UNA METODOLOGÍA DE CONSTRUCCIÓN DE SISTEMAS DE CLASIFICACIÓN BASADOS EN REGLAS DIFUSAS

José María Fernández Garrido José Manuel Benítez Sánchez Ignacio Requena Ramos
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada
garrido@condor.ugr.es jmbs@decsai.ugr.es requena@decsai.ugr.es

Resumen:

En este trabajo, se presenta una metodología para obtener un conjunto de reglas difusas para sistemas de clasificación. El sistema se representa en una red difusa, en el que los antecedentes de las reglas son arcos de entrada a los nodos ocultos, y los consecuentes son arcos de salida. Se utilizan algoritmos genéticos específicos, en dos fases, para extraer las reglas. En la primera fase se extraen las reglas, y en la segunda se refinan las etiquetas. Como ejemplos para un test de la metodología, se aplica a los problemas Iris y Pima.

Palabras clave: Sistemas de clasificación, Sistemas basados en reglas difusas, Algoritmos genéticos.

1.- Introducción

En este trabajo se propone una metodología de construcción de sistemas de clasificación fácilmente interpretables. Para ello, se hace uso fundamentalmente de dos herramientas: los sistemas basados en reglas difusas, que proporcionan modelos fácilmente interpretables, y los algoritmos genéticos como método de búsqueda de soluciones. En el proceso de construcción, se emplea un enfoque de modelado descriptivo, en el que todas las reglas del sistema utilizarán una misma definición de las variables lingüísticas que intervienen en el problema.

Feldman [Feld93] propuso un modelo aplicado a problemas de control, en el que fijaba a priori el número de reglas del SBRD y la definición de las etiquetas lingüísticas, y aplicando un algoritmo genético binario clásico obtenía redes difusas (ver figura 1)

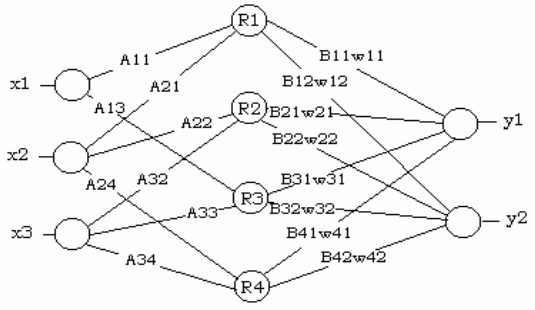


Figura 1: Estructura de la red difusa

En ellas, cada unidad R_i representa el núcleo de una regla difusa del tipo

SI x_1 es A_1 *y* x_2 es A_2 *y* .... *y* x_n es A_n ENTONCES y_1 es B_1 *y* y_2 es B_2 *y* .... *y* y_m es B_m,

donde A_{ij} y B_{jk} son etiquetas lingüísticas y w_{ij} pesos reales asociados a cada término del consecuente.

En este trabajo, tomando como referencia esta idea, se va a diseñar una metodología para construir SBRD aplicados a problemas de clasificación.

Para ello, se va a asignar una salida y_k a cada clase, de forma que al aplicar una muestra al sistema se concluirá que pertenece a la salida con mayor y_k. En el caso de que ante una muestra no se dispare ninguna regla, se empleará un esquema de razonamiento por defecto, para asignar dicha muestra a la clase representada por la salida y_1.

Como método de búsqueda de las bases de reglas se utilizarán algoritmos genéticos con codificación real. Destacaremos la introducción de operadores genéticos que puedan modificar el número inicial de reglas (conjunto de reglas que definen el sistema), para que sea el propio proceso quién determine el número final de reglas que se necesitan. También se estudia la aplicación de un operador de mutación con un número de cambios decreciente conforme avanza el algoritmo.

Se consideran dos tipos de reglas:

a) Reglas difusas clásicas del tipo:
SI x_1 es A_1 *y* x_2 es A_2 *y* .... *y* x_n es A_n
ENTONCES y_1 es B_1 *y* y_2 es B_2 *y* .... *y* y_m es B_m

b) Reglas con grados de certeza en los consecuentes:
SI x_1 es A_1 y ... y x_n es A_n ENTONCES
Y es $Clase_1$ grado r_1, … Y es $Clase_m$ grado r_m

En ambos casos, A_{ij} y B_{jk} son etiquetas lingüísticas, y a cada término del consecuente se le asigna un peso real w_{ij} que da la importancia relativa del término.

