

Clasificador Difuso para Diagnóstico de Enfermedades

Juan A. Contreras¹
Laura B. Martinez²
Yuliana V. Puerta³

Resumen

En este artículo se presenta la aplicación de un nuevo método de identificación difusa para resolver problemas de clasificación. El modelo o clasificador difuso obtenido después del proceso de entrenamiento, contiene conjuntos triangulares con solapamiento de 0.5 para el antecedente y conjuntos tipo singleton para el consecuente. En la evaluación de las reglas se emplea un operador promedio en vez de una T-norma. Los consecuentes son ajustados empleando mínimos cuadrados recursivos. El método propuesto consigue una mayor precisión que la alcanzada con los métodos actuales existentes, empleando un número reducido de reglas y parámetros, sin sacrificar la interpretabilidad del modelo difuso. El enfoque propuesto es aplicado a dos problemas clásicos de clasificación: el *Pima Indian Diabetic* y el *Dermatology Problem*, para mostrar el desempeño del método propuesto y comparar los resultados con los alcanzados por otros investigadores.

-
- 1 Coordinador de Investigación, Facultad de Ingeniería de Sistemas, Corporación Universitaria Rafael Núñez, epcontrerasj@yahoo.com
 - 2 Docente Investigador, Facultad de Ingeniería de Sistemas, Corporación Universitaria Rafael Núñez, lauramargarcia@gmail.com
 - 3 Docente Investigador, Facultad de Ingeniería de Sistemas, Corporación Universitaria Rafael Núñez, puertacruz@gmail.com

Fecha de recepción: 29 de Agosto de 2010
Fecha de aceptación: 13 de Octubre de 2010

Palabras clave

Dermatología, diabetes, identificación difusa, mínimos cuadrados recursivos.

Abstract

This paper presents the application of a new fuzzy identification method to solve classification problems. The model or fuzzy classifier, obtained after training process, contains triangular sets with 0.5 overlapping to the antecedent and singleton sets for the consequent. In the evaluation of the rules is used an average operator instead of a T-norm. The consequent are adjusted using recursive least squares. The proposed method achieves higher accuracy than others methods, using a small number of rules and parameters, without sacrificing the interpretability of the fuzzy model. The proposed approach is applied in two classic classification problems: *Pima Indian Diabetic* and *Dermatology Problem*, to show the performance of the proposed method and compare the results with other researchers.

Keywords

Dermatology, diabetes, fuzzy identification, recursive least square method.

1. INTRODUCCIÓN

Las aplicaciones biomédicas han despertado gran interés en el desarrollo de nuevos algoritmos de clasificación basados en lógica difusa y en otras técnicas de inteligencia artificial IA, como redes neuronales artificiales y algoritmos genéticos. El diagnóstico de enfermedades de la piel y la diabetes, requieren de la manipulación de un gran número de información imprecisa, con marcado grado de incertidumbre, lo que ha incentivado el desarrollo de técnicas de IA para apoyar en su detección o diagnóstico (Castellanos et al., 2003; Karahoka et al., 2009).

En este artículo se detalla un algoritmo basado en lógica difusa, a partir de datos experimentales de entrada y salida que han sido utilizados en diversos procesos de diagnóstico, con el objetivo de mejorar el grado de exactitud al diagnosticar la presencia de diabetes y enfermedades de la piel, como herramienta de apoyo al especialista de la salud, en la toma de decisiones.

Estos datos experimentales, fueron tomados de dos bancos de prueba, provenientes del Repositorio UCI (Repository of Machine Learning Databases), para el entrenamiento del algoritmo difuso propuesto. Las bases de datos seleccionadas son: Pima Indians Diabetes Database (PIDD) y Dermatology. Estos conjuntos de datos, tipo benchmark, han sido ampliamente estudiados en proyectos anteriores, de los cuales se destacan por sus resultados, el algoritmo “Neural Network ADAP Algorithm” (Smith & Everhart, 1988), considerado, uno de los primeros proyectos que trabajaron con la base de datos PIDD. Se basa en un modelo de redes neuronales (ADAP) para predecir la aparición de la diabetes en una población de alto riesgo, como lo es la tribu indígena Pima. El algoritmo fue entrenado con 576 casos y se utilizó el Algoritmo ADAP, para predecir si los otros 192 casos de prueba, desarrollaron la enfermedad en un plazo no mayor de 5 años, obteniendo un porcentaje de predicción en la prueba de la presencia de la diabetes en un 76%.

