Capítulo 11

MULTICLASIFICADOR PARA DIAGNOSTICAR HIPERTENSIÓN ARTERIAL INFANTIL COMBINANDO ALGORITMOS GENÉTICOS CON MEDIDAS DE DIVERSIDAD

L. Cabrera-Hernández,L. Denoda-Pérez,A. Morales-Hernández, G.M. Casas-Cardoso, E. González-Rodríguez
Universidad Central "Marta Abreu" de Las Villas

ABSTRACT

Nowadays, deaths by cardiovascular diseases are increasing. It is known that High Blood Pressure producesan increase the probability to suffer a cardiovascular disease. The early diagnostic of hypertension is consequently a very important problem. In this paper a muticlassifier to diagnostic high blood pressure in kids of 10 to 12 years are presented. Genetic Algorithm Metaheuristic is used to combine the information of 18 base classifiers with 13 diversity measures and to obtain a good enough multiclassifier. The accuracy of the new classifier issuperior to the accuracies of all base classifiers.

Key Words: multiclassifier, diversity measures, genetic algorithm, high blood pressure in kids

RESUMEN

En la actualidad, las muertes por enfermedades cardiovasculares van en ascenso. Se conoce que la elevada presión arterial incrementa la probabilidad de padecer una enfermedad cardiovascular. El diagnóstico temprano de la hipertensión arterial es un problema muy importante. En este artículo se presenta un multiclasificador para diagnosticar hipertensión en niños de 10 a 12 años de edad. Se utiliza la metaheurística basada en algoritmos genéticos para combinar la información de 18 clasificadores bases con 13 medidas de diversidad y así obtener un buen multiclasificador. La exactitud de este nuevo clasificador es superior a la de los clasificadores individuales.

Palabras claves: multiclasificador, medidas de diversidad, algoritmos genéticos, hipertensión arterial en niños

1. INTRODUCCIÓN

El término de hipertensión arterial es cada vez más común en nuestra sociedad y su identificación como factor de riesgo cardiovascular. A pesar de ello, no todo el mundo traslada esta preocupación a los niños. La prevención de las enfermedades cardiovasculares no queda limitada a la edad adulta, sino que debe iniciarse en la edad pediátrica.

Las técnicas de clasificación están siendo muy utilizadas en la solución de diferentes problemas de la sociedad. Existen varios modelos de clasificación reportados en la literatura como las redes neuronales, árboles de clasificación y análisis discriminante. En investigaciones recientes muchos autores introducen el término multiclasificador como un "clasificador" que combina las salidas de un conjunto de clasificadores individuales, utilizando algún criterio (ej.; promedio, voto mayoritario, mínimo, etc.). Cuando se combinan clasificadores es importante garantizar la diversidad entre ellos ya que no tendría sentido combinar clasificadores cuya clasificación sea la misma. Existen varios modelos para construir un multiclasificador y todos garantizan esta diversidad de diferentes formas. En el caso de aquellos que usan distintos clasificadores bases, existen algunas medidas estadísticas que pueden ser usadas para estimar cuán diversos son. Muchos estudios teóricos y empíricos han sido publicados demostrando las ventajas del paradigma de combinación de clasificadores sobre el de clasificadores individuales.

Además de los modelos tradicionales de construcción de multiclasificadores recientemente se ha comenzado con la utilización de varias meta-heurísticas combinadas con este tema pues debido a la gran cantidad de clasificadores individuales existentes se hace cada vez mas difícil escoger cuál de ellos combinar, dentro de las meta-heurísticasexistentes se destacan los Algoritmos Genéticos por su capacidad de adaptabilidad a varios problemas, específicamente al nuestro.

En este trabajo se presentan algunas medidas de diversidad y se implementa una variante de Algoritmo Genético con el objetivo de obtener una combinación de clasificadores diversos y una exactitud del sistema multiclasificador superior a la mejor exactitud individual. Además se muestra una aplicación del algoritmo propuesto para predecir el riesgo de hipertensión arterial en niños.

El siguiente epígrafe describe los métodos de clasificación más comunes. Posteriormente se enuncian y describen las medidas de diversidad usadas en este trabajo, luego se presentan algunos conceptos básicos de algoritmos genéticos y su modelación para resolver el problema. Para finalizar se muestra una aplicación real con escolares de la ciudad de Santa Clara, Cuba para predecir el riesgo de hipertensión arterial en la edad pediátrica. El trabajo culmina con conclusiones y referencias bibliográficas.

2. MÉTODOS DE CLASIFICACIÓN

Los métodos matemáticos de clasificación están caracterizados fundamentalmente porque se conoce la información acerca de la clase a la que pertenece cada uno de los objetos. Cuando la variable de decisión, función o hipótesis a predecir es continua, a los algoritmos relacionados con los problemas supervisados se les conoce como métodos de regresión. Si por el contrario la variable de decisión, función o hipótesis es discreta, ellos se conocen como métodos de clasificación o simplemente clasificadores.

En un problema de clasificación se tienen un conjunto de objetos, elementos, instancias u observaciones divididos en clases o etiquetados. Dado un elemento del conjunto, un especialista le asigna una clase de acuerdo a los rasgos, características o variables que lo describen. Esta relación entre los descriptores y la clase puede estar dada por un conjunto de reglas. La mayoría de las veces este conjunto de reglas no se conoce y la única información que se tiene es el conjunto de ejemplos etiquetados, de forma tal que las etiquetas representan las clases.

De manera general, se puede decir que los métodos de clasificación son un mecanismo de aprendizaje, donde la tarea es tomar cada instancia y asignarla a una clase en particular.

La clasificación puede dividirse en tres procesos fundamentales: pre-procesamiento de los datos, selección del modelo de clasificación y, entrenamiento y prueba del clasificador [1].

