

---

# Diseño de un algoritmo genético para generar conocimiento presuntivo del síndrome metabólico

---

## *Desing of a genetic algorithm to generate knowledge presumptive of the syndrome metabolic*

David Mauricio Sánchez, Luis Guerra, Rosa Delgadillo, Percy De la Cruz, Virginia Vera

Universidad Nacional Mayor de San Marcos  
Facultad de Ingeniería de Sistemas e Informática

dms\_research@yahoo.com

---

### RESUMEN

Recientes estudios de la World Health Organization [24] muestran que existen actualmente en el mundo 220 millones de personas con diabetes y que dicho número debe duplicarse para el 2030, de este número aproximadamente el 90% de pacientes tienen alta probabilidad de tener síndrome metabólico [10], siendo este último un síndrome cuyo protocolo de diagnóstico y tratamiento aún no es un estándar para las diversas instituciones internacionales que estudian dicho síndrome, por lo que se hace necesario el uso de técnicas no convencionales para su investigación. Una técnica no convencional e incipiente pero que está presentando buenos resultados en el diagnóstico de algunas patologías en medicina humana es machine learning [16, 2], sin embargo, a la fecha no existe su aplicación para el síndrome metabólico. En el presente trabajo se introduce un machine learning para generar conocimiento para el diagnóstico del síndrome metabólico, a través de una propuesta de algoritmo genético. La generación de conocimiento es realizada a fin de optimizar la tasa de verdaderos y falsos positivos, además, se introduce una representación cromosómica del conocimiento, y los operadores genéticos de crossover y mutation.

**Palabras clave:** Síndrome metabólico, aprendizaje automático, algoritmo genético, descubrimiento de conocimiento.

### ABSTRACT

Recent studies by the World Health Organization [24] shows that currently exist in the world 220 million people with diabetes and that number is going to double by 2030, of this number approximately 90% of patients have a high probability of having metabolic syndrome [10], being this syndrome whose diagnosis and treatment protocol is not yet a standard for international institutions who study this syndrome, so it is necessary to use unconventional techniques for its research. One unconventional technique and emerging, but is presenting good results in the diagnosis of certain diseases in human medicine is machine learning [15, 2], however, to the date there is not its application for metabolic syndrome. In this work is introducing a machine learning to create knowledge for the diagnosis of metabolic syndrome, through a proposed genetic algorithm. Knowledge generation is performed so as to optimize the rate of true and false positives, in addition, to introduce a chromosomal representation of the knowledge and genetic operators of crossover and mutation.

**Keywords:** Metabolic syndrome, machine learning, genetic algorithm, knowledge discovery.

---

## 1. INTRODUCCION

La medicina clínica está enfrentando el desafío de descubrir conocimiento a partir de los datos, cuyo volumen se vienen incrementando considerablemente. Las cantidades crecientes de datos de pacientes han hecho del análisis manual realizada por los expertos médicos una tarea tediosa y a veces imposible, haciendo que muchas relaciones ocultas y potencialmente útiles no puedan ser reconocidas por el analista. En ese contexto se vienen desarrollando, en primera instancia, los sistemas de información para el área de medicina y con ella el área de bio-informática también denominada de bio-computación [7], en segunda instancia se desarrollan sistemas basados en el conocimiento para el auxilio al diagnóstico y tratamiento de enfermedades (ver por ejemplo los trabajos publicados en [13, 14]), en tercera instancia se viene desarrollado técnicas de extracción de conocimiento a través de metodologías computacionales de minería de datos (selección, clasificación, comportamiento, pronósticos, tendencias, etc.), y con ello la denominada machine learning aplicada al área médica (ver por ejemplo [18]).

Uno de los problemas más importante de salud y de orden mundial es la diabetes. Según estimaciones de la World Health Organization existen en la actualidad alrededor de 220 millones de personas con diabetes, cifra que se estima doblar para el 2030

[24]. Un síndrome muy relacionado con la diabetes es el denominado síndrome metabólico (SM), diversos estudios muestran que el 90% de personas que padecen de diabetes presentan dicho síndrome [10], y a pesar de los grandes esfuerzos que se vienen realizando por diversas organizaciones internacionales como la World Health Organization (WHO), European Group for the Study of Insulin Resistance (EGIR), American Heart Association (AHA), a la fecha no existe un conocimiento unificado para su diagnóstico. Una revisión de machine learning en medicina es el trabajo de Kononenko [15]. Espejo [6] presenta una revisión de aplicaciones de programación genética y señala algunas aplicaciones en el área médica. Entretanto no existe a la fecha el uso de machine learning para el diagnóstico del síndrome metabólico, por lo que en el presente trabajo se propone un algoritmo genético como un método no tradicional para generar conocimiento para el diagnóstico del síndrome metabólico.

El presente trabajo se organiza como sigue: en la sección 2 se describe sucintamente el síndrome metabólico y su diagnóstico dado por diversas organizaciones internacionales, en la sección 3 se presenta la propuesta del algoritmo genético y sus operadores genéticos de casamiento y mutación, finalmente, en la sección 4 se presentan las conclusiones.