La red neuronal ARTMAP-IC (Carpenter & Markuzon, 1998), es un sistema predictivo difuso, que combina dos módulos no supervisados para llevar a cabo aprendizaje supervisado. El sistema aprende a clasificar una secuencia arbitraria de vectores

de entrada en categorías, basándose en el éxito de la predicción. Para comprobar su eficiencia trabajaron con cuatro bases de datos del repositorio UCI, entre ellas la base de datos Pima Indians Diabetes, obteniendo un porcentaje de acierto en la prueba del 81%.

También, se destaca el proyecto “Rough Sets to PIDD using ROSETTA software” (Breault, 2002), dado que obtuvo uno de los mejores porcentajes hasta ahora en el grado de acierto en la prueba (82,6%), utilizando la base de datos Pima Indians. Este resultado es consecuencia de utilizar 392 casos de los 768 que contiene la base de datos original. De los 392 casos, utilizaron 300 para entrenamiento y los restantes 92, como conjuntos de prueba. Para validar su técnica, utilizaron el Software Rosetta.

En relación a proyectos que utilizaron la base de datos Dermatology, se cita el proyecto “Genetic Programming System for Discovering Classification Rules” (Bojarczuk et al., 2001), donde los autores basados en una técnica de programación genética, dividieron la base de datos de forma aleatoria en particiones, de las cuales cinco se utilizaron como conjunto de prueba y cuatro como conjunto de entrenamiento. Este modelo obtuvo un porcentaje de acierto en la prueba del 96,64%.

Por su parte, el Algoritmo de selección de rasgos basado en la teoría de conjuntos aproximados, del autor (Caballero et al., 2007), obtuvo uno de los mejores porcentaje de acierto en la prueba, utilizando diversas bases de datos, entre ellas Dermatology. Su éxito radica en la utilización del método de selección de rasgos a través del concepto de reducto, RSReduct, los cuales realizaron comparaciones con otros métodos, obteniendo un 98% de significancia estadística que indica importantes diferencias entre otros métodos de predicción. Para probar, que no existían pérdidas de información sobre el sistema en el reducto obtenido, se realizaron pruebas con las medidas de inferencia de calidad y precisión de los conjuntos aproximados sobre dicho reducto y sobre la base completa, y obtuvieron los mismos resultados.

En cuanto al empleo de la lógica difusa para construir modelos o clasificadores a partir datos experimentales de entrada y salida (Contreras & Puerta, 2007), los algoritmos de Fuzzy C-Means (Bezdek, 1987) y de Gustafson-Kessel (1979) son unos de los más

empleados. Se han realizado diversas variaciones a estos algoritmos de agrupamiento (Nauck & Kruse, 1995; 1999; Díez et al., 2004) que consiguen una buena precisión pero sacrifican la interpretabilidad, ya que presentan interpolación de más de dos funciones de pertenencia y, por lo general, la suma de los valores de pertenencia para un dato de entrada es muy superior a 1. Uno de los principales criterios de la interpretabilidad en sistemas difusos establece que las funciones de pertenencia deben ser claramente diferentes; es decir, que tengan un rango definido en el universo de discurso y significado semántico (Espinosa & Vandewalle, 2000; Paiva & Durado, 2004).

En Fig. 1 se aprecia una clara y frecuente violación al concepto de interpretabilidad en sistemas difusos. Las funciones de pertenencia etiquetadas como Pequeño (P), Mediano (M) y Grande (G) no son claramente diferentes. Es así como un valor de 0.5 tiene una pertenencia alta en los 3 conjuntos, lo que se aparta del sentido común del usuario.

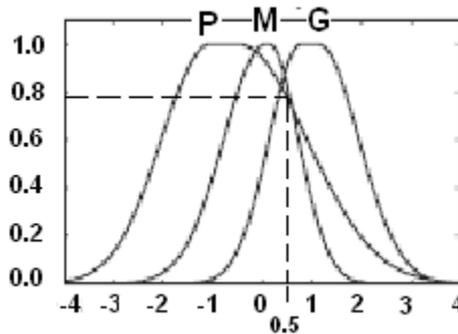


Fig. 1. Violación de la interpretabilidad en sistemas difusos

2. DESCRIPCIÓN DEL PROBLEMA

El desarrollo de sistemas que apoyen a los especialistas de la salud en el diagnóstico de enfermedades, ha tenido un especial interés por los investigadores, dado el elevado y variado número de entradas (signos, síntomas, etc.) que hay que considerar antes de tomar una decisión. Es así, como se han puesto a disposición de

la comunidad científica importantes bases de datos (como la *UCI Machine Learning Repository* y *Delve Datasets*, entre otras) para que desarrollen algoritmos que ayuden a la detección y clasificación de enfermedades. Los datos utilizados en esta investigación corresponden a dos bases de datos experimentales que se detallan a continuación.

