OPTIMIZACIÓN BASADA EN ENJAMBRES DE PARTÍCULAS PARA AJUSTAR LOS PARÁMETROS DE LOS MÉTODOS SCAN

Laureano Rodríguez Corvea[*1], Yunier Rojas García[**2], Gladys Casas Cardoso[***3]
[*] Instituto Superior de Ciencias Médicas, Sancti Spíritus, Cuba
[**]Desarrollo de Aplicaciones, Tecnologías y Sistemas, Santa Clara, Villa Clara, Cuba
[***] Laboratorio de Bioinformática, Universidad Central Marta Abreu de Las Villas, Cuba

ABSTRACT
Classic and fuzzy Scan methods are widely used for cluster detection over a lineal or circular data sequence. This sequence has been previously processed and transformed into a binary sequence. The value number one represents the category that is of interest and the zero value represents the rest. The objective of these methods is to detect a cluster of ones. These methods depend on some parameters: the width of the mobile window that covers the entrance sequence, the step of the window's movement and the size of the fuzzy part in the Fuzzy Scan methods. In this paper it is shown how with the combination of a bioinspired algorithm (particle swarm optimization) is useful to find the adequate values of the parameters of the Scan techniques. Also, a Bioinformatics problem is solved as an example**.**

KEY WORDS: Scan techniques, Fuzzy Scan techniques, particle swarm optimization.

MSC: 62P10, 90C70 y 90C59

RESUMEN
Los métodos Scan clásicos y borrosos se utilizan ampliamente para la detección de conglomerados sobre una secuencia linear o circular de datos. Tal secuencia ha sido procesada previamente y transformada en una secuencia binaria. El valor uno representa la categoría de interés y el valor cero representa lo demás. El objetivo de los métodos es detectar un conglomerado de unos. Estos métodos dependen de varios parámetros: el ancho de la ventana móvil que recorre la secuencia de entrada, el paso con el que la ventana se mueve y el tamaño de la parte borrosa en las variantes borrosas de los métodos Scan. En este trabajo se muestra como la unión de un algoritmo bioinspirado: la optimización mediante enjambre de partículas, es útil para encontrar valores adecuados para los parámetros de las técnicas Scan. Además se resuelve un problema de bioinformática a manera de ejemplo.

1. INTRODUCCIÓN

Los métodos Scan se utilizan para detectar conglomerados. Inicialmente surgieron para resolver un problema epidemiológico: la determinación de si las fechas de diagnósticos o de primeros síntomas de un conjunto de pacientes estaban más próximas que lo que podía esperarse por la casualidad [3]. Si la respuesta era positiva, había una alarma epidemiológica con la enfermedad en cuestión.

Los métodos Scan se han generalizado y han aumentado notablemente su campo de aplicación. Ya no se habla de detectar conglomerados de enfermos, sino conglomerados de cualquier tipo en las distintas ramas del saber. Para generalizar supóngase que se tiene una secuencia de ceros y unos, donde los unos tienen cierta importancia (categoría de interés). El objetivo de los métodos Scan es entonces detectar conglomerados de unos. Ejemplos concretos los podemos encontrar en las aplicaciones bioinformáticas [9, 10, 11].

La probabilidad *p* hallada para un caso particular, depende de sus parámetros seleccionados por el investigador. Resulta imposible determinar los valores ideales para cada parámetros en los diferentes problemas, por lo que se propone utilizar un método de optimización que utilice el Método Scan como función objetivo, en particular la función relacionada con el cálculo de la significación p según fórmulas publicadas en [8].

[1] corvea@uclv.edu.cu
[2] yunier.rojas@datys.cu
[3] gcasas@uclv.edu.cu

2. EL MÉTODO SCAN CLÁSICO Y SU GENERALIZACIÓN

Comienza definiendo un intervalo o ventana de tamaño fijo (parámetro más importante) que se mueve discretamente, según el valor de un determinado paso (segundo parámetro a considerar), por la línea del tiempo [3] o por la secuencia binaria en el caso del Scan Clásico Generalizado [11].
El estadístico: η_{max} representa el número máximo de casos (o de unos) que aparecen en una ventana de tiempo (o de longitud en la secuencia binaria).

La idea del método radica en que si existe un conglomerado, el número máximo de casos, o categoría de interés, hallados en una ventana, debe ser muy grande al compararla con las cantidades que aparecen en la mayoría de las ventanas restantes. El cálculo del valor de *p* asociado aparece reportado en la literatura en [2].

Pueden definirse las hipótesis de la forma siguiente:
H_0 : Los casos se distribuyen uniformemente dentro de la secuencia considerada.
H_1 : Existe al menos un conglomerado dentro de alguna de las ventanas analizadas.