Entre los métodos de clasificación más usados están los algoritmos basados en casos, los árboles de decisión, las redes bayesianas, las redes neuronales artificiales, el análisis discriminante y la regresión logística, pero estos no son los únicos. A continuación se presenta una breve descripción de los mencionados anteriormente.

Algoritmos basados en casos

El razonamiento basado en casos se basa en el principio de usar experiencias viejas para resolver problemas nuevos. Muchos algoritmos usan este razonamiento para resolver los problemas y entre los más comunes están los de clasificación.

Aunque todos los métodos de clasificación se basan en casos, existe un conjunto que se conoce como algoritmos basados en casos, o también como métodos de aprendizaje perezoso. Estos algoritmos deben contar con una serie de ejemplos ya conocidos y cuando van a resolver un problema nuevo, lo hacen buscando la semejanza entre éste y los ejemplos almacenados. No necesitan crear reglas, ni árboles, ni ajustar parámetros. A cada ejemplo se le conoce como instancia y a la colección de ejemplos como base de casos.

Una nueva instancia se compara con el resto de la base de casos a través de una medida de similitud. La clase de la nueva instancia será la misma que la del caso que más cercano esté a la nueva instancia. A este proceso se le conoce con el nombre de método del "vecino más cercano" (nearestneighbor). Si en lugar de usar el caso más cercano se utilizan los *k* casos más similares, entonces se habla de los k-vecinos más cercanos[1] y la clase asignada a la nueva instancia será la más común entre las *k* instancias más cercanas encontradas en la base de casos [2].

Árboles de decisión

Un árbol de decisión es un modelo matemático que clasifica las instancias ordenándolas de la raíz a las hojas. Cada nodo interior del árbol especifica una prueba de algún atributo o variable y las hojas son las clases en las cuales se clasifican las instancias u observaciones Cada rama descendiente de un nodo interior corresponde a un valor posible del atributo probado en ese nodo. Un árbol de decisión representa una disyunción de conjunciones sobre los valores de los atributos. Así, cada rama, de la raíz a un nodo hoja, corresponde a una conjunción de atributos y el árbol en sí, a una disyunción de estas conjunciones.

La familia de algoritmos ID3 [3] es el paradigma de los métodos para descubrir reglas usando árboles de decisión; a pesar de esto, tiene algunas limitaciones. Una variante para la solución de estas limitaciones es el algoritmo C4.5 [4][2], que usa puntos de corte e introduce varias medidas para evitar el sobre entrenamiento, en particular los criterios de parada de la división y de poda del árbol.

Otros árboles de decisión son el CHAID (Chi SquareAutomaticInteraction Detector) en el que la segmentación ocurre siguiendo criterios chi-cuadrados y el CRT (Classification and RegressionTree) en el que se dividen los casos en segmentos que son lo más homogéneos posibles con respecto a la variable dependiente. Varios de estos árboles se pueden encontrar en WEKA; por ejemplo: J48, Id3, BFTree, NBTree, entre otros.

[1] kNN por sus siglas en inglés (*k*NearestNeighbors). Conocido además como IBk (IB1 cuando el número de vecinos es uno) en la plataforma inteligente para aprendizaje *WaikatoEnvironmentforKnowledgeAnalysis*-WEKA.

[2] Conocido como ADTree en WEKA.

Redes bayesianas

Una red bayesiana es un modelo gráfico probabilístico que representa un conjunto de variables y sus dependencias probabilísticas. Las redes bayesianas permiten declarar supuestos de independencia condicionales que son aplicados a subconjuntos de variables. Son representadas por un gráfico acíclico dirigido, donde cada variable se representa por un nodo de la red, y de ella se especifican dos tipos de información:

- la estructura de dependencias condicionales que son los arcos de la red
- las distribuciones de probabilidad correspondientes.

Una red bayesiana puede calcular la distribución de probabilidad para cualquier subconjunto de variables de la red, dado los valores o distribuciones de las variables restantes [2]. Cuando no se conocen todos los valores de las variables en el conjunto de entrenamiento, el aprendizaje con una red bayesiana puede ser más difícil.

Este tipo de clasificador no es muy sensible a los cambios de sus parámetros, ya que se basa en información de toda la base, lo cual hace que pequeños cambios en la base no sean necesariamente significativos [5]. En WEKA hay varias de estas redes implementadas, las más sobresalientes son NaiveBayes y sus variantes.

Redes neuronales artificiales

Una red neuronal es un modelo computacional que pretende simular el funcionamiento del cerebro a partir del desarrollo de una arquitectura que toma rasgos del funcionamiento de las neuronas sin llegar a desarrollar una réplica del mismo [6]. Es una herramienta matemática para la modelación de problemas, que permite obtener las relaciones funcionales subyacentes entre los datos involucrados en problemas de clasificación, reconocimiento de patrones, regresión, etc. Este tipo de método se considera como un excelente aproximador de funciones, esencialmente no lineales, siendo capaces de aprender las características relevantes de un conjunto de datos, para luego reproducirlas en entornos ruidosos o incompletos [7].

En los últimos años se han producido una amplia variedad de arquitecturas de redes neuronales, encontrándose entre las más utilizadas, las redes multicapa de alimentación hacia adelante (Feed-Forward Neuronal Networks, FFN), las cuales se distinguen porque sus neuronas están conectadas a manera de grafo acíclico dirigido (todos los arcos hacia adelante). Las redes MultiLayerPerceptron (MLP) constituyen un ejemplo genérico de las redes FFN, y se encuentran formadas por un conjunto de capas de neuronas ordenadas secuencialmente. Primero una capa de entrada, luego un conjunto de capas intermedias denominadas capas ocultas y por último una capa de salida. Las MLP usando neuronas ocultas con funciones no lineales, son capaces de aproximar cualquier tipo de función continua y brindar excelentes resultados en las tareas de clasificación [8].

