Algoritmo Genético y Testores Típicos en el Problema de Selección de Subconjuntos de Características

María Dolores Torres Soto
Departamento de Sistemas de Información
Universidad Autónoma de Aguascalientes
Aguascalientes, Ags. 20100, México
mdtorres@correo.uaa.mx

Aurora Torres Soto
Departamento de Sistemas Electrónicos
Universidad Autónoma de Aguascalientes
Aguascalientes, Ags. 20100, México
atorres@correo.uaa.mx

Eunice Ponce de León Sentí
Departamento de Sistemas Electrónicos
Universidad Autónoma de Aguascalientes
Aguascalientes, Ags. 20100, México
eponce@correo.uaa.mx

RESUMEN

Un tema común en diferentes investigaciones de años recientes, es el uso de heurísticas y metaheurísticas aplicadas al problema de selección del subconjunto de características que mejor describen un fenómeno, a partir de un conjunto mayor; pues estas técnicas nos permiten evitar la exploración exhaustiva de enormes espacios de soluciones, arrojando resultados muy cercanos al óptimo ó el óptimo mismo; con el consecuente ahorro de tiempo y recursos computacionales.

En este trabajo se presenta el resultado de la combinación del algoritmo genético simple, con el concepto de testor típico. El algoritmo diseñado fue empleado en la determinación de los factores de riesgo durante el embarazo, para calificar el estado de salud del recién nacido durante el primer mes de vida.

Como resultado de esta implementación, el algoritmo siempre encuentra una solución que corresponde a un testor típico, el cual coincide con el conjunto de factores de riesgo reconocidos por los expertos en medicina y reportados en la literatura de esta área.

Palabras Clave:
Testores típicos, Selección de Subconjunto de Características, Factores de riesgo durante el embarazo, Metaheurísticas.

INTRODUCCIÓN

El problema de selección de subconjuntos de características consiste en seleccionar un subconjunto de variables de un conjunto mayor, de tal forma que el subconjunto seleccionado es suficiente para ejecutar la tarea de clasificación y/o reducción del espacio de características [2]. Este problema es de especial interés, pues repercute directamente en el costo, eficiencia y precisión de la clasificación o descripción de los fenómenos en los que interviene un número considerable de variables. [8].

En este documento se presenta el resultado de la aplicación de un algoritmo genético modificado mediante el concepto de testor típico.

El algoritmo fue alimentado con un conjunto de datos reales correspondientes al historial ginecológico de 701 casos de mujeres embarazadas y su seguimiento durante el parto, así como

valoraciones referentes al estado de salud del recién nacido durante sus primeros 28 días de vida.

De acuerdo a Solís y sus colaboradores "aproximadamente dos tercios de las muertes neonatales, corresponden a las muertes neonatales precoces, reflejando principalmente problemas de calidad de atención del parto, asfixia y malformaciones inviables; el restante de muertes está ocasionado principalmente por problemas infecciosos, prematurez, y bajo peso al nacer" [14]; por lo que determinar los factores de riesgo durante el embarazo, puede permitir a los expertos de la salud, intervenir de manera oportuna, sencilla y con costos despreciables reduciendo este alto índice de mortalidad, como lo comenta Dawudo en su estudio: Neonatal mortality: Effects of selective pediatric interventions [6].

A mediados de los cincuentas, se comenzó a utilizar el concepto de testor típico en la detección de fallas en circuitos eléctricos, posteriormente éste se difundió a otras áreas de aplicación como clasificación supervisada y selección de variables en el área de la Geología [7].

El uso que damos al testor típico en este trabajo, se refiere a la selección de variables, ya que en esencia, un testor es un conjunto de características (rasgos) que distingue individuos (objetos) de clases distintas.

Desde los años 60's en los que se ubica el nacimiento de los algoritmos genéticos [9] en la comunidad científica promovidos por Holland, se han realizado diferentes esfuerzos para emplearlos en la solución de problemas de optimización, pues han demostrado ser muy eficientes y confiables cuando el espacio de búsqueda esta delimitado y es factible definir una función de aptitud. [12]

ALGORITMO GENÉTICO SIMPLE

Un algoritmo genético, es una estrategia de búsqueda en la que los estados sucesores se generan combinando dos estados anteriores (padres), permitiendo que los programas de computadora imiten el proceso de la selección natural, donde los individuos mas aptos (con mejor función de adaptabilidad), son los que con mayor probabilidad, se reproducen para transmitir sus características a las siguientes generaciones.