La construcción del modelo se hace en dos fases:

Fase 1.- Fijada a priori la definición de las etiquetas lingüísticas, un algoritmo (genético) de aprendizaje va a buscar una base de reglas que obtenga un buen porcentaje de aciertos. Esta fase se puede usar también como un proceso de extracción de características.

Fase 2.- Fijadas las bases de reglas, y usando otro algoritmo genético, se va a refinar la definición de las etiquetas lingüísticas, manteniendo una definición común que utilizarán todas las reglas del sistema.

En lo que sigue se estudian cada una de estas fases, ilustrándolas con los resultados obtenidos al aplicarlas a los problemas del IRIS y del PIMA.

2. Fase 1. Construcción de la base de reglas

La definición de las etiquetas lingüísticas se fija a priori, siendo C_i el número de etiquetas para la variable i. La búsqueda de la base de reglas se realizará con un algoritmo genético, en el que cada cromosoma codifica una base de reglas, y la función de evaluación es el porcentaje de acierto en la clasificación.

Cada regla se codificará como una tupla, en el que los antecedentes A_{ij} tomarán valores en el conjunto $\{0,..,C_i\}$ (el valor 0 indica que la variable no interviene en la regla), los consecuentes B_{jk} en el conjunto $\{1,..,C_j\}$, y los pesos w_{jk} en el intervalo [0,1]. En el caso de reglas con grados de certeza en el consecuente, los grados de certeza r_i tomarán valores en el intervalo [0,1]. La inicialización de la población será aleatoria.

El operador de mutación, actuará sobre una regla del cromosoma al que se aplique, modificando los valores de las variables de entrada y salida según una variable aleatoria uniformemente definida sobre su dominio ($x \sim U([a, b])$. Los pesos se modifican según la ecuación ($u \sim U([-0.1,0.1])$):

$$w_{ij}(t+1) = w_{ij}(t) + u \cdot w_{ij}(t)$$

Los grados de certeza en los consecuentes, se modificarán según la ecuación ($u \sim U([-0.1,0.1])$)

$$r_{ij}(t+1) = r_{ij}(t) + u \cdot r_{ij}(t)$$

El operador de cruce será clásico con dos puntos de cruce, con intercambio de reglas completas.

Para modificar la longitud de los cromosomas, se introducen los operadores *añadir_regla*, que elige aleatoriamente un cromosoma y le añade una regla del mejor cromosoma (aleatoriamente), y *eliminar_regla*, que elimina una regla del cromosoma elegida aleatoriamente según una distribución uniforme.

El algoritmo genético será un modelo elitista, conservando siempre el mejor cromosoma, utilizando como estrategia de selección un modelo de estado estable ("steady state") según el cual los cromosomas generados por *cruce*, *añadir_regla* y *eliminar_regla* reemplazan a los peores de la población y la mutación transforma al cromosoma sobre el que se aplica. Como criterio de parada se utiliza un número fijo de generaciones y el tamaño de la población es fijo.

Resultados experimentales

Para comprobar el comportamiento de este algoritmo, se ha aplicado a los problemas del IRIS y el PIMA. En cada problema, se han generado cinco particiones aleatorias, dividiendo las muestras en un conjunto de entrenamiento con 2/3 de las muestras, y el resto para test. Se han lanzado diez ejecuciones independientes del algoritmo sobre cada partición, realizando un total de 50 ejecuciones por experimento.

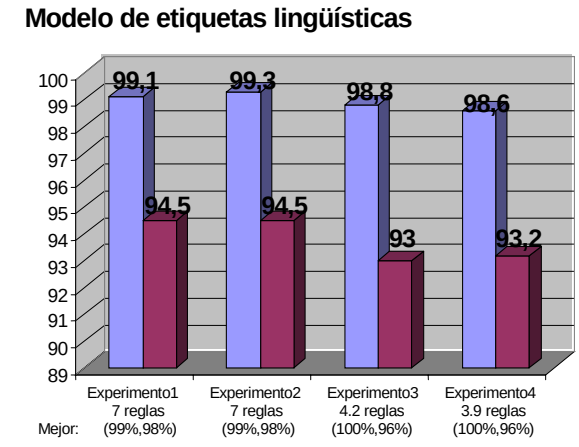
Los experimentos realizados, para cada modelo (etiquetas lingüísticas y grados de certeza) son:

