ALGORITMOS GENÉTICOS CON MEDIDAS DE DIVERSIDAD PARA EL DIAGNÓSTICO DEL RIESGO DE HTA EN MENORES

Leidys Cabrera Hernández[*], Alejandro Morales Hernández[**], Gladys M. Casas Cardoso[*], Lisset Denoda Pérez [*], Emilio F. González Rodríguez[*], Jesús[*]

[*]Departamento de Computación, Facultad Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Cuba.

e-mail: leidysc@uclv.edu.cu, alejandrom@uclv.edu.cu, gcasas@uclv.edu.cu, ldenoda@uclv.edu.cu, eglez@uclv.edu.cu, jesusar@ucm.vcl.sld.cu

Temática: Redes, computación e informática médica Propuesta de modalidad: Ponencia

Resumen

El término de hipertensión arterial 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. La prevención de las enfermedades cardiovasculares no queda limitada a la edad adulta, sino que debe iniciarse en la edad pediátrica.

La combinación de clasificadores es un área activa de investigación en la comunidad del "*machine learning*" y el reconocimiento de patrones. Muchos estudios teóricos y empíricos han sido publicados demostrando las ventajas del paradigma de combinación de clasificadores sobre el de clasificadores individuales. Cuando se combinan clasificadores es importante garantizar la diversidad entre ellos. Algunas medidas estadísticas pueden ser usadas para estimar cuán diverso en el conjunto de clasificadores empleados. Los Algoritmos Genéticos juegan un papel significante como técnica de búsqueda para manejar espacios complicados en distintos campos de aplicación. Ellos están basados en el proceso genético que ocurre en los organismos vivos y en los principios de la evolución natural de las poblaciones. Estos algoritmos procesan una población de cromosomas, los cuales representan las soluciones del espacio de búsqueda, con tres operadores: selección, cruzamiento y mutación. Después de su formulación, el espacio de soluciones es codificado usando el alfabeto binario.

En este trabajo algunas medidas de diversidad son presentadas y una variante de Algoritmo Genético es implementada con el objetivo de obtener, de todas las posibles combinaciones de un número grande de clasificadores, una combinación que asegure simultáneamente la mayor diversidad entre los clasificadores escogidos y exactitud del sistema multiclasificador. Además se muestra una aplicación del algoritmo implementado para predecir el riesgo de hipertensión arterial en niños.

Palabras claves: Algoritmos Genéticos, Medidas de Diversidad, Clasificadores, Multiclasificador

Introducción

Los Algoritmos Genéticos (GAs) surgen como una herramienta para resolver problemas de optimización, como resultado del análisis de un sistema adaptativo en la naturaleza. Los métodos de búsqueda y optimización han sido estudios desde los primeros años de la computación, extendiéndose de los métodos basados en el cálculo para métodos enumerativos, hasta algoritmos de búsqueda aleatoria. Todos estos métodos son analizados y criticados en términos de robustez, pero esto no significa que no sean útiles; ellos pueden ser usados como complemento en esquemas más robustos con el objetivo de crear aproximaciones híbridas. El término de Algoritmo Genético es usado porque se simula el proceso de la evolución Darwiniano a través del uso de operador genéticos que trabajan en una población de individuos que "evolucionan" de una generación a otra. El desarrollo teórico concerniente a este tema no ha servido solamente en métodos eficientes de búsqueda, sino que ha permitido explicar resumida y rigurosamente el proceso adaptativo en sistemas naturales. Además, esto hace posible el diseño de sistemas artificiales que incluyen estos mecanismos naturales [1]

Por otra parte, el asunto de clasificación se ha discutido ampliamente y continúa desarrollándose. Escoger el mejor clasificador depende mayormente del problema a ser solucionado, para cada caso el clasificador seleccionado establece el mejor límite de decisión para separar las clases. En la búsqueda de los mejores métodos de clasificación hay una tendencia a combinar varios clasificadores para obtener la solución de mismo problema. Ésta es la idea en la que se basan los llamados sistemas multiclasificadores. Ellos usan varios clasificadores y

combinan sus salidas con afán de lograr un mejor resultado [2].