2.1 Pima Indians Diabetes Database

La base de datos Pima Indian Diabetic Database, se ha convertido en un estándar para el análisis y precisión en el diagnóstico de la Diabetes, aplicado a diversos algoritmos, con resultados que van desde una tasa de acierto del 66% al 82,6%. Se evidencia en (Knowler et al., 1978), que ha estado bajo estudio continuo desde 1965 por el instituto nacional de la diabetes y de las enfermedades digestivas y del riñón, debido a su alto índice de la incidencia de la diabetes. Esta base de datos cuenta con información de mujeres indígenas proveniente de la tribu Pima de Arizona, igual o mayores a 21 años de edad. A los indios Pima se les conoce por ser genéticamente predispuestos a padecer diabetes (Hanson & Ehm, 1998).

La base de datos está compuesta por 768 instancias, recibiendo como entrada un conjunto de 8 variables o atributos, representados en la Tabla 1.

Tabla 1. Variables de entrada

Variable	Criterios o Atributos de predicción
1	Número de embarazos
2	Concentración plasmática de glucosa a las 2 horas de una prueba de tolerancia a la glucosa oral
3	Presión diastólica de la sangre
4	Espesor de la piel del tríceps
5	Cantidad de insulina en dos horas
6	Índice de Masa Corporal (IMC)
7	Antecedentes Familiares (DPF)
8	Edad

La variable de salida está representada por el diagnóstico de la condición del individuo, de la cual se esperan altos niveles de precisión e interpretabilidad en los datos para esta investigación. Estas son: Tested_Negative o Individuo sano, la cual presenta 500 casos (65%) y toma el valor 0, y Tested_Positive o padecimiento de la enfermedad a través de 268 casos (35%), tomando como valor 1.

2.2 Dermatology Database

Los datos corresponden a 366 registros con 34 atributos de pacientes que primero fueron evaluados clínicamente con 12 características. Después, se tomaron muestras de piel para la evaluación de 22 características histopatológicas. Los valores de las características histopatológicas fueron determinados mediante análisis de las muestras en un microscopio. Las Tablas 2 y 3 muestran estos atributos.

Tabla 2. Atributos clínicos de predicción de la enfermedad

<u>Atributos Clínicos: (toman valores 0, 1, 2, 3)</u>	
1	Erythema
2	Scaling
3	Definite borders
4	Itching
5	Koebner phenomenon
6	Polygonal papules
7	Follicular papules
8	Oral mucosal involvement
9	Knee and elbow involvement
10	Scalp involvement
11	Family history, (0 o 1)
34	Age (lineal)

La característica *Family history* (atributo 11) tiene el valor 1 si cualquiera de esas enfermedades ha sido observada en la familia, y 0 si no. La característica *Age* (atributo 34) representa la edad del paciente. Estas dos características fueron descartadas porque no tenían inherencia en el resultado. Las demás características (clínicas e histopatológicas) tienen valores entre 0 y 3, donde 0

indica que la característica no estuvo presente, 3 indica la mayor cantidad posible y 1 ó 2 indican valores relativos intermedios.

Tabla 3. Atributos hispatológicos de predicción de la enfermedad
Atributos Hispatológicos: (toman valores 0, 1, 2, 3)

12	Melanin incontinence
13	Eosinophil in the infiltrate
14	PNL infiltrate
15	Fibrosis of the papillary dermis
16	Exocytosis
17	Acanthosis
18	Hyperkeratosis
19	Parakeratosis
20	Clubbing of the rete ridges
21	Elongation of the rete ridges
22	Thinning of the suprapapillary epidermis
23	Spongiform pustule
24	Munro microabcess
25	Focal hypergranulosis
26	Disappearance of the granular layer
27	Vacuolization and damage of basal layer
28	Spongiosis
29	Saw-tooth appearance of retes
30	Follicular horn plug
31	Perifollicular parakeratosis
32	Inflammatory mononuclear infiltrate
33	Band-like infiltrate

3. DESCRIPCIÓN DEL SISTEMA DIFUSO DE DIAGNÓSTICO DE ENFERMEDADES

La estructura del modelo del sistema difuso se ha fundamentado en los siguientes criterios:

3.1 Funciones de Pertenencia

Se emplearon funciones de pertenencia triangulares normales con solapamiento de 0.5 para las particiones de los universos de las variables de entrada en el aprendizaje. Además, de permitir la reconstrucción del valor lingüístico en el mismo valor numérico, luego de aplicar un método de defuzzyficación (Pedriycz, 1994). Se escogió solapamiento de 0.5 para asegurarse que los soportes de los conjuntos difusos sean diferentes. Para el caso de las variables de salida se emplearon conjuntos difusos tipo singleton.