Regresión logística

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas o categóricas. En este último caso, se requiere que sean transformadas en variables "*dummy*"; es decir, variables simuladas.

El propósito del análisis consiste en predecir la probabilidad de que a alguien le ocurra cierto "evento". Puede, además, determinar cuáles variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión. Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso de cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

La regresión logística sólo resuelve problemas de clasificación binarios. Si el problema fuese más general, entonces se puede aplicar un modelo más general basado en los mismos principios, denominado regresión multinomial, precisamente este criterio es el que utiliza la función *Logistic*, implementada en WEKA.

Como se ha visto, se han desarrollado un gran número de clasificadores, pero determinar cuál de ellos logra encontrar una mejor frontera de decisión para separar las clases es el mayor problema. En la búsqueda de mejores métodos de clasificación aparece una tendencia a combinar varios de estos clasificadores. Los algoritmos llamados multiclasificadores se basan en esta idea; utilizar varios clasificadores y combinar sus diferentes salidas [9] con el objetivo de alcanzar un mejor resultado.

3. MEDIDAS DE DIVERSIDAD COMO CRITERIO PARA SELECCIONAR LOS CLASIFICADORES DE BASE

Existen dos tipos de medidas de diversidad: las medidas pareadas (pairwise) y las medidas grupales (non pairwise).

Las medidas en forma de pares se calculan por pares de clasificadores usando sus salidas, las cuales son binarias (1,0) que indica si la instancia fue correctamente clasificada o no por el clasificador.

A continuación se indica en la Tabla 1 el resultado de dos clasificadores (C_i, C_j) para una instancia en cuanto a si la clasificaron correctamente o no.

Tabla 1: Matriz binaria para una instancia

	C_j correcto (1)	C_j incorrecto (0)
C_i correcto (1)	a	b
C_i incorrecto (0)	c	d
a + b + c + d = 1		

Si se suman para todas las instancias los valores de a, b, c, d se obtendrán los resultados mostrados en la Tabla 2:

Tabla 2: Matriz binaria para N instancias

	C_j correcto (1)	C_j incorrecto (0)
C_i correcto (1)	A	B

C_i incorrecto (0)	C	D
A +B + C + D = N		

Donde N es el número total de instancias. Un conjunto de L clasificadores produce L (L – 1)/2 pares de valores. Para obtener un único resultado habría que promediar estos valores.

A continuación se presentan las medidas pareadas tenidas en cuenta en este trabajo.

Coeficiente de correlación ρ

Entre las medidas de diversidad está el coeficiente de correlación [10] el cual se calcula como,

$$\rho_{ci,cj} = \frac{A \times D - B \times C}{\sqrt{(A+B) \times (C+D) \times (A+C) \times (B+D)}} \quad (1)$$

Mientras menor sea el valor de ρ, mayor será la diversidad. Los valores de ρ estarán en el intervalo [-1, 1].

El estadístico Q

El estadístico Q (Q Statistics) es otra de las medidas para pares de clasificadores. Se calcula de la siguiente forma:

$$\mathrm{Qci, cj} = \frac{A \times D - B \times C}{A \times D + B \times C} \quad (2)$$

Para cualquier par de clasificadores, los valores de ρ y Q tendrán el mismo signo y se puede probar que $|\rho| \leq |Q|$[11].

La medida de diferencias

La medida de diferencias (TheDisagreementMeasure) introducida por Skalak[12], es igual a la probabilidad de que los dos clasificadores discrepen en sus predicciones. Mientras mayor sea su valor mayor será la diversidad.

$$D_{ci,cj} = \frac{B + C}{N} \quad (3)$$

La medida de doble fallo

La medida de doble fallo (TheDouble- FaultMeasure) introducida por Giacinto y Roli[13] considera el fallo de los dos clasificadores al mismo tiempo. Está basada en el concepto de que es más importante conocer cuando errores simultáneos son cometidos que cuando ambos tienen clasificación correcta. Mientras menor sea el valor mayor será la diversidad.

$$D_{ci,cj} = \frac{\mathrm{D}}{\mathrm{N}} \tag{4}$$

A continuación se enuncian las medidas grupales utilizadas en este trabajo.

La medida de Entropía

La medida de Entropía (TheEntropyMeasure) [11] se basa en la idea intuitiva de que en un conjunto de N casos y L clasificadores la mayor diversidad se obtendrá si L/2 de los clasificadores clasifican una instancia correctamente y los otros L- L/2 la clasifican incorrectamente. Fue introducida por Cunningham y Carney en [14].

$$E = \frac{1}{N} \cdot \frac{2}{L-1} \sum_{j=1}^{N} min\left\{\left(\sum_{i=1}^{L} y_{j,i}\right), \left(L - \sum_{i=1}^{L} y_{j,i}\right)\right\} \quad , y_{j,i} \in \{0,1\}, \;\; 0 \leq E \leq 1 \tag{5}$$

Varianza de Kohavi-Wolpert

La varianza de Kohavi-Wolpert (Kohavi-WolpertVariance), fuepropuesta en[15]. Esta medida es originada de la descomposición de la varianza del sesgo del error de un clasificador.

$$KW = \frac{1}{NL^2} \sum_{j=1}^{N} Y(z_j)\left(L - Y(z_j)\right) \, , 0 \leq \mathrm{KW} \leq 1 \;\; donde Y(z_j) = \sum_{i=1}^{L} y_{i,j} \tag{6}$$

Con esta medida, la diversidad disminuye a medida que el valor de KW aumenta.