Los componentes de un algoritmo genético son:

1. Una representación para soluciones potenciales al problema que se pretende resolver.
2. Una manera de crear una población inicial de soluciones potenciales.
3. Una función de evaluación que juega el papel del entorno, ponderando las soluciones en términos de su adaptabilidad.
4. Operadores genéticos que alteran la composición de la descendencia.
5. Valores para diferentes parámetros que utiliza el algoritmo (Tamaño de la población, probabilidad de aplicar los operadores genéticos, etc..)

[5. Coello, 1996]

El algoritmo genético que se está poniendo a prueba, es el SGA de Goldberg con algunas modificaciones:

```
BEGIN /* SGA*/
  Generar una población inicial
  Computar la función de evaluación de cada individuo
  Escoger los 2 mejores individuos para la siguiente generación según su fitness
  Insertar los 2 individuos en la nueva generación
  While NOT termino do
    BEGIN /* Nueva generación*/
     FOR (tamaño de la población –2)/2 DO
        BEGIN /*Reproducción */
          Seleccionar 2 individuos de la generación anterior. (probabilidad de selección proporcional al fitness)
          Cruzar los 2 individuos seleccionados con cierta probabilidad, obteniendo dos descendientes.
          Mutar los 2 descendientes con cierta probabilidad.
          Computar la función de evaluación de los 2 descendientes.
          Insertar los 2 descendientes en la nueva generación
        END
      IF se han producido N generaciones THEN
        Termino := TRUE
      Escoger los 2 mejores individuos para la siguiente generación según su fitness
     Insertar los 2 individuos en la nueva generación
    END
END.
```

La generación, tanto de la población inicial como de las descendencias de ésta, están siendo construidas de manera que todos sus individuos son testores. La función de adaptabilidad utilizada, garantiza que

aquellas soluciones que correspondan a testores típicos, serán mejor evaluadas que las que no lo sean.

TESTOR TÍPICO

Supongamos que U es un conjunto de objetos descritos por N características, que están agrupados en K clases. En base a la comparación de cada característica de los objetos que pertenecen a una clase contra los que pertenecen a las demás (tomando dos objetos a la vez), se confecciona la matriz de diferencias MD. Esta matriz se construye mediante algún criterio de comparación por rasgos. En nuestro caso, la matriz fue construida empleando el criterio de comparación de igualdad estricta.

Una vez que se tiene la MD, se construye la matriz básica MB, que está constituida por todas las filas de MD que son básicas, es decir:
Una fila iq es básica si no existe fila alguna ip que sea subfila de iq.
Sean ip e iq filas de MD.
Decimos que ip es subfila de iq si para todo elemento de iq =0 se cumple que ip =0 y además, existe al menos un elemento de iq=1 en el que ip=0. [11].

El subconjunto de rasgos T de una matriz básica es un testor, si al eliminar de la MB todas las características, excepto las de T, no existe ninguna fila de ceros.

T es un testor típico, si al quitarle cualquiera de sus características, deja de ser testor.

EXPERIMENTO

Para construir la Matriz Básica (MB) que guió la generación de las soluciones en el algoritmo genético, así como la evaluación de la función de adaptabilidad, se utilizó la base de datos correspondiente a 701 casos de mujeres embarazadas con un total de 30 variables, de las cuales, la última es el descriptor de la clase.

La matriz básica obtenida, esta formada por 29 variables y 32768 filas, lo que implica un problema de tamaño considerable, ya que en investigaciones concernientes a testores típicos, el tamaño de las matrices de entrenamiento va de 40x42 a 1215x105 según Alba y colaboradores en 2000.

El problema se codificó utilizando una cadena binaria de 29 bits, donde 1 representa la presencia del factor de riesgo y 0 la ausencia del mismo. Los factores de riesgo considerados para este problema son los mostrados en la tabla 1. Donde el número menor representa al bit más significativo y el mayor al menos significativo.