Dietterich [3] sugiere tres tipos de razones por las cuales un sistema multiclasificador puede ser mejor que un clasificador simple. La primera es estadística, pues si efectivamente por cada clasificador se tiene una hipótesis, la idea de combinar estas hipótesis, da como resultado una hipótesis que puede no ser la mejor, pero al menos evita seleccionar la peor de ellas. La segunda justificación es computacional, ya que algunos algoritmos ejecutan búsquedas que pueden llevar a diferentes óptimos locales: cada clasificador comienza la búsqueda desde un punto diferente y termina cercano al óptimo. Existe la expectativa de que alguna vía de combinación puede llevar a un clasificador con una mejor aproximación. La última justificación es figurativa ya que es posible que el espacio de hipótesis considerado no contenga la hipótesis óptima; pero la aproximación de varias fronteras de decisión puede dar como consecuencia una nueva hipótesis fuera del espacio inicial y que se aproxime más a la óptima.

Existen varias formas en las cuales se pueden construir multiclasificadores. Hay una serie de algoritmos desarrollados, algunos para problemas generales como bagging y boosting y otros para problemas específicos, pero todos tienen como partes fundamentales: la selección de los clasificadores de base y la elección de la forma de combinar las salidas [4].

La selección de los clasificadores de base es el primer paso a la hora de construir un multiclasificador. Entre las variantes para combinar los clasificadores existen algunos usados por bagging y boosting quienes usan el mismo modelo de clasificación entrenado con diferentes subconjuntos de casos. El primero hace una selección aleatoria de subconjuntos de casos y el segundo selecciona iterativamente subconjuntos basado en el resultado de la iteración previa. Otra variante es la usada por Stacking, quien usa modelos diferentes de clasificación entrenado con la misma base inicial.

Se pudiera decir que estos dos paradigmas son los más usados en general en la construcción de sistema multiclasificadores. Aunque no se ha demostrado cuál de las dos variantes es mejor. De igual forma, los sistemas multiclasificadores no son intrínsecamente mejores que los clasificadores individuales, le elección de uno u otro modelo deberá ser hecha para cada problema en específico [5].

La diversidad entre los clasificadores base es muy importante ya que de esto depende grandemente el resultado final del sistema multiclasificador. Cada clasificador consigue un porciento de casos correctamente clasificados. Mientras más diverso sean los clasificadores bases mayor probabilidad habrá de que cubrir un alto por ciento de casos bien clasificados, combinando la salida de los clasificadores base.

Algunos sistemas multiclasificadores aseguran la diversidad usando diferentes conjuntos bases de entrenamiento, pero esto solo funciona para clasificadores que son sensibles a los cambios como los árboles de decisión. Otros usan diferentes conjuntos de rasgos y por tanto, también varía la base de entrenamiento. Otros usan diferentes clasificadores bases. En este último caso, es importante conocer cuándo se garantiza gran diversidad, haciendo necesario el uso de varias medidas estadísticas que ayudan a determinar cuán diverso es un conjunto de clasificadores. Algunas medidas son descritas por Kuncheva en [6]. Ellas pueden ser clasificadas como: medidas en forma de pares (pairwise) y medidas grupales (nonpair-wise). En este trabajo solo se tendrán en cuenta las medidas en forma de pares debido a su sencillez.

Sección 1: Medidas en forma de pares (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 (Ci, Cj) para una instancia en cuanto si la clasificaron correctamente o no.

Si se suman para todas las instancias los valores de a, b, c, d entre el par de clasificadores (Ci, Cj) se obtendrán los resultados mostrados en la Tabla 2:

	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		

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 = N		

Tabla 2: Matriz binaria para N instancias

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

Coeficiente de correlación ρ

Entre las medidas de diversidad está el coeficiente de correlación [5] 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:

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

Para cualquier par de clasificadores, los valores de ρ y Q tendrán el mismo signo y se puede probar que $|\rho| \leq |Q|$ [6].

La medida de diferencias

La medida de diferencias (The Disagreement Measure) introducida por Skalak [7], es la más intuitiva de las medidas entre un par de clasificadores, y 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

Otra de las medidas que se analizará se conoce como medida de doble fallo (The Double- Fault Measure) introducida por Giacinto y Roli [8] considera el fallo de los dos clasificadores al mismo tiempo. Ruta y Gabrys [9] definen a esta medida como una medida no-simétrica. Esto quiere decir que si se intercambian los unos con los ceros en los resultados de los clasificadores, el valor de la medida no va a ser el mismo. Esta medida 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.