3. EL MÉTODO SCAN BORROSO

Este método es una modificación del anterior usando la teoría de la lógica borrosa [1]. Consiste en tomar el parámetro más importante del Scan: el ancho de la ventana móvil y suavizarlo definiendo una ventana borrosa. Dicha ventana tiene una función de pertenencia en cada uno de sus extremos. La ventana borrosa se define de la siguiente forma:

$$\text{Ventana borrosa}_{k} = \begin{cases} (i - k + g + 1) * \dfrac{s_i}{(g+1)} & i = k - g, \ldots, g \\ s_i & i = k, \ldots, k + t - 1 \\ (k + t + g - i) * \dfrac{s_i}{(g+1)} & i = k + t, \ldots, k + t + g - 1 \end{cases}$$

Donde:

- Secuencia: binaria o cantidad de casos reportados por unidad de tiempo, dada por:
  - Scan Lineal:
    - $s_1, s_2, \ldots, s_n$
    - If i<1 then $s_i = 0$
    - If i > n then $s_i = 0$
  - Scan Circular:
    - $s_1, s_2, \ldots, s_n, s_{n+1}, s_{n+2}, \ldots, s_{n+t-1}$ secuencia circular formada por:

      $s_1, s_2, \ldots, s_n$ secuencia y $s_{n+j} = s_j$ para j = 1 to t –1
    - If i<1 then $s_i = s_{n-i}$
    - If i > n + t –1 then $s_i = s_{i-n}$
- k: variable que toma valores desde uno hasta el total de ventanas formadas
- t: longitud de la ventana fija,
- g: longitud de la parte borrosa de la nueva ventana. A esta parte se le llamará suavizado.

El estadístico: η^*_{max} representa igualmente el número máximo de casos (o de unos) que aparecen en una ventana de tiempo o de longitud según el caso. Téngase en cuenta que este valor ahora puede ser real, ver figura 1 [10].

La significación del método Scan Clásico o Generalizado se basa en las distribuciones de Poisson que esta definida para variables aleatorias discretas, luego hay que realizar transformaciones al cálculo de la significación en el método clásico. Concretamente se proponen tres formas diferentes de calcular la significación [9].

- Aproximado
- Aproximar el valor real usando una combinación de dos distribuciones: Poisson hasta el valor entero inferior y uniforme para estimar la parte decimal

- Aproximar el valor real utilizando funciones de interpolación.

Posteriormente se busca el grado de pertenencia de los conjuntos borrosos significativo y no significativo con función de pertenencia S, pudiéndose mover los puntos de corte según los intereses del problema. Para el resultado final se utiliza el método del máximo [7].

Obsérvese que el método Scan Borroso tiene tres parámetros: la parte fija de la ventana móvil (t), la parte borrosa de la ventana móvil (g) y el paso con el que ella se mueve por toda la secuencia.

	Método Scan (t = 5)	
	Clásico	**Borroso (g = 2)**
Secuencia tiempo:	5 0 1 **2 0 2 0 1** 0 1 0 4 0 3	5 0 1 **2 0 2 0 1** 1 1 0 4 0 3
Secuencia binaria:	1 0 1 **1 1 1 1 1** 0 1 0 1 0 1	1 0 1 **1 1 1 1 1** 1 1 0 1 0 1
Ventana:	**5**	**0+.67 + 5 + .67+.33**
Estadígrafo:	$\eta_{max} = 5$	$\eta^*_{max} = 6.67$

Figura 1. Representación gráfica de las ventanas de los métodos Scan Clásico y Borroso

4. OPTIMIZACIÓN BASADA EN ENJAMBRE DE PARTÍCULAS

En la actualidad los modelos bioinspirados se muestran eficientes en la solución de problemas prácticos de diversas áreas. Dentro de los algoritmos bioinspirados usados para la selección de rasgos, la Inteligencia de Enjambres ha sido objeto de estudio, investigación y de aplicación por su simplicidad y robustez [2, 5].

Esta metaheurística está inspirada en el comportamiento social observado en grupos de individuos tales como bandadas de pájaros o enjambres de insectos. Un enjambre se define como una colección estructurada de organismos (agentes) que interactúan. La inteligencia no está en los individuos sino en la acción de todo el colectivo. Tal comportamiento social se basa en la transmisión del éxito de cada individuo a los demás del grupo, lo cual resulta en un proceso sinergético que permite a los individuos satisfacer de la mejor manera posible sus necesidades más inmediatas, tales como la localización de alimentos o de un lugar de cobijo. Cada organismo (partícula) se trata como un punto en un espacio N dimensional el cual ajusta su propio vuelo de acuerdo a su propia experiencia y la experiencia del resto de la banda. La bandada vuela por el espacio de búsqueda localizando regiones o partículas prometedoras. [5].