## 2 . EL SÍNDROME METABOLICO

El concepto de síndrome metabólico es quizá el acontecimiento más importante en el tratamiento de las enfermedades cardiovasculares (ECV) en los últimos 15 años. Antes de esto, los médicos con frecuencia trataban la diabetes, hipertensión o dislipidemia como enfermedades separadas y realmente no consideraban el impacto del tratamiento de una de estas condiciones sobre las otras condiciones coexistentes. Avogaro fue el primero en describir el síndrome hace más de 40 años, pero la prevalencia e importancia del concepto en la práctica clínica diaria fue primero colocado en relieve por Gerald Reaven en 1988, quien llamó la atención al mostrar características asociadas a la enfermedad cardíaca coronaria (ver Tabla 1). Reaven introduce el nombre de Síndrome X, también sugirió que la resistencia a la insulina juega un papel etiológico esencial para aportar un vínculo entre estos componentes [5].

La forma de presentación clínica del síndrome metabólico tiene una variación fenotípica. Puede manifestarse inicialmente con hipertensión u obesidad, y en otros como alteración de la regulación de la glucosa o de los lípidos u otros trastornos, como por ejemplo la asociación de hipertrigliceridemia y el perímetro de la cintura (con un punto de corte de 102 cm. en el hombre y de 88 cm. en la mujer) [17] [25].

En la década subsecuente, se ha hecho evidente que el síndrome X abarca mucho más alteraciones complejas que lo previsto originalmente (Tabla 1). Además, es bien conocido por diversos términos, como se muestra en la Tabla 2.

Se puede identificar tempranamente a los individuos portadores de una tríada metabólica aterogénica – hiperinsulinemia en ayuno, incremento de apolipoproteína B e incremento de la LDL pequeñas y densas que está asociado con un marcado incremento en el riesgo de enfermedad coronaria [11] [17] [25].

En la Tabla 3 se muestra las características clínicas antropométricas y bioquímicas que definen el síndrome metabólico según WHO.

Según la Revista Mexicana de Cardiología de Enero-Marzo del 2002 y el Consenso Mexicano sobre el Tratamiento Integral del Síndrome Metabólico, el diagnóstico clínico evidente del síndrome metabólico se basa en

los criterios propuestos por la WHO, y además en los diagnósticos de sospecha y de portador de dicho síndrome, ya que de acuerdo a la evidencia, los individuos de estos dos últimos grupos tienen un alto riesgo de desarrollar a futuro diabetes y enfermedades cardiovasculares y, por lo tanto, no pueden quedar excluidos de dicho diagnóstico [10].

Tabla 1  
El síndrome metabólico: 1988 y 2003 [5].

Metabolic syndrome 1988	Metabolic syndrome 2003
Resistance to insulin-simulated glucose uptake	Resistance to insulin-simulated glucose uptake
Glucose intolerance	Glucose intolerance
Hyperinsulinemia	Hyperinsulinemia
Increased VLDL triglyceride	Increased VLDL triglyceride
Decreased HDL cholesterol	Decreased HDL cholesterol
Hypertension	Hypertension
	Central obesity
	Microalbuminuria
	High plasminogen activator inhibitor-1
	Hyperleptinemia
	Hyperuricemia
	Hypoadiponectinaemia
	Subclinical information

Tabla 2.  
Sinónimos para el síndrome metabólico.

Metabolic syndrome
Syndrome X
CHO syndrome
Reavans syndrome
Chronic cardiovascular risk factor syndrome
Insulin resistance syndrome
Increased VLDL triglyceride
Decreased HDL cholesterol
Obesity syndrome
Hypertension

Tabla 3.  
Características que definen el síndrome metabólico – who, 1999 [26].

Hipertensión arterial:	Presión sistólica <sup>3</sup> 140 mm Hg
	Presión diastólica <sup>3</sup> 90 mm Hg o bien
Si tiene tratamiento antihipertensivo	
Dislipidemia:	Triglicéridos <sup>3</sup> 150 mg/dL
Colesterol HDL: Hombres <35 mg/dL, Mujeres < 45 mg/dL	
Obesidad:	Índice de masa corporal <sup>3</sup> 30 kg/m <sup>2</sup>
Relación cintura cadera: Hombres <sup>3</sup> 0.90, Mujeres <sup>3</sup> 0.85	
Microalbuminuria > 20 µg/minuto	
Trastornos en la homeostasis de la glucosa:	
En la glucemia en ayunas:	
Glucemia en ayunas alterada (GAA): 111 a 125 mg/dL	
Diabetes mellitus (ADA): <sup>3</sup> 126 mg/dL	
	En la curva de tolerancia a la glucosa:
	Intolerancia a la glucosa (ITG): 140 a 1999 mg/dL a las 2 horas
	Diabetes mellitus (OMS): <sup>3</sup> 200 mg/dL
Resistencia a la insulina: definida por el modelo homeostático para valorar resistencia a la insulina (HOMA) y se calcula con la fórmula siguiente:	
	$\text{HOMA IR} = \frac{\text{Insulina en ayunas (en mU por mL)} \times (\text{glucemia en ayuna (en mg/dL)} / 18)}{22.5}$
Criterios diagnósticos: se considera que una persona con glucemia en ayunas normal, con intolerancia a la glucosa o con diabetes mellitus tipo 2, tiene síndrome metabólico si se establece asociación con dos de los componentes anotados. Una persona con homeostasis normal a la glucosa (glucemia en ayunas de 70 a 110 mg/dl; glucemia posprandial < 140mg/dl a las dos horas en la curva de tolerancia oral a la glucosa) tendrá síndrome metabólico si se confirma el diagnóstico de resistencia a la insulina por el modelo HOMA, además de dos de los otros componentes señalados.	