3.2 Distribución de las Funciones de Pertenencia

La distribución de los conjuntos difusos triangulares de las variables de entrada estará dada de forma simétrica en cada universo, buscando así que cada elemento del universo de discurso de cada variable pertenezca al menos a un conjunto difuso (Espinosa & Vandewalle, 2000).

3.3 Operadores

Se utilizó operador tipo OWA o promedio ponderado para la combinación de los antecedentes de las reglas. Este operador permite promediar los grados de pertenencia de cada antecedente evaluado (Espinosa & Vandewalle, 2000).

3.4 Método de Inferencia

El método está dado por (1)

$$f(x^{(i)}) = \frac{\sum_{j=1}^L y_j^{-j} m_j(x^{(i)})}{\sum_{j=1}^L m_j(x^{(i)})} \quad (1)$$

donde

$$m_j(x^{(i)}) = u_{A_1^j}(x_1^{(i)}) \cdot u_{A_2^j}(x_2^{(i)}) \cdot \dots \cdot u_{A_n^j}(x_n^{(i)}) \quad (2)$$

es el grado de cumplimiento de los antecedentes de la j-ésima regla de un sistema difuso tipo Sugeno, \bar{y}^j es el valor del singleton correspondiente a la regla j. El denominador siempre arroja un valor igual a 1 cuando se trata de particiones suma 1, el cual es el caso del método de aproximación propuesto en este artículo.

3.5 Descripción del Algoritmo

El algoritmo diseñado para la generación de sistemas borrosos interpretables a partir de los datos (Contreras & Acuña, 2009), está basado en la minimización del error de inferencia. Se requieren como entradas, los datos de las variables de entrada y salida que se van a evaluar.

En el algoritmo inicialmente se determinan los rangos de cada variable, luego se distribuyen las funciones de pertenencia en los universos de cada variable de entrada. Se ubican los consecuentes de tipo singleton en el espacio de salida, se determinan las reglas y se ajusta la ubicación de los consecuentes, empleando mínimos cuadrados, para minimizar el error de aproximación.

El algoritmo se detiene cuando se ha alcanzado una métrica de error menor a la requerida por el usuario o cuando el número de conjuntos borrosos por variable de entrada es mayor a 9. De otra manera se debe incrementar en 1 el número n de conjuntos de la variable de entrada y se redistribuyen las funciones de pertenencia. En Fig. 2 se describe paso a paso el algoritmo a través de un diagrama de flujo.

4. RESULTADOS Y DISCUSIÓN

Inicialmente, el algoritmo es utilizado para aproximar una función dada por los diagnósticos reales consignados en la base de datos, los cuales constituyen el objeto de análisis. A cada diagnóstico realizado por el experto se le asigna un valor numérico, o clase, los cuales son considerados como un arreglo de

datos, que representan una función, a la cual el modelo difuso obtenido debe aproximar con la mayor precisión posible. Al finalizar el entrenamiento, se genera un modelo difuso que tiene una partición de funciones de pertenencia triangulares por cada variable de entrada y un consecuente tipo singleton asociado a cada función de pertenencia triangular. Es decir, la salida del modelo difuso estará dada por (3)