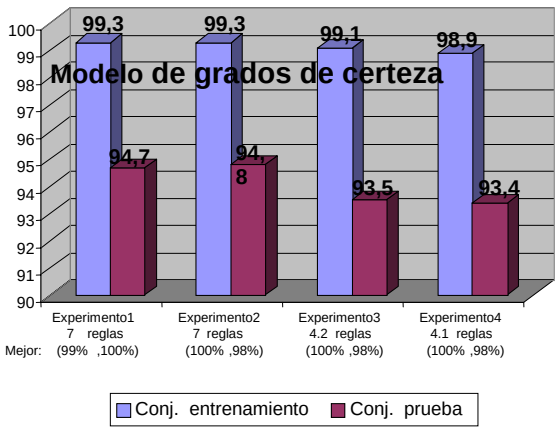
Número de reglas	Nº de cambios en la mutación	
	Fijo	Decreciente
Fijo	Experimento1	Experimento2
Variable	Experimento3	Experimento4

Los parámetros usados han sido: $C_i=C_j=3$ (3 etiquetas por variable) en el Iris y $C_i=C_j=5$ en el Pima; 500 iteraciones; Tamaño de la población 100.

Los resultados obtenidos se pueden ver en las siguientes gráficas, donde en las barras se representa la media del porcentaje de acierto de las 50 ejecuciones realizadas para cada experimento, sobre el conjunto de entrenamiento y sobre el conjunto de prueba. En la leyenda de las gráficas aparece el número medio de reglas utilizadas y los resultados de la mejor ejecución para cada experimento (aprendizaje, test)

A) Para el problema del IRIS



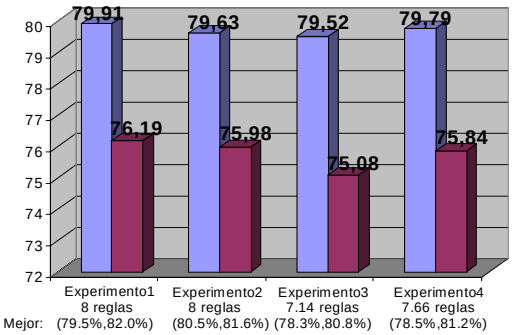


Como se puede observar, se obtienen porcentajes de acierto por encima del 99% sobre el conj. de entrenamiento y en torno al 94.5% sobre el conj. de prueba. Como mejores ejecuciones, se han obtenido varias con sólo 1 error sobre el total de muestras.

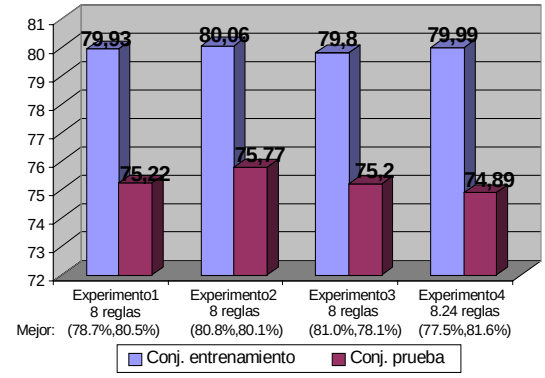
B) Para el problema del PIMA

En el pima, se obtienen resultados cercanos al 80% para el entrenamiento y al 76% para el test.

Modelo de etiquetas lingüísticas



Modelo de grados de certeza



Como mejores ejecuciones se obtienen resultados por encima del 80% tanto sobre el conjunto de entrenamiento como sobre el conjunto de prueba.

3. Fase 2. Optimización de las etiquetas lingüísticas

El objetivo en esta fase es refinar las etiquetas lingüísticas que intervienen en el problema, manteniendo la interpretabilidad de las reglas. El algoritmo usado será también genético, donde cada cromosoma codificará la definición de las etiquetas lingüísticas que usan las bases de reglas. Tendrá como entrada los M mejores conjuntos de reglas obtenidos con el algoritmo anterior, y utilizará como función de evaluación la expresión

$$f_2(C_i) = \max_{j=0}^{M-1} f(R_j, C_i, X)$$

con *f(R,C,X)* el porcentaje de acierto en la clasificación del conjunto de muestras X con el conjunto de reglas R y la definición de las etiquetas lingüísticas C.