El Grupo de Estudio de México (GEM) recomienda seguir los criterios dados en la Tabla 4 para el diagnóstico del síndrome metabólico [10].

La Asociación Americana del Corazón (AAC) indica que el síndrome metabólico es caracterizado por un grupo de factores de riesgo metabólicos en una persona [1], las mismas que se muestran en la Tabla 5.

En los Estados Unidos, de acuerdo con las definición de Síndrome metabólico hecha por la NCEP (siglas del inglés: National Cholesterol Education Program-Adult Treatment Panel III), alrededor del 25% de la población mayor de 20 años padece del síndrome metabólico [8]. En el Perú, reciente estudio muestra que 16.8% de la población de mayor de 20 años, esto es 2'680,000 personas presentan síndrome metabólico [19].

Tabla 4. Características que definen el síndrome metabólico – GEM, 2002 [10].

<p>Personas con antecedentes familiares de diabetes tipo 2 (padres, hermanos):</p> <ol style="list-style-type: none"> <li>1. Sedentarismo</li> <li>2. Tabaquismo</li> <li>3. Multiparidad o menopausia precoz</li> <li>4. Triglicéridos en ayuno &gt; 150 mg/dl</li> <li>5. Hiperinsulinemia</li> <li>6. Tensión arterial normal alta 130-134/85-89 mmHg</li> <li>7. Edad &gt; 45 años o menor de 45 años + otros marcadores</li> <li>8. Sobrepeso: IMC <math>\geq</math> 25 kg/m<sup>2</sup></li> <li>9. ICC en hombres &gt; 0.85</li> <li>10. ICC en mujeres &gt; 0.80</li> <li>11. Diámetro cintura hombres <math>\geq</math> 102 cm; mujeres <math>\geq</math> 88 cm</li> <li>12. Anovulación crónica y/o hiperandrogenismo</li> <li>13. Menos de 30 minutos de actividad física x 5 días por semana.</li> </ol> <p>Evaluación clínica integral:</p> <ol style="list-style-type: none"> <li>14. HDL &lt; 35 mg/dl en hombres</li> <li>15. HDL &lt; 45 mg/dl en mujeres</li> <li>16. Acantosis nigricans</li> <li>17. Poliquistosis ovárica</li> <li>18. Antecedentes de diabetes gestacional</li> <li>19. Productos macrosómicos</li> <li>20. Hiperuricemia</li> <li>21. Hiperinsulinemia</li> </ol> <p>Ambos: personas con antecedentes y evaluación integral</p> <ol style="list-style-type: none"> <li>22. Glucemia capilar en ayuno &gt; 95 mg/dl.</li> <li>23. Curva de intolerancia a la glucosa oral 140-199 mg/dl a las 2hrs.</li> <li>24. Intolerancia a la glucosa.</li> <li>25. Niveles séricos de insulina HOMA.</li> </ol>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Tabla 5. Características que definen el síndrome metabólico – AAC, 2010 [1].

<ol style="list-style-type: none"> <li>1. Obesidad central (tejido fino gordo excesivo en y alrededor del abdomen).</li> <li>2. Nivel de triglicéridos mayor o igual a 150 mg/dL</li> <li>3. HDL colesterol menor de 40 mg/dL para varón y 50 mg/dL para mujer.</li> <li>4. Presión arterial sistólica (130/85 mm Hg o más alto).</li> <li>5. Glucosa en ayunas mayor o igual a 100 mg/dL</li> <li>6. Resistencia a la insulina o tolerancia a la glucosa.</li> </ol> <p>Además, las causas subyacentes de este síndrome son: sobrepeso/obesidad, inactividad física y factores genéticos. Los pacientes con síndrome metabólico están en el riesgo creciente de la enfermedad cardíaca coronaria, de otras enfermedades relacionadas con las acumulaciones de la placa en las paredes de la arteria (movimiento y enfermedad vascular periférica) y de la diabetes de tipo 2.</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

### 3 ALGORITMO GENÉTICO PROPUESTO

#### 3.1 Hallazgos y su codificación

Un hallazgo o conjunto de hallazgos clínicos dan a conocer la naturaleza de una enfermedad o diagnóstico. En el síndrome metabólico se han identificado 32 hallazgos reconocidos por la WHO, el GESM y la AAC, las mismas que han sido codificadas y descritas en la Tabla 6. Observe que los valores 1 y 0 indican si un hallazgo está presente o no en el diagnóstico del síndrome metabólico para una de las instituciones mencionadas. Así por ejemplo, el hallazgo H1 es considerado por el GESM, pero no por la WHO y la AAC.

#### 3.2 Cromosoma

En los algoritmos genéticos, un cromosoma es la representación de un individuo y que es una solución (viable o no) al problema. En nuestro caso, un cromosoma estará dado por un conocimiento que permite diagnosticar el síndrome metabólico. En ese sentido, algunos individuos para nuestro problema son dados por los conocimientos establecidos por las diversas instituciones que investigan este síndrome. A su vez, el conocimiento para el diagnóstico médico engeneral se puede colocar como un conjunto de reglas de inferencia en donde los antecedentes están conectados por operadores "and", y donde cada regla corresponde a un conocimiento suficiente para el diagnóstico (esto es, no hay encadenamiento de reglas). cada regla corresponde a un conocimiento suficiente para el diagnóstico (esto es, no hay encadenamiento de reglas). Por otro lado, es fácil observar que una regla definida por an-

tecedentes conectados por operadores “and” es equivalente a una lista de dígitos binarios en donde cada dígito corresponde a un antecedente del conjunto de todos los posibles antecedentes, y cuyo valor es “1” o “0” si dicho antecedente se encuentra presente o no en

la regla respectivamente; en ese sentido, el hecho de que una regla se verifica (esto es, todos sus antecedentes se verifican) es equivalente a decir que todos los dígitos correspondiente a los antecedentes de dicha regla presentan valor 1

Tabla 6. Hallazgos para el síndrome metabólico.