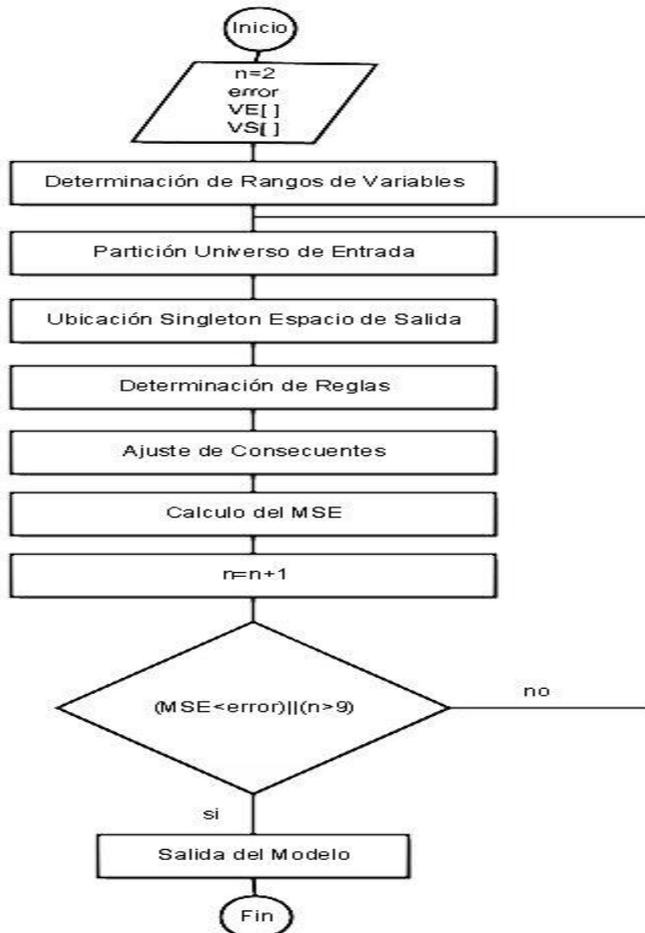


Fig. 2. Diagrama de flujo del modelo difuso

$$y = W \cdot \theta = [u_1(x_1) \ u_2(x_1) \ \dots \ u_n(x_1) \ \dots \ u_{n-1}(x_m) \ u_n(x_m)] \begin{bmatrix} \theta_{1-1} \\ \theta_{2-1} \\ \vdots \\ \theta_{n-1} \\ \vdots \\ \theta_{(n-1)-m} \\ \theta_{n-m} \end{bmatrix} \quad (3)$$

donde W es un vector que contiene los grados de pertenencia de cada dato de cada una de las m entradas disponibles, en las n funciones de pertenencia que contiene cada variable de entrada respectiva y θ es el vector de consecuentes tipo singleton asociado a cada función de pertenencia.

La salida y , será por lo tanto un valor numérico real que luego debe ser convertido al número entero más cercano a cada una de las clases especificadas.

4.1 Modelo Difuso para Diagnóstico de la Diabetes

Con la información contenida en la base de datos Pima Indian Diabetic Database se procedió a realizar el entrenamiento. Para poder establecer una comparación con otros métodos se utilizaron 576 instancias (75%) en el proceso de entrenamiento y 192 instancias (25%) en el proceso de validación. Inicialmente se procedió a obtener un modelo difuso, que aproximará la función dada por los datos correspondientes al diagnóstico médico. La Fig. 3 muestra la variación del error cuadrático medio MSE con respecto al número de funciones de pertenencia triangulares consideradas en cada iteración, durante el proceso de entrenamiento.

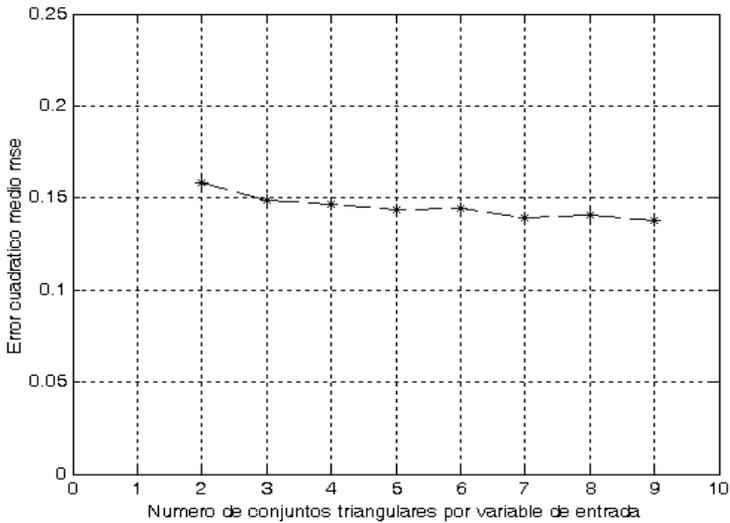


Fig. 3. Variación del error cuadrático medio con respecto al número de funciones de pertenencia triangulares. Caso Pima Indian Diabetic

El modelo difuso obtenido tiene 9 funciones de pertenencia triangulares por cada variable de entrada con solapamiento de 0,5. La salida del modelo difuso se obtiene empleando la ecuación (3). Sin embargo, como no se trata de una tarea de aproximación, sino de clasificación, se procedió a incluir un pequeño algoritmo, así:

*Si y^L es mayor a 0,5, entonces
 y^L se hace igual a 1
 En caso contrario
 y^L se hace igual a 0*

donde y^L es la salida del modelo difuso. Considerando 9 conjuntos triangulares por variable de entrada, la precisión alcanzada fue de 83,5% en el proceso de entrenamiento y de 83,1% con los datos de validación. La Fig. 4 muestra una comparación entre el diagnóstico del modelo difuso (o) y el diagnóstico real (*) para ambas clases, en el proceso de entrenamiento.