Cada cromosoma codifica todas las etiquetas lingüísticas que intervienen en los M conjuntos de reglas, y cada etiqueta se representa por 3 números reales (a, b, c), que deben cumplir ciertas restricciones:

a) Limite inferior (a) ≤ Moda (b) ≤ Limite superior (c)

b) a,b,c pertenecen al dominio de valores de la variable.

Para la inicialización de la población, el primer cromosoma será la definición de etiquetas empleada en el algoritmo anterior, y los demás cromosomas se generarán a partir de este, modificando los parámetros *p* de las etiquetas lingüísticas según la expresión:

$$p(t+1) = p(t) + u \cdot anchura$$

con $u \sim U([-0.2,0.2])$ y *anchura=lim_sup-lim_inf*

El operador de mutación realizará varios cambios que consistirán en seleccionar una variable, un conjunto difuso y un parámetro *p* según una v.a.u.d. y modificarlos, de acuerdo con las restricciones descritas antes, usando la expresión anterior con (u~U([-0.1,0.1]).

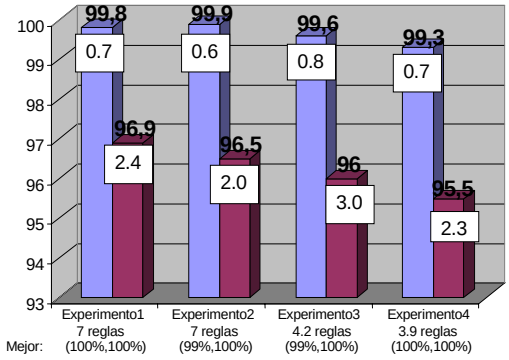
El operador de cruce, clásico por 2 puntos, intercambia variables lingüísticas completas. El algoritmo genético será un modelo elitista, conservando siempre el mejor cromosoma, utilizando como estrategia de selección un modelo de estado estable ("steady state") como antes. El criterio de parada es un número fijo de generaciones y el tamaño de la población es fijo.

Resultados experimentales

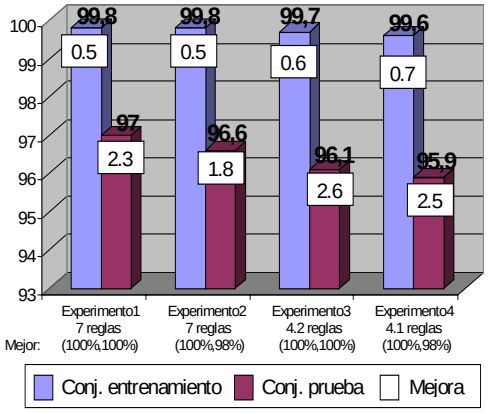
Los resultados obtenidos al aplicarlo sobre las ejecuciones del algoritmo anterior, se pueden ver en las siguientes gráficas, donde también se indica la mejora que ha conseguido el algoritmo de optimización.