$$D_{ci,cj} = \frac{D}{N} \quad (4)$$

Seccion 2: Algoritmos Genéticos

Los algoritmos genéticos (AGs) son métodos de búsqueda basados en los principios generales de la genética natural, son algoritmos de búsqueda basados en los mecanismos de selección natural y la genética. Los algoritmos genéticos son un ejemplo de un método que sacan provecho de la búsqueda aleatoria "dirigida", que ha ganado popularidad en estos últimos años debido a su aplicabilidad en una gran variedad de campos y a los pocos requisitos impuestos por el problema [1, 10],

La idea básica es mantener una población de cromosomas, que represente las soluciones candidatas del problema concreto. Esta población evoluciona con el paso del tiempo a través de un proceso de competición y variación controlada. Cada cromosoma en la población tiene una adaptabilidad asociada para determinar cuáles cromosomas son usados para formar los nuevos cromosomas en el proceso de competición, el cual es llamado selección. Los nuevos son creados usando a los operadores genéticos como cruzamiento y la mutación. Los AGs han tenido un gran éxito en los problemas de búsqueda y de optimización. La razón de gran parte de su éxito es su habilidad para sacar provecho de la información acumulada acerca de un espacio de búsqueda inicialmente desconocido para influenciar subsiguientes búsquedas en subespacios útiles, es decir, su adaptación. Éste es su característica principal, particularmente en grandes, complejos y pobremente entendidos espacios de búsqueda, donde las herramientas clásicas de búsqueda (enumerativas, heurísticas) son inapropiadas, ofreciendo enfoque válido para problemas que requieren técnicas de búsquedas eficientes y efectivas.

Para usar los AGs es necesario encontrar una estructura para representar las soluciones posibles. Pensando en esto como el problema de búsqueda en un espacio establecido, una instancia de esta estructura representa un punto o una condición en el espacio de búsqueda de todas las soluciones posibles. Así, una estructura de datos consiste de uno o más cromosomas, quien es comúnmente representado como una cadena de bits. Cada cromosoma es una concatenación de un número de subcomponentes llamados genes. La posición de un gene en el cromosoma es conocida como el locus del alelo. En cadena de bits, un gen es un bit, el locus es la posición en la cadena y el alelo es su valor (0 o 1 si es un bit).

El tamaño fijo y la codificación binaria de las cadenas para la representación de las soluciones han

dominado la investigación de los AGs desde que hay resultados teóricos que los muestran como uno de los más apropiados [11], y fáciles para implementar.

Para optimizar la estructura de los AGs, una medida de la calidad de cada solución en el espacio de búsqueda es necesaria. La función de adaptabilidad es responsable de esta tarea. En una función de maximización, la función objetiva a menudo actúa como la función de adaptabilidad. Los AGs usualmente trabajan con funciones de maximización, para los problemas de minimización los valores objetivos de la función puede ser negados y transferido para tomar valores positivos, produciéndose adaptabilidad [11].

En esta metaheurística poblaciones de soluciones son construidas, ella es estocástica porque las probabilidades son usadas para tomar decisiones en el proceso de búsqueda, y por supuesto es bioinspirada porque proviene de un proceso natural.

El mecanismo simple de los AGs es el siguiente:

- Los AGs simples generan aleatoriamente una población de n estructuras (cadenas, cromosomas o individuos)
- Los operadores de la población actúan transformando la población. Una vez que la aplicación de estos operadores es completada, se puede decir que un ciclo generacional ha concluido.
- Entonces el paso anterior es repetido mientras que el criterio de parada del AG no esté garantizado.

El operador de selección hace la selección las cadenas según su adaptabilidad para los siguientes pasos. El operador de cruzamiento realiza la recombinación de material genético a partir de dos cadenas padre. El operador de mutación, al igual que la mutación natural, realiza la mutación de un gen dentro de un cromosoma.

Una probabilidad es asociada a cada uno de estos operadores. El modo de operación de una AG puede ser resumido como se muestra en la Figura 1. El AG se ejecuta para un número fijo de generaciones o hasta que algún criterio de parada es satisfecho.

La mayoría de expertos en este tema están de acuerdo en que los AGs pueden solucionar las dificultades representadas en los problemas reales de la vida que algunas veces no tienen solución por otros métodos. El foco de investigación en los AGs es la robustez: el balance entre la efectividad y la eficiencia necesitada para sobrevivir en muchos ambientes diferentes.