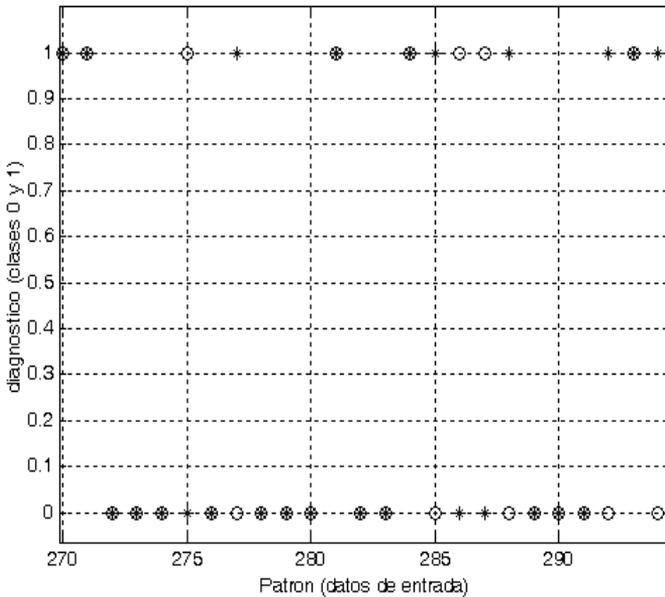


Fig. 4. Comparación entre el diagnóstico del modelo difuso (o) y el diagnóstico real del experto (*)

La Tabla 4, muestra la comparación de los resultados de esta investigación con respecto a estudios anteriores, obteniendo un porcentaje de precisión mayor en el diagnóstico temprano de la Diabetes.

4.2 Modelo Difuso para Diagnóstico de Enfermedades de la Piel

Con la información contenida en la base de datos *Dermatology Database* se procedió a realizar el entrenamiento. Para poder establecer una comparación con otros métodos se utilizaron todas las instancias (100%) en el proceso de entrenamiento. Inicialmente se procedió a obtener un modelo difuso que aproximará a la función dada por los datos correspondientes al diagnóstico médico.

Tabla 4. Comparación de métodos de clasificación que utilizan Pima Indians Diabetes, tomando como referencia porcentajes de acierto mayores a 75%

Algoritmo – Técnica	Porcentaje de acierto	Autor
Neural Network ADAP Algorithm	76%	Smith & Everhart (1988)
22 algorithms with 12-fold cross validation	Logdisc 77,7% SMART 76,8% DIPOL92 77,6%	Michie & Spiegelhalter (1994)
Bayesian neural	79,5%	Bioch & Meer (1996)
532 cases that excluded missing insulin levels	Logistic regression 80,2% k-NN 75,3% OLVQ 78,9% CART 75,6%	Ripley (1996)
ARTMAP-IC neural network	81%	Carpenter & Markuzon (1998)
Multiplier-free feedforward networks (MFN)	78,0%	Khan (2002)
Continuous-weight networks (CWN)	78,4%	Khan (2002)
14 algorithms discarded the insulin variable with the most missing cases, leaving n=532	S-Plus 79% PcOLPARS 81% PRW 80% PolyNet 78% Gnosis 81% KnowledgeMiner 78%	King & Elder IV (1998)
Rough Sets to PIDD using ROSETTA software	82,6%	Breault (2002)
Modelo Propuesto	83,1%	

En la Fig. 5 se aprecia que el menor error cuadrático medio (MSE) se alcanzó con cuatro funciones de pertenencia triangulares por variable de entrada. El modelo resultante tiene 4 funciones de pertenencia triangulares por cada una de las 32 variables de entrada (no se consideraron las entradas: family history, age), por lo tanto el vector de consecuentes tendrá 128 singletons.