Hallazgos		WHO	GESM	AAC
Cód.	Descripción			
H1	Paciente con antecedentes familiares de diabetes: padre y hermanos	0	1	0
H2	Varón	1	1	0
H3	Edad > 45 años	0	1	0
H4	Presión arterial elevada $\geq 130 / 85$ mmHg	1	1	1
H5	Colesterol Total $\leq 200$ mg/dL	0	0	0
H6	Triglicéridos en ayuno > 150 mg / d L	1	1	1
H7	Glucemia capilar en ayunas > 100 mg/dL (5.6 mmol/L)	1	1	1
H8	Curva de tolerancia a la glucosa oral 140-199 mg / dL a las 2 hrs	1	1	0
H9	Intolerancia a la glucosa	1	1	1
H10	HOMA IR=(Insulina en ayunas)(en $\mu$ m/ml)xglucemia en ayunas (en Mg/d L)/18)/22.5	1	1	0
H11	Hiperuricemia	0	1	0
H12	Hiperinsulinemia	0	1	0
H13	Poliquistosis ovárica	0	1	0
H14	HDL $\text{♂} < 40$ mg / d L	1	1	1
H15	HDL $\text{♀} < 50$ mg / d L	1	1	1
H16	LDL Colesterol $\leq 165$ mg/dL	0	0	0
H17	VLDL Colesterol	0	0	0
H18	ICC $\text{♂} \geq 90$ cm.	1	1	1
H19	ICC $\text{♀} \geq 80$ cm.	1	1	1
H20	IMC $\text{♂} \geq 30$ Kg / m 2	1	1	1
H21	IMC $\text{♀} \geq 27$ Kg / m 2	1	1	1
H22	Acanthosis nigricans	0	1	0
H23	Antecedentes diabetes gestacional.	0	1	0
H24	Productos Macrosómicos	0	1	0
H25	Multiparidad o menopausia precoz	0	1	0
H26	Microalbuminuria > 20 $\mu$ g / minuto	1	0	0
H27	Diabetes mellitus ADA $\geq 126$ mg / dL	1	1	0
H28	Diabetes mellitus OMS $\geq 200$ mg / dL (glucemia)	1	0	0
H29	Estado de prothrombotic	0	0	1
H30	Estado de proinflammatory	0	0	1
H31	Tabaquismo	0	1	0
H32	Sedentarismo. Menos de 30' de actividad física x días x semana	0	1	0

Por lo expuesto, podemos representar el conocimiento asociado al síndrome metabólico como un conjunto de listas de dígitos binarios, donde cada lista corresponde a una regla de dicho conocimiento; y cada dígito corresponde a un hallazgo, cuyo valor es "1" o "0" si dicho hallazgo se encuentra presente o no respectivamente en dicha regla. En la Figura 1 se muestra un cromosoma conformado por 3 reglas.

	R1	R2	R3
H <sub>1</sub>	0	1	0
H <sub>2</sub>	1	1	1
H <sub>3</sub>	1	0	0
H <sub>4</sub>	1	0	0
...	...	...	...
H <sub>32</sub>	0	1	0

Figura 1. Cromosoma correspondiente a un conocimiento conformado por 3 reglas.

Consideremos que el conocimiento es dado por un conjunto de "k" reglas  $R_1, R_2, \dots, R_k$ , entonces el cromosoma puede estar dado por una matriz  $C = [c_{ij}]_{32 \times k}$

donde:

$$c_{ij} = \begin{cases} 1, & \text{si } H_i \text{ se encuentra en } R_j \\ 0, & \text{en caso contrario} \end{cases}$$

### 3.3 Fitness

Denotemos por n el número de pacientes (experiencias)  $E_1, E_2, \dots, E_n$ , m el número de hallazgos, y por  $E = [e]_{m \times n}$  la matriz de ocurrencias de hallazgos en los pacientes (historia clínica), en donde:

$$e_{ij} = \begin{cases} 1, & \text{si el paciente } E_j \text{ presenta} \\ & \text{el hallazgo } H_i \\ 0, & \text{en caso contrario} \end{cases}$$

Denotemos también por  $D = [d]_n$  y  $S = [s]_n$  el vector de diagnóstico del experto y del sistema respectivamente asociado a los n pacientes, en donde:

$$d_j = \begin{cases} 1, & \text{si el médico diagnostica} \\ & \text{síndrome metabólico para } E_j \\ 0, & \text{en caso contrario} \end{cases}$$

$$s_j = \begin{cases} 1, & \text{si el sistema diagnóstica} \\ & \text{síndrome metabólico para } E_j \\ 0, & \text{en caso contrario} \end{cases}$$