Medida de desacuerdo entre expertos

La medida de desacuerdo entre expertos (Measurement interrateragreement) [16]. Se desarrolla como una medida de fiabilidad entre clasificadores. La diversidad disminuye cuando el valor de k aumenta. El k se calcula por:

$$k = 1 - \frac{\frac{1}{L}\sum_{j=1}^{N} Y(Z_j)\left(L - Y(Z_j)\right)}{N(L-1)p(1-p)}, -1 \leq k \leq 1 \tag{7}$$

Donde el término de la derecha es la medida de concordancia de Kendall y p es la media de la exactitud de la clasificación individual, y se calcula como:

$$p = \frac{1}{N \cdot L} \cdot \sum_{j=1}^{N} \sum_{i=1}^{L} y_{j,i} \tag{8}$$

Medida de dificultad

La medida de dificultad (The Measure of "difficulty" ϴ) viene del estudio realizado en[17]. Se calcula a través de la varianza de una variable aleatoria discretay denota la probabilidad de que exactamente *i* clasificadores hayan clasificado bien todas las instancias. La diversidad del ensamblado aumenta con el decremento del valor de la medida de dificultad.

$$\theta = Var(x) \tag{9}$$

Medida de diversidad generalizada

La medida de diversidad generalizada (GeneralizedDiversity) se enunció en[18].

$$GD = 1 - \frac{p(2)}{p(1)}, 0 \leq GD \leq 1, \qquad \text{donde} \tag{10}$$

$$p(1) = \sum_{i=1}^{L} \frac{i}{L} * p[i], \quad p(2) = \sum_{i=1}^{L} \frac{i * (i-1)}{L * (L-1)} * p[i] \tag{11}$$

El valor de *GD* varía entre 0 y 1, siendo 0 la menor diversidad cuando p(2)=p(1) y 1 la mayor diversidad cuando p(2)=0 y L la cantidad de clasificadores.

Medida de diversidad de coincidencia de fallos

Esta medida (CoincidentFailureDiversity) se enuncia también en[18], como una mejora a la medida anterior.

$$CFD = \begin{cases} 0, & p[0] = 1 \\ \frac{1}{1 - p[0]} * \sum_{i=1}^{L} \frac{L - i}{L - 1} \times p[i], & p[0] < 1 \end{cases} \tag{12}$$

El valor de CFD está entre 0 y 1 y mientras mayor sea, mayor será la diversidad.

Medida de diversidad de distintos fallos

Esta medida (DistinticFailureDiversity) fue igualmente enunciada en[18], como una mejora a la medida anterior.

$$DFD = \begin{cases} 0, & t[i] = 0 \\ \sum_{i=1}^{L} \frac{L-i}{L-1} \times t[i], & t[i] > 0 \end{cases} \quad (13)$$

El valor de DFD está entre 0 y 1 y mientras mayor sea, mayor será la diversidad.

Medida de la diversidad global

La medida de la diversidad global (OverallDiversity) fue enunciada en[19] como una versión "*pesada*" de la medida de diversidad de distintos fallos.

$$DFD = \begin{cases} 0, & t[i] = 0 \\ \sum_{i=1}^{L} \frac{L-i}{L-1} \times t[i] \times w[i], & t[i] > 0 \end{cases} \quad (14)$$

Cada posición de *w* representa el promedio de valores *d* para cada fila donde *i* clasificadores fallaron. Los valores *d* se calculan para cada instancia omo:

$$d_i = \sum_{j=0}^{K} \left[\sqrt{\frac{C_{ki}}{{n_i}^2}} \right] \quad (15)$$

El valor de OD está entre 0 y 1 y mientras mayor sea, mayor será la diversidad.

Medida de variabilidad

Esta medida (TheMeasure of Variability*)* tiene en cuenta si las clases asignadas por los clasificadores en cada instancia son distintas o no. Mientras mayor sea suvalor, mayor es la diversidad.

$$Var = \frac{\sum_{i=1}^{N} a}{N} \; dondea = \begin{cases} 0 & si E_1(i) = E_2(i) = \cdots = E_L(i) \\ 1 & e.o.c \end{cases} \quad (16)$$

Para lograr combinar varias medidas de diversidad y obtener un solo valor se tiene en cuenta el operador **Fuzzy**. Este utiliza la teoría de los conjuntos borrosos para calcular el promedio de pertenencia de cada una de las medidas a los conjuntos borrosos y retorna el máximo estandarizado de esos valores. Los términos lingüísticos manejados por este operador son *baja diversidad* y *alta diversidad.*

Las funciones de pertenencia usadas fueron triangulares y su formulación se define en la figura 1.

a)

$$triangle(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{x}{0.6}, & 0 \le x \le 0.3, \\ \dfrac{0.6 - x}{0.3}, & 0.3 \le x \le 0.6, \\ 0, & 0.6 \le x \end{cases}$$

a)

$$triangle(x) = \begin{cases} 0, & x \le 0.4, \\ \dfrac{x - 0.4}{0.3}, & 0.4 \le x \le 0.7, \\ \dfrac{1 - x}{0.3}, & 0.7 \le x \le 1, \\ 0, & 1 \le x \end{cases}$$

b)

Figura 1: Funciones de pertenencia para los términos lingüísticos *baja diversidad* y *alta diversidad*. a) Representación gráfica, b) y c) Representación analítica de estos términos

Con lo anterior se puede determinar la pertenencia a cada uno de los conjuntos de las medidas de diversidad que se estén analizando.

4. ALGORITMOS GENÉTICOS COMBINADO CON MEDIDAS DE DIVERSIDAD PARA OBTENER UN BUEN MULTICLASIFICADOR

Los algoritmos genéticos (AGs) son métodos de búsqueda basados en los principios generales de la genética natural, se ejecuta para un número fijo de generaciones o hasta que algún criterio de parada es satisfecho.