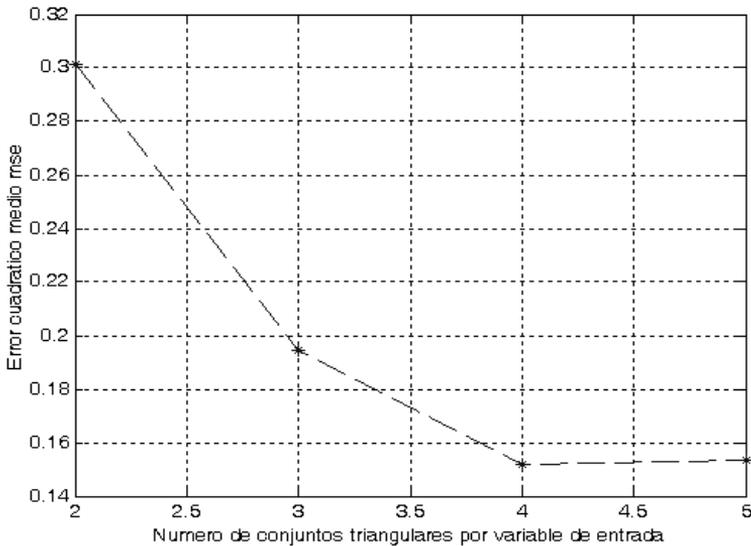


Fig. 5. Variación del error cuadrático medio con respecto al número de funciones de pertenencia triangulares. Caso Dermatología

Posteriormente se procedió a convertir cada valor de salida del modelo difuso en el número entero más cercano a cada una de las seis clases especificadas. En la Fig. 6 se muestra una comparación entre el diagnóstico del modelo difuso (o) y el diagnóstico real (*) para las seis clases de enfermedades de la piel analizadas.

La Tabla 5, muestra la comparación de los resultados de esta investigación en el diagnóstico de enfermedades de la piel, utilizando la base de datos Dermatology, obteniendo un porcentaje de precisión mayor que el obtenido en estudios anteriores.

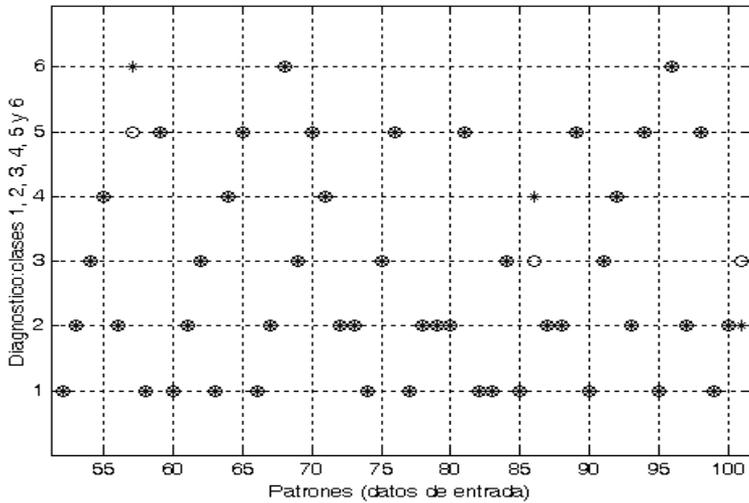


Fig. 6. Comparación entre el diagnóstico del modelo difuso (o) y el diagnóstico real del experto (*). Caso Dermatología

Tabla 5. Estudios realizados con la base de datos Dermatology Database

Algoritmo – Técnica	Porcentaje de acierto	Autor
Genetic programming system for discovering classification rules	96,64%	Bojarczuk et al., 2001
Classification rule discovery with ant colony optimization and improved quick reduct algorithm	82,37%	Jaganathan et al., 2004
Algoritmo de selección de rasgos basado en la teoría de los conjuntos aproximados	98%	Caballero et al., 2007
Algoritmo propuesto	97,2%	

5. CONCLUSIONES

Se presentó un sencillo pero efectivo algoritmo de identificación difusa para aproximación de funciones que puede ser aplicado a sistemas de clasificación. El modelo difuso obtenido es empleado como clasificador (diagnóstico) con la inclusión de un pequeño algoritmo que lleva (aproxima) los valores de salida arrojados por el modelo difuso a los valores establecidos por cada clase.

El método de identificación propuesto fue aplicado a problemas clásicos de clasificación como el *Pima Indian Diabetic* y el *Dermatology*, alcanzando unos de los valores de precisión más altos alcanzados hasta la fecha. Se presentaron comparaciones con los resultados obtenidos por otros autores.

El método de identificación difusa propuesto es parte de una investigación orientada a desarrollar una herramienta de apoyo al especialista médico en la toma de decisiones. En tal sentido, a futuro se procederá a recopilar datos en centros médicos para continuar con la validación del método presentado.