Una función muy usada para evaluar sistemas inteligentes para el área médica es el índice de acuerdo (ver por ejemplo [2], [6], [23]); usaremos esta función para evaluar el conocimiento generado por el sistema. En ese sentido, deseamos que el sistema presente la mayor tasa de aciertos, este indicador conocido en la literatura como índice de acuerdo se obtiene como:

$$\text{Índice de acuerdo} = \frac{a + d}{a + b + c + d}$$

En donde:

- a: número de verdaderos positivos
- b: número de falsos positivos
- c: número de falsos negativos
- d: número de verdadero negativos

Para nuestro caso en particular, los valores a, b, c, y d se pueden obtener como sigue:

$$a = \sum_{j=1}^n d_j s_j$$

$$b = \sum_{j=1}^n (1 - d_j) s_j$$

$$c = \sum_{j=1}^n d_j (1 - s_j)$$

$$d = \sum_{j=1}^n (1 - d_j) (1 - s_j)$$

Observe que  $a + b + c + d$  es el número total de instancias de pruebas, esto es n, por lo que la función fitness es dado por:

$$f(C, E, D) = \frac{n + \sum_{j=1}^n (2d_j s_j - d_j - s_j)}{n}$$

Observe que D es el vector de diagnóstico del experto y es conocido, y S es el vector de diagnóstico obtenido por el sistema y depende de C y E. Observe también que el sistema diagnóstica que un paciente  $E_i$  tiene síndrome metabólico si alguna regla de C se verifica, esto es, si existe  $\exists l \in \{1, \dots, k\}$  tal que  $c_{il} = e_{il} \quad \forall i \in \{1, \dots, m\}$  donde  $c_{il} = 1$ , en otras palabras:

$$s_i(C, E_i) = \begin{cases} 1, & \text{Si } \exists l \in \{1, \dots, k\} \text{ tal que } c_{il} = e_{il} \\ 0, & \text{Si } \forall i \in \{1, \dots, m\} \text{ donde } c_{il} = 1 \\ & \text{caso contrario} \end{cases}$$

Entonces el problema de optimización correspondiente consiste en encontrar un conocimiento C que maximice el índice de acuerdo. Observe también que  $0 \leq f(D, S) \leq 1$  y los valores próximos a 1 significa un conocimiento muy bueno.

### 3.4 Algoritmo Genético

A continuación se propone un algoritmo genético para generar conocimiento sobre el diagnóstico del síndrome metabólico.

```

Algoritmo GCSM
1. read (parameters)
2. Choose an initial population  $(C^1, C^2, \dots, C^t)$ 
3. while  $\neg$  (Stop condition)
4. repeat
5. If (crossover condition satisfied) then
6. Select  $(C^u, C^v)$ 
7. Crossover  $(C^u, C^v, \bar{C}^u, \bar{C}^v)$ ;
8. If (mutation condition satisfied) then
9. Select  $(C)$ 
10. Mutation  $(C)$ ;
11. Evaluate fitness offspring created  $(\bar{C}^u, \bar{C}^v, C)$ ;
12. until sufficient offspring created;
13. Select new population
14. end -while
    
```

El algoritmo inicia con la lectura de parámetros (paso1). En el paso 2 se elige una población inicial de cromosomas. El proceso de evolutivo se realiza en los pasos 3-14, hasta que determinada condición de parada sea verificada (paso 3). Cada iteración del proceso de evolución consiste en la posible aplicación de los operadores de crossover (casamiento y generación de

dos hijos – pasos 5-7) y mutation (mutación de algunos códigos de un cromosoma seleccionado – pasos 8-10); este proceso se repite hasta que el tamaño de la población de hijos tenga el tamaño de la población inicial (paso 12). El último paso del proceso evolutivo consiste en determinar la nueva población desde la población obtenida en la iteración anterior y la población de hijos obtenida en la presente iteración (paso 13). El algoritmo termina retornando la mejor solución obtenida (paso 15).

A continuación se describe las condiciones y los operadores mostrados en el algoritmo propuesto.

### 3.5 Descripción del algoritmo

Parameters

- t:** Tamaño de la población
- m:** Número de hallazgos
- n:** Número de pacientes
- k:** Número de reglas para cada conocimiento
- x:** Probabilidad para aplicar crossover
- μ:** Probabilidad para aplicar mutation

#### Choose an initial population

Considerando que existe un conocimiento establecido por diversas organizaciones internacionales (OMS, GESM y AAC), se construirá una población inicial a partir de estos conocimientos, de la siguiente forma:

```

Choose an initial population  $(C^1, C^2, \dots, C^t)$ 
1. read  $(t)$ ;
2.  $R \leftarrow \emptyset$ ;
3. for each organization
4. Compute rules
5.  $R \leftarrow R + rules$ 
6. end -for;
7. for  $(i = 1, \dots, t)$ 
8. for  $(j = 1, \dots, k)$ 
9.  $rule \leftarrow Random(R)$ ;
10.  $C_j^i \leftarrow rule$ ;
11. end -for;
12.  $C^i \leftarrow (C_1^i, C_2^i, \dots, C_k^i)$ 
13. end -for;
14. return  $(C^1, C^2, \dots, C^t)$ .
    
```

Comenzamos con la lectura del tamaño de la población (paso 1). Luego inicializamos el conjunto de reglas (paso 2). Para cada organización (OMS, GESM y AAM) generamos sus reglas y lo registramos en  $R$  (pasos 3-6). Seleccionamos al azar sin reposición  $k$  reglas de  $R$  y con ellos conformamos un conocimiento (pasos 8-12), y repetimos este proceso hasta completar la población requerida (pasos 7). Observe que  $C_j^i$  corresponde a una regla del conocimiento  $C^i$ .