Bit	Variable
1	Edad de la madre
2	Peso de la madre
3	Índice de masa corporal de la madre
4	Talla de la madre
5	Número de embarazos
6	Número de partos
7	Número de abortos
8	Número de cesáreas
9	Último período intergenésico
10	Toxemia
11	Polihidramnios
12	Sangrado
13	Hipertensión no totémica
14	Infecciones de vías urinarias
15	Antecedentes de malformaciones
16	Tabaquismo
17	Alcoholismo
18	Género del recién nacido
19	Peso del recién nacido
20	Índice de masa corporal del recién nacido
21	Talla del recién nacido
22	Edad de gestación
23	Tipo de nacimiento
24	Presentación
25	Tipo de parto
26	Apgar al minuto
27	Apgar a los 5 minutos
28	Forceps
29	Malformaciones

Tabla 1. "Variables consideradas para el problema de selección de características en morbilidad neonatal"

Los parámetros empleados en el Algoritmo Genético fueron los siguientes:

1. Tamaño de la población: 20
2. Número de generaciones 10
3. Probabilidad de cruzamiento 60%
4. Probabilidad de mutación 3.3%

El algoritmo se probó con diferentes tamaños de población (50, 30 y 20) y diferentes números de generaciones (30, 20 y 10); sin embargo, observamos que en todas las ejecuciones encontró

un testor típico de 22 variables como la mejor solución.

En vista de que el tiempo de ejecución es muy sensible a los parámetros comentados anteriormente, se seleccionaron los valores mínimos.

RESULTADOS Y CONCLUSIONES

El algoritmo genético se ejecutó en un tiempo de 17 segundos con 89 centésimas en una computadora con procesador Intel Pentium 4 a 3.2 GHz y 1GB de memoria Ram. Éste es un tiempo de ejecución muy bueno si consideramos que otras implementaciones que realizan esta tarea demoraron 1024 segundos [13], por otro lado, el algoritmo UMDA que reportan [1] tardó más de 3 horas en encontrar una cantidad considerable de testores típicos, mientras el algoritmo determinista DA tarda más de 2 días en ejecutar la misma tarea. A diferencia de nuestro algoritmo, éstos últimos pretenden encontrar todos los testores típicos de una matriz básica.

En todas las ejecuciones, el algoritmo encontró un testor típico constituido de 22 variables, el individuo que apareció en el mayor número de ejecuciones es el que se muestra en la figura 1.

1111110111010101111111101001

Figura 1. "Testor típico encontrado por el algoritmo"

El significado de esta cadena corresponde a la presencia de los factores de riesgo en el embarazo, que se muestran en la tabla 2.

Bit	Variable
1	Edad de la madre
2	Peso de la madre
3	Índice de masa corporal de la madre
4	Talla de la madre
5	Número de embarazos
6	Número de partos
8	Número de cesáreas
9	Último período intergenésico
10	Toxemia
12	Sangrado
14	Infecciones de vías urinarias
16	Tabaquismo
17	Alcoholismo
18	Género del recién nacido
19	Peso del recién nacido
20	Índice de masa corporal del recién nacido
21	Talla del recién nacido
22	Edad de gestación
23	Tipo de nacimiento
24	Presentación
26	Apgar al minuto
29	Malformaciones

Tabla 2. "Decodificación del testor típico encontrado".

Esto implica que el problema original, conformado de 30 variables, puede ser representado por únicamente 22 de ellas sin perder información relevante.

Todas las variables que constituyen el testor típico encontrado por el algoritmo, han sido reconocidas por diferentes fuentes como factores de riesgo en el embarazo.

Por ejemplo, el índice de masa corporal, que es una medida bruta del estado nutricional de una persona (IMC) y que corresponde a la razón de peso entre el cuadrado de la altura, se considera un factor de riesgo para el embarazo, pues de acuerdo a Jiménez y Gay (1997) "La malnutrición de la madre antes o durante el embarazo, contribuye al nacimiento de niños de bajo peso o con peso insuficiente" [10].

Por otro lado, la Organización Mundial de la Salud, y otras instancias expertas en el área, establecen que los límites de este indicador para una persona adulta son de 18.5 a 25. Por encima de este rango, se considera exceso de peso y por debajo de él, falta de peso.