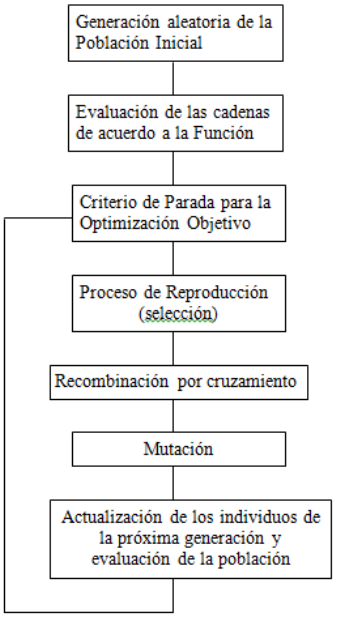

Figura 1: Diagrama funcional de un algoritmo genético

Sección 3: Algoritmo genético para detectar un buen sistema multiclasificador

La configuración del Algoritmo Genético depende del tipo de problema a resolver. Si la configuración y la representación de todos los elementos han sido definidas, entonces es necesario definir los operadores genéticos, los cuales son responsables de la evolución que cada población tendrá, proceso mediante el cual la solución esperada debería ser encontrada. Estos operadores tienen que responder a las restricciones del problema y por consiguiente, tendrán que ser adaptados en algunas ocasiones.

En nuestro caso, el algoritmo genético es presentado usando medidas de diversidad para 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:

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

Descripción de la función objetivo

En nuestro caso, se quiere 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. Luego, el valor de *f* es la suma entre la exactitud del sistema multiclasificador y el resultado de las medidas de diversidad, según la configuración del cromosoma.

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

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

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

En nuestro problema pueden existir casos en los cuales el primer parámetro de la función objetivo podría ser pequeño y el valor de $f(C_x)$ ser alto porque 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 encontrar la diversidad superior entre los clasificadores, entonces otra restricción se agrega. Esta restricción manifiesta que el resultado final será la combinación donde la exactitud del multiclasificador sobrepase la mejor exactitud obtenida con los clasificadores individuales; y 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.

Existen varios trabajos relativos a la influencia del tamaño de la población en la convergencia del AG. En principio, es lógico pensar que el trabajo con poblaciones pequeñas tiene el riesgo de representar pobremente el espacio de soluciones. Por otra parte, las poblaciones de gran tamaño consumen más tiempo computacional. En esta alternativa y como un trabajo teórico, Goldberg obtuvo en su investigación que el buen tamaño de una población de cadenas binarias, crece exponencialmente con la longitud de la cadena [2]. Sin embargo, en diferentes resultados empíricos muchos autores sugieren tamaños de poblaciones tan pequeños como de 30 individuos.

Se implementó el algoritmo para un tamaño de población, igual a $2^{\frac{L}{2}}$ dónde la 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, dónde cada valor del gen es 0 o 1 dependiendo de la presencia del clasificador en la combinación; es decir, un 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 el proceso de recombinación genética por medio del cruzamiento concluye, si los cromosomas nuevos ya existen en la población,

entonces una mutación es realizada 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 es definido: aleatoriamente se escoge un cromosoma, aleatoriamente se escoge un gen para mutar y cambiar su estado: 0 por 1 o 1 por 0, lo cual significa que la inclusión del clasificador cambia 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

En este proceso, una población intermedia de cromosomas es formada, donde los operadores previamente mencionados son aplicados para obtener una nueva población con cromosomas que tienen más calidad que los previos. Para la selección de los cromosomas que serán parte de la población intermedia es usada la función objetivo, la cual evaluada en cada uno de los individuos determinará su selección para participar en recombinación genética, los mejores cromosomas son seleccionados, es decir, los cromosomas que obtengan los valores más altos cuando la función objetivo es evaluada.

Luego, los operadores previos son aplicados en esta población intermedia para obtener nuevos cromosomas que se agregan a la población inicial. Ellos son añadidos en la población inicial porque para nuestro problema las combinaciones que no fueron seleccionada para la población intermedia, algunas veces pueden ofrecer mejores soluciones cuando son combinadas con los nuevos cromosomas y por consiguiente deben ser tenidos en cuenta. Ahora el tamaño de la población es mayor, denotándose su tamaño como *P'*.

Teniendo en cuenta las características del problema, en este proceso un paso más es añadido después de la recombinación, con el objetivo de obtener una población con el tamaño establecido. Esta reducción del tamaño de la población es realizada usando una selección nueva, aplicando 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:

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

Resumiendo, cada iteración simple del AG 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 al menos una de las siguientes condiciones sea cierto:

- 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.

Sección 4. Hipertensión Arterial 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 [12].

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 (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 y, por lo 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, intervenir a tiempo es un componente importante en la salud de los niños y adolescentes [13].

