptación que orienta el proceso evolutivo se basa en la precisión aportada por la regla 1-NN -insensible a variables redundantes- y esto hace que, a pesar de la penalización incluida en la función de adaptación, el AG propuesto no consiga una reducción más efectiva del número de variables.

Tabla 2. Problema 1. Porcentajes de predicción y número medio de variables

	Media Mejores Resultados		Media Resultados Medios	
Algoritmo	Nº Var.	% Predicción	Nº Var.	% Predicción
LVF	21.4	89.33	21.54	86.11
Greedy	25	95.56	-	-
AG codificación binaria	8.7	100	11.94	96.02
AG codificación entera	3	100	3	92.45

Los mejores resultados se obtienen con los dos AGs frente a los algoritmos LVF y greedy. El AG de codificación entera determina un subconjunto de 3 variables que permite clasificar con un porcentaje de predicción correcta del 100%.

Una vez analizada la capacidad de generalización de los sistemas de clasificación obtenidos con las propuestas evolutivas se ha ejecutado el AG con codificación entera con todos los datos para determinar las variables más significativas de este problema. De esta experimentación se obtiene que los mejores conjuntos de dos y tres elementos están formados por las variables *"control de impulsos"* y *"apertura expresiva"* en el primer caso y *"animación", "control de impulsos"* y *"apertura expresiva"* en el segundo, lo que subraya la relevancia de estos rasgos de personalidad en el diagnóstico de la enfermedad.

En la Tabla 3 se muestran los resultados de los procesos de selección de características y selección de prototipos relativos al problema 2.

El conjunto de variables más predictivo se obtiene con el AG con codificación binaria que determina conjuntos de aproximadamente 40 variables que permiten predecir la clase de ejemplos desconocidos con un 98.63% de acierto.

El AG con codificación entera para selección de características y la propuesta evolutiva de selección de características y prototipos conjunta permiten determinar sistemas de clasificación con porcentajes de predicción superiores al 90%, un valor muy superior al porcentaje que se obtiene con las 87 variables y todos los ejemplos (60.97%, Tabla 1), por lo que se puede considerar que las variables y los prototipos seleccionados tienen una capacidad de predicción adecuada.

Tabla 3. Problema 2. Porcentajes de predicción y número medio de variables

	Media Mejores Resultados			Media Resultados Medios		
Algoritmo	NºVar.	NºProt.	%Predic.	NºVar.	NºProt.	%Predic.
LVF	25.9	64	71	25.8	64	63.03
Greedy	25	64	67.33	-	-	-
AG codificación binaria	39.6	64	98.63	43.87	64	68.9
AG codificación entera	6	64	90.67	6	64	66.69
AG codif. binaria SP	87	50.7	80.00	87	51.91	69.09
AG codif. Binaria SC+SP	28.6	46.6	94.00	27.02	45.75	67.49

Si ejecutamos el AG de selección de características y prototipos con todos los ejemplos para determinar los más significativos en el estudio del problema 2, se obtiene un conjunto de 59 patrones que, considerando 41 variables, permiten clasificar con un porcentaje de acierto del 100%. (En [1] se puede encontrar una descripción más detallada de esta experimentación).

3 Extracción de Conocimiento en el Problema de Urgencias Psiquiátricas

En el problema de urgencias psiquiátricas el objetivo es obtener información sobre ritmos horarios en la afluencia de un servicio de urgencias psiquiátricas. 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 del servicio de urgencias del Hospital Ramón y Cajal de Madrid. En el Apéndice 2 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 que ingresan 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 [39].

En las siguientes secciones se describe el proceso evolutivo presentado para la obtención de reglas de asociación y los resultados obtenidos.

3.1 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 greedy 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 [20][17] como desde el enfoque Pittsburgh [10][25].

En este problema se intenta extraer información significativa representada como reglas de asociación, por lo 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 [14]. 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 [31][9][12][35].

Sobre el problema presentado hemos realizado una primera aproximación desarrollando un AG de extracción de reglas difusas de asociación que intenta optimizar la precisión, generalidad e interés de las reglas difusas de asociación. El criterio de interés 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 contínuas.