4.1 FUNDAMENTOS GENERALES DEL ALGORITMO

Sean:

$\mathbf{R}^N$ – R espacio de búsqueda designado, N: cantidad de dimensiones que cuenta dicho espacio

$\mathbf{x}^k_i = (x^k_{i1}, x^k_{i2}, \ldots, x^k_{iN})$ – Posición de la i-ésima partícula en R^N de la iteración k.

$\mathbf{v}^k_i = (v^k_{i1}, v^k_{i2}, \ldots, v^k_{iN})$ – Velocidad de la i-ésima partícula en R^N de la iteración k.

$\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iN})$ – Mejor posición de la i-ésima partícula en R^N de las k iteraciones.

$\mathbf{p}_g = (p_{i1}, p_{i2}, \ldots, p_{iN})$ – Mejor posición del grupo. (Mejor partícula entre las k iteraciones)

$\mathbf{f}^k_i$ – Valor de la función objetivo evaluada en $\mathbf{x}^k_i$.

$\mathbf{f}^{best}_i$ – Mejor valor de la función objetivo evaluada en la i-ésima partícula de las k iteraciones.

$\mathbf{f}^{best}_g$ – Mejor valor de la función objetivo evaluada en el grupo.

$\mathbf{V}_{max}$ – Velocidad máxima que puede alcanzar una partícula

$-\mathbf{V}_{max}$ – Velocidad mínima que puede tener una partícula

$\boldsymbol{\omega}$ – Coeficiente de inercia: valor aleatorio en el rango [0.5, 1]

$\mathbf{c}_1, \mathbf{c}_2$ – Parámetros sociales y cognoscitivos.

$\mathbf{r}_1, \mathbf{r}_2$ – Números aleatorios entre [0, 1].

A continuación se describen los pasos del algoritmo:

1. Inicializar.
   a. Darle valores a las variables $\mathbf{k}_{max}$, $\mathbf{c}_1$, $\mathbf{c}_2$.

b. Inicializar aleatoriamente la posición de las partículas $\mathbf{x}^i_0 \in \mathbf{D}$ en $\mathbf{R}^n$ for i = 1, .., p.
c. Inicializar aleatoriamente la velocidad de las partículas $0 \leq \mathbf{v}^i_0 \leq \mathbf{v}_{max}$ for i = 1,.., p.
d. k = 1

2. Optimizar.
   a. Calcular los valores de $\mathbf{f}^i_k$.
   b. Si $\mathbf{f}^i_k \leq \mathbf{f}^i_{best}$ entonces $\mathbf{f}^i_{best} = \mathbf{f}^i_k$, $\mathbf{p}^i_k = \mathbf{x}^i_k$.
   c. Si $\mathbf{f}^i_k \leq \mathbf{f}^g_{best}$ entonces $\mathbf{f}^g_{best} = \mathbf{f}^i_k$, $\mathbf{p}^g_k = \mathbf{x}^i_k$.
   d. Si se cumple la condición de parada entonces ir a 3.
   e. Actualizar la velocidad de las partículas como sigue $\mathbf{v}^i_{k+1} = \mathbf{v}^i_k + \mathbf{c}_1 \mathbf{r}_1(\mathbf{p}^i_k - \mathbf{x}^i_k) + \mathbf{c}_2 \mathbf{r}_2 (\mathbf{p}^g_k - \mathbf{x}^i_k)$.
   f. Actualizar la posición de las partículas como sigue $\mathbf{x}^i_{k+1} = \mathbf{x}^i_k + \mathbf{v}^i_{k+1}$ para d = 1, . . , N
   g. Incrementar k.
   h. Ir a 2(a).
3. Terminar.
4.

La fórmula de velocidad se explica de la siguiente forma:

$\mathbf{v}^i_{k+1}$ = velocidad anterior de esa partícula + influencia personal+influencia social
Matemáticamente queda [6, 12] de la forma:
$\mathbf{v}^i_{k+1} = \mathbf{v}^i_k + \mathbf{c}_1 \mathbf{r}_1(\mathbf{p}^i_k - \mathbf{x}^i_k) + \mathbf{c}_2 \mathbf{r}_2 (\mathbf{p}^g_k - \mathbf{x}^i_k)$.
Siguiendo recomendaciones de la literatura consultada: $\mathbf{c}_1$=2.08 y c_2=2.08

5. APLICACIÓN DEL ALGORITMO DE OPTIMIXACIÓN BASADA EN ENJAMBRE DE PARTÍCULAS PARA AJUSTAR LOS PARÁMETROS DE LOS MÉTODOS SCAN

Se han desarrollado numerosos experimentos de simulación en los que se le presentan a los métodos Scan secuencias binarias con verdaderos y falsos conglomerados. Los detalles de esos resultados pueden consultarse en [9, 10].