A) Para el problema del IRIS

Modelo de etiquetas lingüísticas



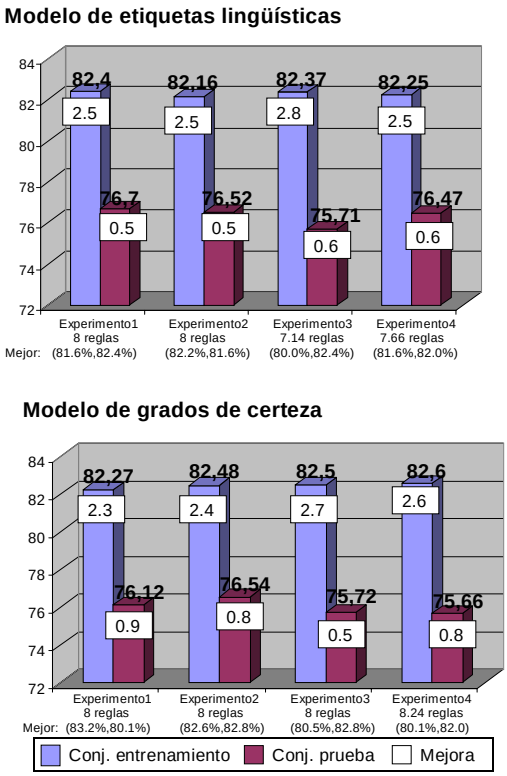
Modelo de grados de certeza



Tras el algoritmo de optimización, se produce una mejora de los resultados anteriores, consiguiendo resultados cercanos al 100% sobre el conjunto de entrenamiento y en torno al 97% sobre el conjunto de prueba. Como mejor resultado, se ha obtenido en varias ejecuciones el 100% de acierto sobre el conjunto de entrenamiento y sobre el conjunto de prueba.

B) Para el problema del PIMA

Los valores de los parámetros utilizados en la fase 2, son básicamente los mismos que en la fase 1.



Tras el algoritmo de optimización, en el PIMA también se produce una mejora, consiguiendo resultados por encima del 82% sobre el conjunto de entrenamiento y en torno al 76.5% sobre el conjunto de prueba. Como mejor resultado, en todos los experimentos se han obtenido en varias ejecuciones porcentajes de acierto superiores al 80% tanto en el conjunto de entrenamiento como sobre el conjunto de prueba.

4. Conclusiones

En este trabajo, se propone una metodología de construcción de SBRD para clasificación, representados como redes difusas, a partir del modelo propuesto por Feldman. El método permite el desarrollo de SBRDs que emplean dos tipos de reglas: con etiquetas lingüísticas y con grados de certeza en los consecuentes.

Se ha implementado un método para la obtención de un conjunto de reglas, una vez fijada la definición de las etiquetas lingüísticas que intervienen en el problema, que obtiene buenos resultados.

Se ha implementado un método de refinamiento de las etiquetas lingüísticas, basado en la evaluación sobre varios conjuntos de reglas simultáneamente, con el objetivo de evitar un posible sobreaprendizaje del conjunto de entrenamiento, y dar un carácter de búsqueda más global a la definición resultante. Al aplicarlo a los problemas del IRIS y el PIMA se han obtenido muy buenos resultados, estando a la altura de los mejores resultados publicados.

Como líneas para futuros trabajos nos proponemos aplicar el modelo con grados de certeza a la obtención de clasificaciones difusas, mejorar los algoritmos genéticos utilizados, hacer uso de medidas de similaridad para elegir en la fase de optimización, los M conjuntos de reglas utilizados, de modo que sean "suficientemente distintos", y considerar el primer algoritmo como un método de selección de características y aplicar métodos de construcción directa de las etiquetas lingüísticas en la segunda fase.

Bibliografía

[Ben98] Benítez Sánchez, J.M. *"Extracción de reglas difusas con y de redes neuronales artificiales"*. Tesis doctoral. 1998. Dpto. Ciencias Computación e I.A. Universidad Granada. 19

[Feld93] Feldman, D.S. *"Fuzzy Network Synthesis with Genetic Algoritms"*. Fifth International Conference on Genetics Algorithms

[Gold89] Goldberg, D.E. *"Genetic Algorithms in Search, Optimization and Machine Learning"*. Addison-Wesley, New York. 1989

[Ma75] Mamdani, E.H.; Assilian, S. *"An experiment in linguistic synthesis with a fuzzy logic controller"* Int. Journal of Man-Machine Studies, 7(1). Pags 1-13. 1975

[Mic92] Michalewicz, Z.; *"Genetic Algorithms + Data Structures = Evolution Programs"*, Springer-Verlag. New York, 1992

[Zad75a] Zadeh, L.A. *"The concept of a linguistic variable and its applications to aproximate reasoning. Part I"* Information Sciencies 8. Pags 199-249

[Zad75b] Zadeh, L.A. *"The concept of a linguistic variable and its applications to aproximate reasoning. Part II"* Information Sciencies 8. Pags 301-357

[Zad75c] Zadeh, L.A. *"The concept of a linguistic variable and its applications to aproximate reasoning. Part III"* Information Sciencies 9. Pags 43-80