La configuración del Algoritmo Genético depende del tipo de problema a resolver. En este caso, el algoritmo genético es presentado usando medidas de diversidad con el objetivo de combinar clasificadores diversos y proveer la mejor exactitud posible. El conjunto de todos los parámetros del algoritmo genético y la definición de la función objetivo son:

Configuración del cromosoma

El cromosoma representará las posibles soluciones de nuestro problema.

Gen: variable binaria que toma valor 1 si el clasificador pertenece a la combinación y 0 en otro caso.

Cromosoma: Secuencia de genes que representan el conjunto de todos los clasificadores base que serán usados en el sistema multiclasificador.

La siguiente ecuación muestra los aspectos anteriores:

$$C_x = (g_1, g_2, \dots, g_L) g_i = \begin{cases} 0 & \text{, clasificador } i \text{ no está presente} \\ 1 & \text{, clasificador } i \text{ está presente} \end{cases}$$

Descripción de la función objetivo

Como se desea obtener de forma simultánea la mejor diversidad entre los clasificadores que son usados en el sistema multiclasificador y la mejor exactitud que se pueda obtener con él, se calcula el valor de *f*quees la suma entre la exactitud del sistema multiclasificador y el resultado de las medidas de diversidad, según la configuración del cromosoma.

$\mathrm{f}(\mathrm{C_x}) = \mathrm{Exactitud}(\mathrm{C_x}) + \mathrm{Diversidad}(\mathrm{C_x})$,donde$\mathrm{C_x}$ es el cromosoma

Por tanto, la función objetivo en el proceso evolucionario será:

$\max_{0 \le x \le P} f(C_x)$,donde$P$ es el tamaño de la población

Dado que pueden existir casos en los cuales el primer parámetro de la función objetivo podría ser pequeño y el valor de $f(C_x)$ser altoporque hay una gran diversidad entre los clasificadores.Como el principal objetivo en este trabajo es encontrar la combinación con exactitud superior en la clasificación y al mismo tiempo hallar la diversidad superior entre los clasificadores,entonces se agrega otra restricción,la cual manifiesta que el resultado final será la combinación donde la **exactitud del multiclasificador** sobrepase la mejor exactitud obtenida con los clasificadores individualesy entre ellas,la combinación con más **diversidad**.

Configuración de la población

En la configuración de la población varios elementos son necesarios;por ejemplo,el número de individuos en la población y el número de ellos que serán reemplazados en cada iteración.

Se utiliza un tamaño de población, igual a$2^{\frac{L}{2}}$donde L es el número de clasificadores;es decir, el tamaño de los cromosomas.Este tamaño fue sugerido tratando de evitar pequeños espacios de soluciones o un alto tiempo computacional mientras se analiza este espacio.

La población inicial será generada usando un híbrido entre la generación aleatoria y el sembrado de individuos.

Cada cromosoma es generado aleatoriamente yel valor del gen es 0 o 1 dependiendo de la presencia del clasificador en la combinación,para elloun número aleatorio *r* es generado,si *r* es mayor que 0.5 el clasificador será incluido y por eso el gen será 1;en otro caso, el clasificador no es incluido y el gen será 0.Después de que todos los cromosomas son generados como se explica antes,los mejores clasificadores individuales también serán incluidos en la combinación,poniendo el valor correspondiente de este gen igual a 1.

Los operadores de selección,cruzamiento y mutación son explicados a continuación,ellos son usados para simular la recombinación genética y el mecanismo de selección natural.

Operador de cruzamiento

En el caso del cruzamiento,es permitido seleccionar fragmentos del genotipo de cromosomas que no son muy buenos independientemente,pero cuando son mezclados,pueden ser una mejor solución respecto a la anterior.Hay varias formas de definir este operador;en nuestro caso,se usó el operador clásico de cruzamiento en un punto y el cruzamiento uniforme.

En el cruzamiento en un punto,dos cromosomas son seleccionados de forma aleatoria a partir de la población intermedia;estos dos cromosomas actuarán como padres.Una posición del gen es escogida aleatoriamente y como resultado de este cruzamiento dos nuevos cromosomas son obtenidos.

En el cruzamiento uniforme,cada padre tiene la misma probabilidad de contribuir con sus genes para el único individuo resultante.Si un número generado aleatoriamente es más pequeño o igual que 0.5,entonces el gen será tomado del primer padre; en otro caso,será tomado del segundo padre.

Una vez que concluye el proceso de recombinación genética por medio del cruzamiento,si los cromosomas nuevos ya existen en la población, entonces es realizada una mutación para obtener cromosomas nuevos y diferentes.

La probabilidad de ocurrencia del cruzamiento estará definida por el usuario.

Operador de mutación

La implementación de este operador es muy simple.El operador tradicional de mutación se define de la siguiente forma:aleatoriamente se escoge un cromosoma,aleatoriamente también se seleccionael gen a mutar y se cambia su estado:0 por 1 o 1 por 0, lo cualgarantiza que la inclusión del clasificador se modifica en la combinación.Si el cromosoma resultante existe previamente,entonces se escoge otro punto de mutación y se repite el proceso. Si como resultado de explorar todos los puntos de mutación no se obtuviese ningún cromosoma nuevo, se selecciona otro cromosoma para mutar.La probabilidad de ocurrencia de la mutación estará definida por el usuario.

Operador de selección

Teniendo en cuenta las características del problema se utiliza el método de la ruleta,que no permite la selección de un individuo más de una vez.