6. REFERENCIAS

- Bezdek J.C. (1987); Pattern recognition with Fuzzy Objective Function Algorithms. Ed. Plenum Press.
- Bioch, J.C. et al., (1996); Classification using Bayesian neural nets. The 1996 IEEE International Conference on Neural Networks, p. 1488-1493, Washington, DC, Institute of Electrical and Electronics Engineers.
- Bojarczuk, C., Lopes, H., Freitas, A., (2000); Genetic Programming System for Discovering Classification Rules. IEEE Eng Med Biol Mag, 19:38-44.
- Breault, J.L., (2002); Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?. Department of Health Systems Management, Tulane University.
- Caballero, Y., Alvarez, D., Baltá, A., Bellor, R., García, M., (2007); Un nuevo algoritmo de selección de rasgos basado en la teoría de los

- conjuntos aproximados. Revista de Ingeniería de la Universidad de Antioquia. ISSN 0120-6230, Número 41, pp. 130-142.
- Carpenter, G.A., Markuzon, N., (1998); ARTMAP-IC and medical diagnosis: instance counting and inconsistent cases. *Neural Networks* 11(2): 323-336.
- Castellano, G., Castiello, C., Fanelli, A.M., Leone, C., (2003); Diagnosis of Dermatological Diseases by a Neuro-Fuzzy System. *EUSFLAT Conference 2003*, p.p. 747-750
- Contreras, J., Puerta, Y., (2007); *Introducción a la Lógica Difusa*. Cartagena, Colombia. Editorial Corporación Universitaria Rafael Núñez.
- Contreras, J., Acuña, O., (2009); Generating Interpretable Fuzzy Systems for Classification Problems. *Revista Tecno Lógicas*, Vol. 23, pp. 239-255.
- Díez J.L., Navarro, J.L., Sala A., (2004); Algoritmos de Agrupamiento en la Identificación de Modelos Borrosos. *RIAI: Revista Iberoamericana de Automática e Informática Industrial*.
- Espinosa, J., Vandewalle, J., (2000); Constructing fuzzy models with linguistic integrity form numerical data-afreli algorithm, *IEEE Trans. Fuzzy Systems*, vol. 8, pp. 591 – 600, Oct. 2000.
- Guztafson, E.E., Kessel, W.C., (1979); Fuzzy Clustering with a Fuzzy Covariance Matrix. *IEEE CDC*, San Diego, California, pp. 503 – 516.
- Hanson, R.L., Ehm, M.G., (1998); An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am J Hum Genet* 63(4): 1130-1138.
- Jaganathan, P, Thangavel, K., Pethalakshmi, A., Karnan, M., (2007); Classification Rule Discovery with Ant Colony Optimization and Improved Quick Reduct Algorithm, *IAENG International Journal of Computer Science*, 33:1, IJCS_33_1_9.
- Karahoca, A., Karahoca, D., Kara, A., (2009); Diagnosis of diabetes by using adaptive neuro fuzzy inference systems. *Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009. ICSCCW 2009*.
- Khan, A.H., (2002); Multiplier-free Feedforward Networks, *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, Honolulu, Hawaii, volume 3, pp. 2698-2703.

- King, M.A. et al., (1998); Evaluation of Fourteen Desktop Data Mining Tools. IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA.
- Knowler, W.C., Bennett, P.H., Hamman, R.F., Miller M., (1978); Diabetes incidence and prevalence in Pima Indians: A 19 fold greater incidence than in Rochester, Minnesota. American Journal of Epidemiology, 108(6), 497-505.
- Michie, D. et al., (1994); Machine learning, neural and statistical classification. New York, Ellis Horwood.
- Nauck, D., Kruse, R., (1995); Nefclass - a neuro-fuzzy approach for the classification of data, In Proceedings of the Symposium on Applied Computing.
- Nauck, D., Kruse, R., (1999); Neuro-fuzzy systems for function approximation. Fuzzy Sets and System. 101(2), pp. 261-271. Jan.
- Paiva, R.P., Dourado, A., (2004); Interpretability and Learning in Neuro-Fuzzy Systems, Fuzzy Sets and System. 147, pp. 17-38. 2004.
- Pedrycz, W., (1994); Why Triangular Membership Functions?. IEEE Trans. Fuzzy Sets and System, vol. 64, pp.21-30
- Ripley, B.D., (1996); Pattern recognition and neural networks, Cambridge, New York, Cambridge University Press.
- Smith, J.W. et al., (1988); Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Symposium on Computer Applications and Medical Care (Washington, DC). R. A. Greenes. Los Angeles, CA, IEEE Computer Society Press: 261-265.