El primer paso es pensar en la HTA en niños; en segundo lugar, identificarla en los controles de salud, así como sus factores de riesgo (historia familiar de HTA, obesidad, enfermedades asociadas a HTA secundaria); y, en tercer lugar, una vez diagnosticada, saber la actitud a seguir.

Sección 5. Aplicación

La base de casos HTA-children fue usada como aplicación de este trabajo. Esta base es binaria, tiene siete rasgos nominales, 16 rasgos son numéricos y 626 instancias. Algunos de estos rasgos presentaban gran cantidad de valores perdidos, por lo que pensamos que los bajos por cientos de clasificación se deban a esto. Según la información contenida en la base y después de aplicar un análisis discriminante, las variables más importantes resultaron el color de la piel, el sexo y la actividad de catalasa. La base se obtuvo como resultado de un estudio aplicado para predecir el riesgo de que un niño sea o no hipertenso. El cálculo de la clasificación individual fue efectuado con los clasificadores existentes en la versión 3.7.5 del Weka. Los clasificadores individuales tenidos en cuenta en el estudio fueron:

Clasificador	Exactitud
NaivesBayes	0.6291
Functions.Logistic	0.6479
Lazy.IBK	0.5634
Trees.J48	0.6338
Multilayer Perceptron	0.5681
Trees. ADTree	**0.6620**
Functions.SGD	0.6244
Random Tree	0.6915
Functions.SMO	0.6197
Lazy.KStar	0.5775
Functions.VotedPerceptron	0.6009

Tabla 3: Exactitud de los clasificadores base

Como se observa en la tabla anterior, el mejor por ciento de clasificación obtenido por los clasificadores individuales no supera el 67% (0.66), luego fue aplicado el multiclasificador *Vote*, existente en la versión del Weka mencionada anteriormente, promediando las salidas de los clasificadores base. Los mejores resultados obtenidos se muestran en la Tabla 4, especificándose la configuración del cromosoma encontrada por el Algoritmo Genético.

El cromosoma resultante corresponde a la combinación de los siguientes clasificadores: *weka.classifiers.trees.J48, weka.classifiers.trees.RandomTree, weka.classifiers.lazy.KStar, weka.classifiers.functions.VotedPerceptron*. Este cromosoma provee una combinación de clasificadores que mejora a un 73% la exactitud del sistema multiclasificador, con respecto a los clasificadores individuales, nótese que aún no se logra un buen por ciento de casos correctamente clasificados pero se logra mejorar la clasificación individual en un 6%.

Medida de diversidad usada	Diversidad	$f(C_x)$	C_x	Exactitud del sistema
DF	0.8279	1.5603	00010001011	0.7324

Tabla 4: Resultados del Algoritmo Genético

Conclusiones

Este trabajo muestra una técnica novedosa usando 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 la diversidad entre los clasificadores individuales del sistema.

Un caso de estudio de la base de HTA es usado para ejemplificar esta contribución. Once clasificadores base fueron aplicados y sus resultados individuales no superan el 66%. Usando la propuesta del Algoritmo Genético con medidas de diversidad, se obtiene un

multiclasificador que logra mejorar en un 6% la clasificación anterior.

Referencias

[1] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning* vol. 412: Addison-wesley Reading Menlo Park, 1989.

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

[3] T. G. Dietterich, "*Ensemble methods in machine learning. Multiple Classifier Systems.*," *Berlin: Springer-Verlag Berlin.,* 2000.

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

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

[6] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. ," *Machine Learning,* vol. 51, pp. 181-207., 2003.

[7] D. B. Skalak, "The Sources of Increased Accuracy for Two Proposed Boosting Algorithms," 1996. .

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

[9] D. Ruta and B. Gabrys, "*Analysis of the Correlation Between Majority Voting Error and the Diversity*," 2001.

[10] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*: U Michigan Press, 1975.

[11] D. E. Goldberg, "*Real-coded genetic algorithms, virtual alphabets, and blocking*," *Urbana,* vol. 51, p. 61801, 1990.

[12] G. Mancia, G. De Baker, A. Dominiczak, R. Cifkova, R. Fargard, G. Germano, et al. *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; 25: 1105-87, 2007

[13] A. Ortigado, *Hipertensión arterial sistémica*. En: Del Pozo Machuca J, Redondo A, Gancedo MC, Bolívar V, eds. Tratado de Pediatría Extrahospitalaria. Madrid: Ergon; p. 455-62, 2011.