En el método de la ruleta la probabilidad usada para cada cromosoma es calculada dividiendo el resultado de la función objetivo para el cromosoma entre la suma de la función objetivo de cada cromosoma en la población con tamaño *P'*.Esto se muestra en la siguiente fórmula:

$$p(C_i) = \frac{f(C_i)}{\sum_{i=1}^{P'} f(C_i)}$$

Para resumir se enuncian los pasos necesarios para el funcionamiento del AG.Cada iteración simple comienza con una población que tiene tamaño igual al número previamente especificado,esta población es generada usando un híbrido entre la generación aleatoria y el sembrado de individuos.Después de que una población intermedia es generada por el operador de selección,entonces los cromosomas nuevos son generados por el proceso de la recombinación,se agregarán para la población inicial y podrán estar o no en la nueva población.

La población será limpiada de cromosomas que probabilísticamente tomen los valores más pequeños en la función objetivo,hasta conservar el tamaño establecido (método de la ruleta).

El algoritmo para cuando sea cierto al menos una de las siguientes condiciones:

- El usuario especifica la parada cuando el algoritmo encuentra la primera combinación que satisface las condiciones y restricciones del problema.
- Se ha alcanzado el número de generaciones definidas por el usuario.

5. ANÁLISIS DE HIPERTENSIÓN ARTERIAL EN EDAD PEDIÁTRICA

El término de hipertensión arterial sistémica (HTA) es cada vez más común en nuestra sociedad y su identificación como factor de riesgo cardiovascular, sin embargo, no todo el mundo traslada esta preocupación a los niños. Las guías de la Sociedad Europea de Hipertensión (ESH) y de la Sociedad Europea de Cardiología (ESC) del tratamiento de la HTA, publicadas en 2003 y actualizadas en 2007, no incluyen, lamentablemente, ninguna sección dedicada a la HTA en niños y adolescentes[20].

La prevención de las enfermedades cardiovasculares no queda limitada a la edad adulta, sino que debe iniciarse en la edad pediátrica. La HTA es la mayor causa de morbimortalidad en muchos países, por sus consecuencias sobre el sistema cardiovascular y los accidentes cerebrovasculares. Se ha demostrado que la HTA en la infancia es un factor de riesgo independiente para la hipertensión en la edad adulta y está asociada con marcadores precoces de enfermedad cardiovascular como hipertrofia ventricular izquierda, espesor de la íntima-media, complianza arterial, ateroesclerosis y disfunción diastólica. La prevalencia global de HTA en adultos es del 15-20%; mientras que, en niños con edades entre 4 y 15 años se estima en un 2%.

El diagnóstico de hipertensión en niños es complicado porque los valores normales y anormales de la presión sanguínea varían con la edad, el sexo y la talla, con un amplio rango descrito en tablas y por tanto, son difíciles de recordar. Se ha demostrado que la hipertensión en la infancia es un factor de riesgo independiente para la hipertensión en la edad adulta y está asociada con marcadores precoces de enfermedad cardiovascular. Considerando que la morbilidad y la mortalidad a largo plazo están asociadas a la hipertensión arterial, un componente importante para la salud de los niños y de los adolescentes es intervenir a tiempo [21].

5.1 Descripción de la muestra

En este estudio, la muestra estuvo constituida por un total de 680 niños supuestamente sanos entre 8 a 12 años de edad, de ambos sexos, pertenecientes a 4 escuelas primarias de la ciudad de Santa Clara. Los datos fueron suministrados por el proyecto PROCDEC de la Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba[22]. Se analiza un conjunto de 31 variables aleatorias empleadas en el diagnóstico de riesgo cardiovascular para predecir el riesgo de que un niño sea o no hipertenso. La tabla 3 muestra las características fundamentales de las variables aleatorias que son discretas y la tabla 4 muestra las características fundamentales de las variables aleatorias que son continuas.

Tabla 3: Variables predictoras discretas

Variables	Identificador	Valores	Porcentaje
Sexo	Sexo	Femenino	50.4

		Masculino	49.6
Color de la Piel	ColorPiel	Blanca	81.5
		No blanca	18.5
Diagnóstico	newdiag	Sí	57.5
		No	42.5

Tabla 4: Variables predictoras continuas

Variable	Identificador	Mínimo	Máximo
Edad	EdadA	8	11
Peso actual (en kg)	PesoKg	20.5	75.1
Talla (en cm)	Talla	113	191
Circunferencia Cintura (en cm)	CCintura	37	182
Circunferencia Cadera (en cm)	CCadera	30	106
Índice Cintura Cadera	icc	0.58	2.5
TA Sistólica Miembro Inferior	TSMI	11	150
TA Diastólica Miembro Inferior	TDMI	20	90
TA Sistólica Miembro Superior	T1dBS	81	150
TA Diastólica Miembro Superior	T1dBD	6	99
TA Sistólica 5 min	T1d5S	73	152
TA Diastólica 5 min	T1d5D	46	95
TA Sistólica 10 min (antes 15 min)	T1d10S	11	149
TA Diastólica 10 min (antes 15 min)	T1d10D	45	94
TA 1er día Sistólica Media	T1SMedia	77.33	146
TA 1er día Diastólica Media	T1DMedia	43.33	88.67
TA 2do día Sistólica Basal	T2dBS	82	150
TA 2do día Diastólica Basal	T2dBD	17	94
TA 2do día Sistólica P. Peso Sostenido	T2dPPSS	78	174
TA 2do día Diastólica P. Peso Sostenido	T2dPPSD	48	116
Presión Arterial Media 2d	PAM2d	58	129.67
TA 3er día Sistólica Basal	T3dBS	81	144
TA 3er día Diastólica Basal	T3dBD	40	110
TA 3er día Sistólica P. Peso Sostenido	T3dPPSS	83	160
TA 3er día Diastólica P. Peso Sostenido	T3dPPSD	13	112
Presión Arterial Media 3d	PAM3d	53.33	120
Índice de masa corporal	imc	8.77	255.19

5.2 Modelos matemáticos

Se obtienen varios modelos de clasificación utilizando árboles de decisión, redes bayesianas, redes neuronales, regresión logística y otros incorporados en el software Weka (WaikatoEnvironmentforKnowledgeAnalysis) versión 3.7.5. Este software es libre, de código abierto, y tiene incorporado muchos métodos estadísticos y de Inteligencia Artificial.