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. A continuación describimos el proceso evolutivo a través de sus componentes.

1. Esquema de codificación.

Cada cromosoma codifica sólo el antecedente de la regla ya que el consecuente es fijo durante la ejecución del algoritmo. Se representa con codificación entera de longitud fija el antecedente de una regla con longitud variable añadiendo al conjunto de valores válidos para cualquier variable un valor especial indicador de la ausencia de dicha variable en la regla. En [35] se utiliza codificación entera para el descubrimiento de reglas difusas de asociación pero se codifica la ausencia de la variable mediante un bit adicional.

2. Función de evaluación

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

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

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

- La *confidencia* de una regla es un factor que determina la precisión de la misma ya que indica 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 [33] que se utiliza frecuentemente en la generación de reglas difusas de clasificación [6][7][19]: *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.

- La *completitud* de una regla es una medida del grado de cobertura que la regla ofrece a los ejemplos de la clase y 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 [34].

- Como se ha mencionado, en una regla de asociación el interés puede determinarse 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 potencial interesantes y los criterios subjetivos para que el usuario final determine reglas realmente interesantes [13].

  En nuestra propuesta se sigue este enfoque y en el algoritmo de minería de datos el grado de interés se evalúa de forma objetiva (a diferencia de cómo se realizaba en [35]). Para ello utilizamos el criterio de interés proporcionado en [31] para 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:

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

  Ambos índices se calculan de forma independiente, y mientras la expresión del grado de interés del antecedente se calcula a través de una medida de información, la del consecuente es una medida de frecuencia.

  El grado de interés del antecedente 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 $|Dom(G_k)|$ 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 [13] 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 (por ser menos sorpresivas y aportar menos información) para el mismo. 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 $Pr(G_{kl})$ 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 el esquema de reproducción de estado estacionario modificado, descrito en la sección 2.1.1.

4. Operadores de cruce y mutación

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

A este AG se ha añadido una etapa de post-procesamiento que mejora la regla obtenida mediante un proceso greedy de búsqueda local que modifica la regla manteniendo el grado de confidencia por encima del 90% e incrementando el grado de completitud. El proceso de búsqueda elimina variables del antecedente de la regla con el objetivo de conseguir reglas más generales.

3.2 Experimentación

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

1. 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)

2. 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)

3. 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 se ha diseñado para determinar reglas con confidencia mayor que 0.9, es decir, reglas en las que la información expresada se verifique para el 90% de los ejemplos incluidos en la regla. Además buscamos reglas lo más generales posibles, con el mayor grado de completitud, pero esta meta no se ha alcanzado a un nivel adecuado. La causa de esto es que el algoritmo actual integra en un mismo objetivo diferentes medidas de calidad asignándole pesos. Este 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 el uso de un AG multiobjetivo.

4 Conclusiones

En este trabajo hemos descrito algunas propuestas de algoritmos evolutivos para la resolución de problemas de minería de datos, en concreto, para selección de características, selección de prototipos y extracción de reglas de asociación, y su aplicación a dos problemas médicos.

Para el estudio de la enfermedad de Parkinson, la experimentación realizada subraya la relevancia de dos rasgos de personalidad para el diagnóstico precoz de la enfermedad e indica la posibilidad de diferenciar entre dos subclases dentro de la misma. Además el análisis de los resultados obtenidos para ambos problemas con

otros algoritmos muestra que los conjuntos de variables seleccionados por nuestras propuestas evolutivas tienen una mayor capacidad de predicción.

En el problema de extracción de reglas de asociación para el problema de urgencias psiquiátricas se ha realizado una primera aproximación que nos ha permitido obtener resultados preliminares y determinar la necesidad de desarrollar una propuesta evolutiva que con enfoque multiobjetivo permita la consecución adecuada de las metas fijadas. De igual forma, estamos trabajando en el desarrollo de un algoritmo genético con nichos que permita obtener reglas distintas para un mismo objetivo.

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 desarrollo de operadores específicos para extracción de reglas.