Con el fin de cubrir el espacio de búsqueda, usaremos el principio de Reeves [20], que se expresa para nuestro estudio como: la probabilidad  $p^*$  que al menos un alelo está presente en cada locus (esto es, que en el proceso de búsqueda todos los posibles valores genéticos puedan ser encontrados) es dado por:

$$p^* = \left( 1 - \left( \frac{1}{2} \right)^{t-1} \right)^{k \times m}$$

De donde basta considerar una población de tamaño  $t = 15$ , y  $k = 10$  para garantizar una cobertura del espacio de búsqueda con probabilidad al menos del 98.12%.

**Stop conditions**

Existen varias condiciones de parada para los algoritmos genéticos que bien se pueden aplicar en forma combinada para nuestra propuesta, ellos son:

- Máximo número de iteraciones ejecutadas*
- Tiempo de procesamiento límite*
- Diversidad de la población inferior a un límite establecido*

**Crossover and mutation condition satisfied**

El algoritmo GCSM presentado en 3.4 considera en cada iteración la aplicación de los operadores crossover y/o mutation. La condición de la aplicación de tales operadores es dado por:

Crossover and mutation conditions

```

1.  $r \leftarrow \text{random}(0,1)$ 
2. If  $r \geq x$ 
3. then crossover condition is satisfied;
4. If  $r \geq \mu$ 
5. then mutation condition is satisfied;
```

Es deseable que el parámetro  $x$  presente un valor próximo a 1 al inicio, y que su valor disminuya progresivamente a medida que se llegue a las últimas iteraciones. Al contrario, el parámetro  $\mu$  debe presentar un valor próximo a cero al inicio, y debe aumentar progresivamente a medida que se llegue a las últimas iteraciones. Los valores de convergencia de los parámetros  $x$  y  $\mu$  deberán ser mayor y menor a 0.5, respectivamente.

**Select ( $C^u, C^v$ )**

Existen varias estrategias para la selección de dos individuos de la población, entre ellas destacan las estrategias: (1) la regla de la ruleta, y (2) el ranking. La selección de una de estas estrategias dependerá de los resultados de las pruebas numéricas. Si los resultados numéricos muestran sustancial diferencias en los valores fitness de los individuos de la población, entonces es recomendable el uso de la primera estrategia, de lo contrario deberá aplicarse la segunda estrategia.

Roulette -wheel

```

1.  $A \leftarrow 0; \quad \bar{p}_i \leftarrow 0; \quad i \leftarrow 1;$ 
2. for  $i = 1, \dots, t$ 
3.    $A \leftarrow A + f(C^i, E, D);$ 
4. end -for
5. for  $i = 1, \dots, t$ 
6.    $p_i \leftarrow \frac{f(C^i, E, D)}{A};$ 
7.    $\bar{p}_i \leftarrow \bar{p}_i + p_i;$ 
8. end -for
9.  $r \leftarrow \text{random}(0,1);$ 
10. while  $(r < \bar{p}_i)$ 
11.  $i \leftarrow i + 1;$ 
12. end -while
13. return ( $C^i$ )
```

Goldberg [9] propone un simple algoritmo de ranking aplicable cuando la variación del fitness para los individuos de una población es pequeña respecto a la escala de sus valores.