En la tabla 5 se muestra la exactitud de estos modelos individuales obtenidos

Tabla 5: Exactitud de los clasificadores base

Clasificador	Exactitud
weka.classifiers.trees.FT (*Functional Tree structure*)	0.8831
weka.classifiers.lazy.LWL (*Locally weighted learning*)	0.7749
weka.classifiers.trees.RandomForest	0.8874
weka.classifiers.lazy.IBk	0.8225
weka.classifiers.functions.SPegasos	0.8874
weka.classifiers.trees.REPTree	0.8658
weka.classifiers.bayes.NaiveBayes	0.8225
weka.classifiers.functions.Logistic	0.9091
weka.classifiers.lazy.IBk	0.8182
weka.classifiers.trees.J48	0.8701
weka.classifiers.functions.MultilayerPerceptron	0.8701
weka.classifiers.trees.ADTree	0.9004
weka.classifiers.functions.SGD (*Stochastic Gradient Descent*)	**0.9134**
weka.classifiers.trees.RandomTree	0.8182
weka.classifiers.functions.SMO(*Sequential minimal optimization*)	**0.9134**
weka.classifiers.lazy.KStar	0.7013
weka.classifiers.functions.VotedPerceptron	0.5800
weka.classifiers.trees.LMT (*Logistic Model Trees*)	0.8961

Como se observa el mejor por ciento de clasificación se logra utilizando los clasificadores SGD y SMO, con una exactitud de 0.9134.

Luego combinando las 13 medidas de diversidad mencionadas en el epígrafe 3 (R, p, Q, D, DF, E, KW, k, DIF, GD, CFD, DFD, OD), utilizando el operador fuzzy mencionado anteriormente y teniendo como configuración del Algoritmo Genético: 50 generaciones, 0.25 probabilidad de mutación y 0.75 probabilidad de cruzamiento, se logran los resultados mostrados en la tabla 6.
En ella se muestra el cromosoma resultante del algoritmo genético y la exactitud del sistema multiclasificador, como se observa este es un valor superior al mejor valor obtenido anteriormente.

Tabla 6 Resultados del Algoritmo Genético

C_x	Exactitud del sistema multiclasificador
001001000100101100	0.9480

El cromosoma resultante corresponde a la combinación de los siguientes clasificadores:

weka.classifiers.trees.RandomForest, weka.classifiers.trees.REPTree,
weka.classifiers.trees.J48,
weka.classifiers.functions.SGD, weka.classifiers.functions.SMO,
weka.classifiers.lazy.KStar.

Nótese que con este cromosoma se mejora la clasificación individual en un 3%.

En la figura siguiente se muestra una comparación de los resultados de los clasificadores y el multiclasificador en cuanto a la exactitud en la clasificación. Obsérvese que el sistema multiclasificador supera a los clasificadores individuales.