También Chaviano Quesada encontró en su investigación "Edad materna, riesgo nutricional preconcepcional y peso al nacer" que en los casos de madres con IMC menor de 19.8 kg/m^2, los índices de bajo peso al nacer y peso insuficiente, son el doble con respecto a las madres con IMC mayor o igual que 19.8kg/m^2 [3].

Según varios investigadores, la primiparidad o el primer embarazo, está íntimamente relacionado con bajo peso al nacer y mortalidad neonatal, como es el caso de Cesar CLG, en su estudio "Factores de risco associados á mortalidade infantil em duas áreas da regiao metropolitana de Sao Paulo" [4]. Sin embargo, también la multiparidad se asocia frecuentemente con resultados adversos, como lo mencionan los autores de estudios sobre factores asociados con alto riesgo de mortalidad perinatal y neonatal.

Así como estas, las demás variables encontradas por el testor típico que arrojó el algoritmo, cuentan con

una justificación en el área médica que las reconoce como factores de riego en el embarazo.

Por último, es importante mencionar que cuando un algoritmo genético es armado con una estrategia de exploración del espacio de búsqueda de nivel superior, su eficiencia se incrementa considerablemente.

Es sabido que la debilidad de los algoritmos genéticos es su rápida convergencia a óptimos locales, aspecto que superamos en este trabajo mediante la conducción de la generación de soluciones a través de los conceptos de testor y testor típico.

REFERENCIAS

[1] Alba C. E, Santana R, Ochoa R. A, Lazo C. M. "Finding Typical Testors By Using an Evolitionary Strategy", Proceedings of V Iberoamerican Workshop on Pattern Recognition Lisbon, Portugal, 2000, pp. 267-278

[2] Blue, A.L. y Langley, P. (1997): Selection of revelant features and examples in machine learning", Artificial Intelligence, 97:245-271, 1997.

[3] Chaviano Q y López S. "Edad materna, riesgo nutricional preconcepcional y peso al nacer" \ Centro Provincial de Higiene y Epidemiología. Cuba. 2000

[4] Cesar CLG. "Factores de risco \ associados \`{a} mortalidade infantil em duas áreas da regiao metropolitana de Sao Paulo (Brasil)", 1984-1985.

[5] Coello, C. A.; "An empirical Study of Evolutionary Techniques for Multiobjetive Optimization in Engineering Design", Doctoral Thesis, 1996. Graduate School of Tulane University.

[6] Dawudo A y Effiong C, "Neonatal Mortality: Effects of Selective Pedistric Interventions" Pediatrics 75, 1985.

[7] Dmitriev, A. N.; Zhuravlev, Yu. I. and Krendeleiev, F. P. On the mathematical principles of patterns and
phenomena classification. Diskretnyi Analiz 7. pp. 3-15. Novosibirsk, Russia (In Russian), 1966

[8] García, L. F, Torres M.G, Moreno J.A, Moreno M. "Scatter Search for the Feature Selection Problem". Lecture Notes in Artificial Intelligence, 2004. Vol.3040, pp. 617-525

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

[10] Jiménez S. Gay J. "Vigilancia nutricional materno infantil. Guías para la atención primaria de salud". La Habana; Editorial Caguayo; 1997.

[11] Shulcloper, J. R, Alba C, Lazo C. "Introducción a la tería de testores típicos", Serie Verde No. 50, Cinvestav-IPN, México, 1995a.

[12] Torres A, Torres M.D, Ponce de León E. y Pinales F.J. "Algoritmo Genético para la Selección de Variables de Morbimortalidad Neonatal". Segundo Congreso de Computación Evolutiva COMCEV'05, Aguascalientes. México, 2005. pp. 79-83.

[13] Torres M.D, Torres A, Ponce de León, Ochoa A. "Búsqueda Dispersa y Testor Típico". 11° Simposio de Informática y 6° Mostra de Software Academico SIMS 2006. Uruguaiana, Brasil. Noviembre del 2006
Hífen, Uruguaiana, v. 30, n. 58, 2006. ISSN 0103-1155

[14] Solis F, Mardones G, Castillo B y Romer M.I. "Mortalidad por Inmadurez e hipoxia como causas de atención obstétrica y neonatal". Revista Chilena de Pediatría. 1993