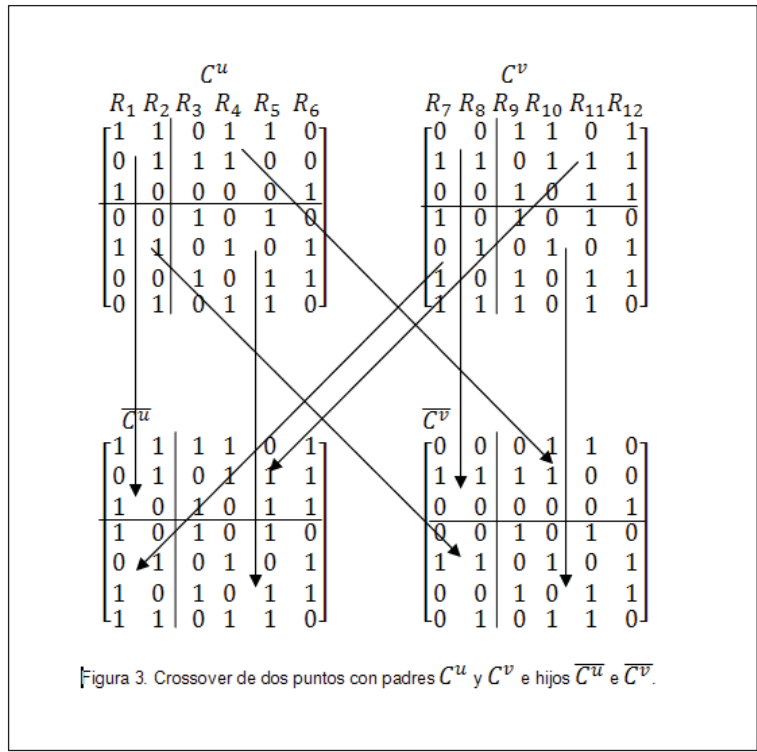
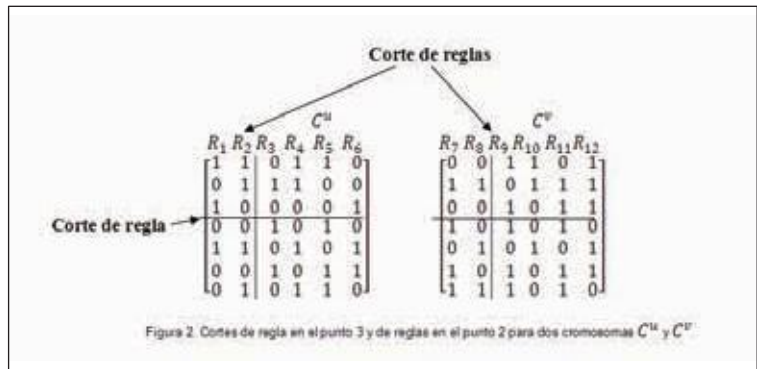
```

Ranking
1.  $A \leftarrow 0;$ 
2. for  $i = 1, \dots, t$ 
3.  $A \leftarrow A + f(C^i, E, D);$ 
4. end -for
5. for  $i = 1, \dots, t$ 
6.  $p_i \leftarrow \frac{f(C^i, E, D)}{A};$ 
7. end -for
8. sort  $(p_1, p_2, \dots, p_t)$  such that
9.  $p_1 \leq p_2 \leq \dots \leq p_t;$ 
10.  $\varphi \leftarrow \frac{p_t}{p_1};$ 
11.  $\alpha \leftarrow \frac{2t - \varphi(t+1)}{t(t-1)}; \quad \beta \leftarrow \frac{2(\varphi-1)}{t(t-1)};$ 
12.  $r \leftarrow \text{random}(0,1);$ 
13.  $i \leftarrow \frac{-(2\alpha+\beta) \pm \sqrt{(2\alpha+\beta)^2 + 4\beta r}}{2\beta};$ 
14. return  $(C^i)$ 
    
```

**Crossover ( $C^u, C^v, \bar{C}^u, \bar{C}^v$ ):**

Usaremos el cruzamiento de dos puntos, uno relacionado al corte de la regla y otro relacionado al corte de las reglas, como se muestra en el siguiente ejemplo:

Considerando  $C^u$  y  $C^v$  como los cromosomas padres, el cruzamiento de dos puntos que estamos adoptando generará dos cromosomas hijos  $\bar{C}^u$  y  $\bar{C}^v$  que se obtendrá combinando parte de las reglas de los cromosomas padres, como sigue:



Los puntos de corte se podrán obtener en forma probabilística y a través de experimentos computacionales.

#### Mutation ( $C$ );

Usaremos la tasa de mutación óptima propuesta por

Bremermann [3], esto es  $\mu = \frac{1}{m} \times k$ , para la aplicación del operador mutation. Sea  $C$  un cromosoma al cual se le aplicará el operador mutation, el procedimiento consiste en seleccionar en forma aleatoria una posición de  $C$  y cambiar el bit de dicha posición de 0 por 1 o 1 por 0, esto se hace con la finalidad de generar diversidad en la población.

Mutation ( $C$ )

1.  $i \leftarrow \text{random}(1, m);$
2.  $j \leftarrow \text{random}(1, k);$
3.  $c_{ij} \leftarrow 1 - c_{ij};$
4. return ( $C$ ).

#### Sufficient offspring created

Limitaremos el número de hijos creados en cada iteración (evolución) del algoritmo por el tamaño de la población de la población inicial, esto es por  $t$ .

#### Select new population

Consideraremos la estrategia usada por Reeves [21] para resolver problemas combinatorios, esto es, la nueva población estará dada por el 50% de los mejores individuos de la población obtenida en la población anterior y de los individuos recientemente creados en la presente iteración.

## 4. CONCLUSIONES

Los diversos estudios sobre el síndrome metabólico revisado en la sección 3 muestran que no existe un conocimiento unificado para su diagnóstico dado por las diversas organizaciones internacionales que estudian este síndrome, a pesar que sus primeros estudios son publicados hace 90 años [12, 16]. Siendo esta la razón por la cual se hace necesario establecer nuevos mecanismos no convencionales para generar conocimiento para su diagnóstico.

Se demuestra que una alternativa no convencional para generar conocimiento en el área médica es machine learning, entretanto no se ha identificado un trabajo de machine learning para el diagnóstico y/o tratamiento del síndrome metabólico.

El presente trabajo introduce un machine learning como una alternativa no convencional para generar conocimiento sobre el diagnóstico del síndrome metabólico. Dicha propuesta está basada en la técnica de optimización de algoritmo genéticos, y se busca a través de ella generar un conocimiento que maximiza la tasa de verdaderos y falsos positivos. La propuesta también introduce una forma genética de representación del conocimiento, y los operadores genéticos de crossover y mutation.

#### AGRADECIMIENTO

El responsable del proyecto agradece al profesor Marcos Negreiros, profesor del Dpto. de Computación de la Universidad de Ceará – Brasil.