Los problemas de selección de características y prototipos se han afrontado como problemas de aprendizaje de sistemas de clasificación ya que la reducción de la dimensionalidad forma parte del modelo de aprendizaje utilizado, la regla del vecino más cercano. No obstante, las propuestas evolutivas pueden extenderse como procesos de pre-procesamiento para cualquier tarea de minería de datos.

5 Bibliografía

1. Aguilera, J.J., del Jesus, M.J., González, R., Herrera, F., Iríbar, C., Navío, M.: Un Estudio sobre la Aplicación de la Selección Evolutiva de Características y Prototipos al Diagnóstico de la Enfermedad de Parkinson. En: Actas del I Congreso Español de Algoritmos Evolutivos y Bioinspirados (2002) 361-368.
2. Babu, T.R., Murty, M.N.: Comparison of Genetic Algorithms based Prototype Selection Schemes. Pattern Recognition **34** (2001) 523-525.
3. Bal, J., de Jong, K., Huang, J., Vafaie, H., Wechsler, H.: Using Learning to Facilitate the Evolution of Features for Recognizing Visual Concepts. Evolutionary Computation **4**(3) (1997) 297-311.
4. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transactions on Neuronal Networks **5** (4) (1994) 537-550.
5. 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** (2001) 135-157.
6. Chi, Z., Yan, H.: Handwritten Numeral Recognition Using Self-organizing Maps and Fuzzy Rules. Pattern Recognition **28** (1) (1995) 59-66.
7. 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) (1998) 1025-1053.
8. Dasarathy, B.V.: Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press (1999).
9. 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** (2000) 251-280.
10. De Jong, K.A., Spears, W.M., Gordon, D.F.: Using Genetic Algorithms for Concept Learning. Machine Learning **13** (1993) 161-188.
11. 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 (1996) 1-34.
12. Fidelis, M.V., Lopes, H.S., Freitas, A.A.: Discovering Comprensible Classification Rules with a Genetic Algorithm. Proc. Congress on Evolutionary Computation (2000) 805-810.
13. Freitas, A.A.: On Rule Interestingness Measures. Knowledge-Based Systems **12** (1999) 309-315.
14. Freitas, A.A.: Understanding the Crucial Differences Between Classification and Discovery of Association Rules-A Position Paper. ACM SIGKDD Explorations **2** (1) (2000) 65-69.
15. 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).
16. Fu, Z.: Dimensionality Optimization by Heuristic Greedy Learning vs. Genetic Algorithms in Knowledge Discovery and Data Mining. Intelligent Data Analysis **3** (1999) 211-225.
17. Giordana, A., Neri, F.: Search-Intensive Concept Induction, Evolutionary Computation **3** (4) (1995) 375-416.
18. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley (1989).
19. González, A., Pérez, R.: Completeness and Consistency Conditions for Learning Fuzzy Rules. Fuzzy Sets and Systems **96** (1) (1998) 37-51.
20. Greene, D.P., Smith, S.F.: Competition-Based induction of decision models from examples. Machine Learning **13** (1993) 229-257.
21. Handels, H., Rob, Th., Kreusch, J., Wolff, H.H., Pöppl, S.J.: Feature Selection for Skin Tumor Recognition Using Genetic Algorithms. Artificial Intelligence in Medicine **16** (1999) 283-297.
22. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975).
23. Ishibuchi, H. Nakashima, T., Nii, M.: Genetic-Algorithm-Based Instance and Feature Selection. En: Liu, H., Motoda, H. (eds) Instance Selection and Construction for Data Mining, Kluwer Academic (2001) 95-112.
24. Inza, I., Merino, M., Larrañaga, P., Quiroga, J., Sierra, B., Girala, M.: Feature Subset Selection by Genetic Algorithms and Estimation of Distribution Algorithms. A Case Study in the Survival of Cirrhotic Patients Treated with TIPS. Artificial Intelligence in Medicine **23** (2001) 187-205.
25. Janickow, C.Z.: A Knowledge-intensive Genetic Algorithm for Supervised Learning. Machine Learning **13** (1993) 189-228.
26. Kuncheva, L.I.: Fitness Function in Editing k-nn Reference Set by Genetic Algorithms. Pattern Recognition **30** (1997) 1041-1049.
27. Kuncheva L.I., Bezdek, J.C.: On Prototype Selection: Genetic Algorithms or Random Search?. IEEE Transactions on Systems, Man, and Cybernetics **28** (1) (1998) 160-164.
28. Kuncheva, L.I., Jain, L.C.: Nearest Neighbor Classifier: Simultaneous Editing and Feature Selection. Pattern Recognition Letters **20** (1999) 1149-1159.
29. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publisher (1998).
30. Navío, M., Aguilera, J.J., del Jesus, M.J., González, R., Herrera, F., Iríbar, C.: Feature Selection Algorithms Applied to Parkinson's Disease. En: Crespo, J., Maojo, V., Martín, F. (eds.): Medical Data Analysis, LNCS 2199 (2001).
31. Noda, E., Freitas, A.A, Lopes, H.S.: Discovering interesting prediction rules with a genetic algorithm. Proc. Congress on Evolutionary Computation (1999) 1322-1329.
32. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann (1999).
33. Quinlan, J.R.: Generating production rules Machine Learning. Morgan Kaufmann (1987).