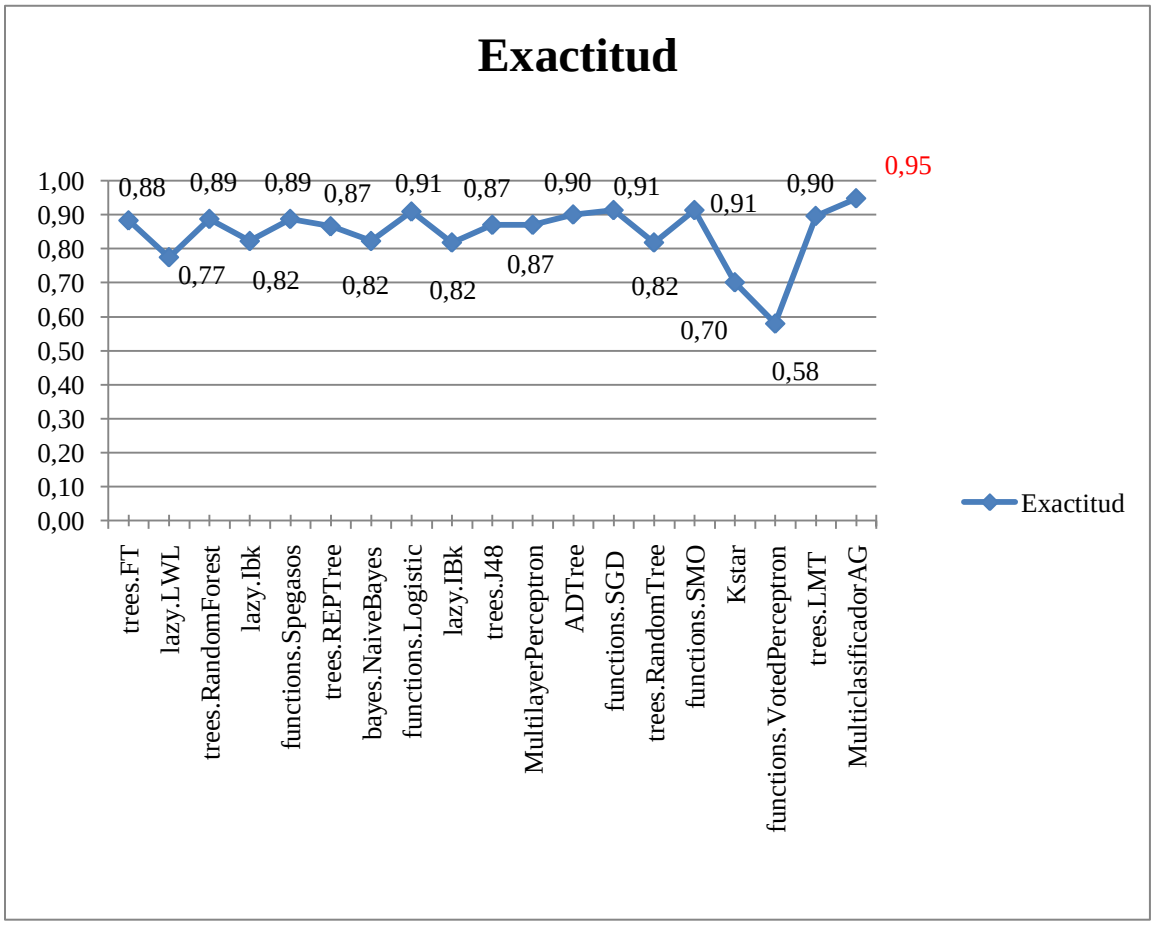

Figura 2: Comparación de los 19 modelos obtenidos en cuanto a la exactitud o por ciento de casos bien clasificados.

6. CONCLUSIONES

El presente trabajo muestra una técnica novedosa que emplea Algoritmos Genéticos para encontrar un buen conjunto de clasificadores diversos. La función objetivo del Algoritmo Genético involucra la

exactitud del sistema multiclasificador y los resultados de medidas de diversidad aplicadas a los clasificadores individuales del sistema.

Un caso de estudio de la base de HTA es usado para ejemplificar esta contribución. Se aplicaron en total 18 clasificadores base y sus resultados individuales no superan el 91%. Usando la propuesta del Algoritmo Genético con medidas de diversidad, se obtiene un multiclasificador que logra mejorar en un 3% la clasificación anterior.

REFERENCIAS

[1] Bonet, I. (2008), *Modelo para la clasificación de secuencias, en problemas de la bioinformática, usando técnicas de inteligencia artificial.*, in *Ciencias de la Computacion*. 2008, Universidad Central "Martha Abreu" de las Villas.: Santa Clara.

[2] Mitchell, T.M. (1997) *Machine Learning.*McGraw-Hill Science/Engineering/Math, 1997: p..

[3] Quinlan, J.R. (1986) *Induction of decision trees.Machine Learning*, 1986: p. 1**,** 81-106.

[4] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning.* San Mateo,CA, Morgan Kaufmann., 1993.

[5] Chavez, M.D.C. (2008), *Modelos de redes bayesianas en el estudio de secuencias genómicas y otros problemas biomédicos.*, in *Ciencias de la Computacion*. 2008, Universidad Central "Martha Abreu" de las Villas.: Villa Clara.

[6] Bello, R., et al. (2001) *Aplicaciones de la IA, Santa Clara.*2001.

[7] Wolpert, D. (1992) *Stacked generalization.*Neural Networks, 1992: p. 5**,** 241-259.

[8] Salazar, S. (2005) *NEngine v1.0. Una Herramienta Software para Redes Neuronales Recurrentes.* .Universidad Central "Marta Abreu" de Las Villas., 2005.

[9] Polikar, R. (2006), *Ensemble based systems in decision making*, in *IEEE Circuits and Systems Magazine.* . 2006.

[10] Kuncheva, L.I. (2004) *Combining Pattern Classifiers: Methods and Algorithms.*New York, NY, Wiley Interscience., 2004.

[11] Kuncheva, L.I. and C.J. Whitaker. (2003) *Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. .MachineLearning*, 2003. **51**: p. 181-207.

[12] Skalak, D.B. (1996) *The sources of increased accuracy for two proposed Boosting algorithms.*1996.

[13] Giacinto, G. and F. Roli. (2001) *Design of effective neural network ensembles for image classification purposes.*2001.

[14] Cunningham, P. and J. Carney. (2000) *Diversity versus Quality in Classification Ensembles Based on Feature Selection, in Machine Learning.*ECML , R. López de Mántaras and E. Plaza, Editors. 2000, Springer Berlin / Heidelberg., 2000: p. p. 109-116.

[15] Kohavi, R. and W. D.H. (1996.) *Bias Plus Variance Decomposition for Zero-One Loss Functions in Machine Learning.*Proceedings of the Thirteenth International Conference. , 1996.

[16] Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions.*John Wiley & Sons., 1981.

[17] Hansen, L.K. and S. P. (1990.) *Neural Network Ensembles.*IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990. : p. 12: p. 993-1001.

[18] Partridge, D. and K. W. (1997) *Software diversity: practical statistics for its measurement and exploitation.*Information and Software Technology, 1997: p. 39(10): p. 707-717.

[19] Partridge, D. and W. Krzanowski. (1997) *Distinct failure diversity in multiversion software.*Res. Rep, 1997. **348**.

[20] G. Mancia, et al. (2007) *Guidelines for the management of arterial hypertension: The Task Force for the Management of the Arterial Hypertension of the European Society of Hypertension (ESH) and the European Society of Cardiology (ESC).*J Hypertens, 2007. **25**.
[21] Ortigado, A. (2011) *Hipertensión arterial sistémica.* Tratado de Pediatría Extrahospitalaria. Madrid: Ergon, 2011: p. 455-62.
[22] UCLV and U.D.O *Proyección del Centro de Desarrollo Electrónico hacia la Comunidad (PROCDEC)*