#### REFERENCIAS

- [1] American Heart Association. About metabolic syndrome, [http://www.heart.org/HEARTORG/Conditions/More/MetabolicSyndrome/About Metabolic - Syndrome\\_UCM\\_301920\\_Article.jsp](http://www.heart.org/HEARTORG/Conditions/More/MetabolicSyndrome/About%20Metabolic%20Syndrome_UCM_301920_Article.jsp). Visited 15.12.2010.
- [2] Beligiannis Grigorios, Hatzilygeroudis Ioannis, Koutsojannis Constantinos, Prentzas Jim. A GA Driven Intelligent System for Medical Diagnosis. In Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Springer 2006, Volume 4251/2006, 968-975.
- [3] Bremermann H.J., Rogson J. and Salaff S. (1964) Global properties of evolution processes. In H.H. Pattee (ed.), Natural Automata and Useful Simulations, pp. 3-42.
- [4] Bojarczuk C.C., Lopes H.S., Freitas A.A., and Michalkiewicz E. L., A constrained-syntax genetic programming system for discovering classification rules: Application to medical data sets, Artificial Intelligence in Medicine, 30 (1), pp. 27-48, 2004, ISSN 0933-3657.
- [5] Creely S.J., Anwan A.J., Kumar S. The Metabolic Syndrome and Vascular Disease. Diabetes and

- Cardiovascular Disease. Edit. Michael T. Johnstone and Arisitidis Veves. Humana Press, Second edition, 2005; 281 - 306.
- [6] Espejo P.G., Ventura S., and Herrera F., A Survey on the Application of Genetic Programming to Classification. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Review*, 40(2), March 2010.
- [7] European Bioinformatics Institute (2006). «What is Bioinformatics?». <http://www.ebi.ac.uk/>. Visited 30-12-2010.
- [8] Ford ES, Giles WH, Dietz WH. Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey. *JAMA*. 2002;287:356-359.
- [9] Goldberg D.E. (1989) Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, Massachusetts.
- [10] Grupo de Estudio del Síndrome Metabólico. Consenso Mexicano sobre el Tratamiento Integral del Síndrome Metabólico. *Rev. Mex. Cardiología* 2002; 13 (1), 4-30.
- [11] Han TS. Waist Circumference Action Levels in the Identification of Cardiovascular risk factors; prevalence study in a random sample. *BMJ* 1995; 311: 1401-1405.
- [12] Joslin EP. The prevention of diabetes mellitus. *JAMA* 1921;76: 79–84.
- [13] Journal of Expert System. <http://www.blackwellpublishing.com/journal.asp?ref=0266-4720>. Visited 30-12-2010.
- [14] Journal of Expert System with Applications. [http://www.elsevier.com/wps/find/journaldescription.cws\\_home/939/description#description](http://www.elsevier.com/wps/find/journaldescription.cws_home/939/description#description). Visited 30-12-2010.
- [15] Kononenko Igor, Machine learning for medical diagnosis: history, state of art and perspective. *Artificial Intelligence in Medicine* 23 (2001) 89-109.
- [16] Kylin E. [Studies of the hypertension-hyperglycemia-hyperuricemia syndrome] (German). *Zentralbl Inn Med* 1923;44: 105-27.
- [17] Lean Me. Waist Circunference as a Measure for Indicating need for Weight Management. *BMJ* (1995); 311: 158-161.
- [18] Nikhil R. Pal and Lakhmi Jain. *Advanced Techniques in Knowledge Discovery and Data Mining*. Springer. 2005, DOI:10.1007/1-84628-183-0
- [19] Pajuelo Jaime, Sánchez José. El síndrome metabólico en adultos, en el Perú. *Anales de la Facultad de Medicina, Universidad Nacional Mayor de San Marcos*. ISSN 1025-5583, pp. 38-46.
- [20] Reeves C.R. (1993) Using genetic algorithms with small populations. In S. Forrest (ed.) *Proceeding of 5th International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, C.A. pp. 92-99.
- [21] Reeves C.R. (1995) A genetic algorithm for flowshop sequencing. *Computers & Operations Research*, 22, 5-13.
- [22] Shen S., Sandham W., Granat M.H., Dempsey M. F., and Patterson J., A new approach to brain tumour diagnosis using fuzzy logic based genetic programming, in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 1, Piscataway, NJ: IEEE, Sep. 2003, pp. 870–873.
- [23] Tan K.C., Yu Q., Heng C.M., Lee T.H., Evolutionary computing for knowledge discovery in medical diagnosis. *Artificial Intelligence in Medicine*, 27 (2003) 129-154.
- [24] World Health Organization, World diabetes day. [http://www.who.int/mediacentre/events/annual/world\\_diabetes\\_day/en/index.html](http://www.who.int/mediacentre/events/annual/world_diabetes_day/en/index.html). Visited 14-11-2010.
- [25] World Health Organization. Obesity: preventing and managing the global epidemic. Report of WHO consultation. *World Health Organ Tech Rep Ser*. 2000; 894:i-xii, 1-253. <http://www.ncbi.nlm.nih.gov/pubmed/11234459>. Visited 15.12.2010.
- [26] World Health Organization. Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Part 1: diagnosis and classification of diabetes mellitus. *Technical Report WHO/NCD/NCS/99.2*. (1999).
- [27] Yu J., Yu J., Almal A.A., Dhanasekaran S.M., Ghosh D., Worzel W.P., and Chinnaiyan A.M., Feature selection and molecular classification of cancer using genetic programming, *Neoplasia*, 9(4), pp. 292–303,(2007).