34. 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** (2001) 215-231.
35. 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).
36. Weiss, S.M., Kulikowski, C.A.: Computer Systems that Learn. Morgan Kaufmann (1991).
37. Wilson, D.R., Martínez, T.R.: Edition Techniques for Instance Based Learning Algorithms. Machine Learning **38** (2000) 257-286.
38. Yang, J., Honavar, V.: Feature Subset Selection using a Genetic Algorithm. IEEE Intelligent Systems **13** (1998) 44-49.
39. Zadeh, L.A.: Fuzzy Sets. Information and Control **8** (1965) 338-353.

Apéndice 1: Variables Consideradas para el Estudio de la Enfermedad de Parkinson

Nº	Descripción	Nº	Descripción	Nº	Descripción
0	Sexo	29	Extraversión-PC	58	Dominancia-Fi-T
1	Edad	30	Dureza	59	Cooperación
2	Tiempo evoluc.	31	Dureza-PC	60	Cooperación-T
3	Edad comienzo	32	Sinceridad	61	Cordialidad
4	Tabaco	33	Sinceridad-PC	62	Cordialidad-T
5	Alcohol	34	Afabilidad	63	Escrupulosidad
6	Café o te	35	Razonamiento	64	Escrupulosidad-T
7	Agua de pozo	36	Estabilidad	65	Perseverancia
8	Tóxicos	37	Dominancia	66	Perserverancia-T
9	Escolaridad	38	Animación	67	Control emocional
10	Encanecimiento	39	Atención normas	68	Control emocional-T
11	Síntomas emoc.	40	Atrevimiento	69	Control impulsos
12	Antidepresivos	41	Sensiblidad	70	Control impulsos-T
13	Ansiolíticos	42	Vigilancia	71	Apertura cultural
14	Depresión	43	Abstracción	72	Apertura cultural-T
15	Cambios mental.	44	Privacidad	73	Apertura exp.
16	Actividad diaria	45	Aprensión	74	Apertura exp.-T
17	Exploración mot.	46	Apertura Camb.	75	Energía
18	Compl. tratam.	47	Autosuficiencia	76	Energía-T
19	Escala independ.	48	Perfeccionismo	77	Afabilidad-Five
20	Estadio H-Y	49	Tensión sanguin.	78	Afabilidad-Five-T
21	Temblor	50	Extraversión-16	79	Tesón
22	Rigidez	51	Ansiedad	80	Tesón-T
23	Acinesia	52	Dureza-16	81	Estab. emocional
24	Reflejos postur.	53	Independencia	82	Estab. emocional-T
25	M.M.E.	54	Autocontrol	83	Apertura mental
26	Emocionalidad	55	Dinamismo	84	Apertura mental-T
27	Emocionalidad-PC	56	Dinamismo-T	85	Distorsión
28	Extraversión	57	Dominancia-Five	86	Distorsión-T

Apéndice 2: Variables Consideradas para el Estudio de Ritmos Horarios en Urgencias Psiquiátricas

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