Tales estudios demuestran que los métodos Scan Clásico y Borroso responden muy bien ante falsos conglomerados. La respuesta de no existencia de conglomerados en esas secuencias es correcta casi en el 100% de los casos, con independencia de los valores de los parámetros utilizados.

Las dificultades surgen al analizar secuencias en las que exista al menos una aglomeración. En estos casos el método Scan Borroso supera al método clásico, pero falla cuando se consideran tamaños de ventanas grandes. En estos casos, se hace necesario un estudio adicional para determinar valores óptimos para esos parámetros. Con tal objetivo se decide aplicar a ese problema el algoritmo basado en la optimización de enjambre de partículas. Cada partícula se define por:

- $\mathbf{x}^i_k$ – Es el vector **(ventana móvil**, **paso**, **suavizado)** en la iteración k, las restricciones pueden definirlas el usuario, aunque las implícitas son las siguientes:
  - 1 ≤ **Ventana móvil** ≤ Tamaño de la secuencia
  - 1 ≤ **Paso** ≤ Ventana móvil
  - 0 ≤ **Suavizado** ≤ (Ventana móvil) / 2
- $\mathbf{p}^i_k$ – Es el mejor vector **(mejor ventana móvil**, **mejor paso**, **mejor suavizado)** de la partícula i, hasta la iteración k
- $\mathbf{p}^g_k$ – Es el mejor vector **(la mejor ventana móvil**, **el mejor paso**, **el mejor suavizado)** hasta la iteración k
- $\mathbf{v}^i_k$ – Velocidad de la partícula i en la iteración k. Como se explicó anteriormente, la velocidad se define por: $\mathbf{v}^i_{k+1} = \mathbf{v}^i_k + \mathbf{c}_1 \mathbf{r}_1(\mathbf{p}^i_k - \mathbf{x}^i_k) + \mathbf{c}_2 \mathbf{r}_2 (\mathbf{p}^g_k - \mathbf{x}^i_k)$.
- $\mathbf{f}^i_k$ – Valor de la función objetivo evaluada en $\mathbf{x}^i_k$.
- $\mathbf{f}^i_{best}$ – Mejor valor de la función objetivo evaluada en la partícula i.
- **fgbest** – Mejor valor de la función objetivo evaluada en el grupo.

6. IMPLEMENTACIÓN COMPUTACIONAL: OPTIMUS

Para la implementación computacional se elaboró el sistema Optimus que incorpora entre otros el algoritmo de optimización de enjambre de partículas, que tiene como función objetivo los métodos Scan y determinando los

parámetros adecuados necesarios para el análisis de una secuencia binaria que posea al menos un conglomerado. Este sistema se implementó utilizando el paradigma de la programación orientado a objetos en el lenguaje Borland Delphi 7, con el propósito de facilitar su explotación a cualquier usuario, ejecutando solo el fichero: Optimus.exe, en cualquier equipamiento que posea sistema operativo Windows o Linux, además utiliza las facilidades de las componentes visuales del lenguaje en aras de brindar un ambiente cómodo y sencillo. De forma general se encuentran entre otras las siguientes facilidades:

- ✓ Los datos de entrada son ficheros textos que poseen una secuencia binaria sin restricciones de longitud.
- ✓ Selección del método Scan a utilizar.
- ✓ Selección del algoritmo basado en optimización de enjambres de partículas
- ✓ Otras facilidades generales, tales como guardar los resultados del proyecto, abrir un proyecto, etc.

7. ESTUDIO DEL GENOMA DE LA ESCHERICHIA COLI

Recientemente se han publicado numerosos estudios relacionados con el ADN de la Escherichia Coli. En este trabajo se pretende analizar el genoma completo de la E. Coli con el objetivo de demostrar la existencia de conglomerados de sitios Dam. Los sitios Dam son subsecuencias pequeñas (GACT) dentro del genoma completo que tienen una importancia especial desde el punto de vista bioquímico [4]. El genoma de la E. Coli tiene una longitud aproximada de 4.7 millones de pares de bases. Los resultados con el software Optimus aparecen recogidos en la tabla 1.

Tabla 1. Resultados del método Scan Borroso al detectar conglomerados de sitios Dam en el genoma de la Escherichia Coli

Escherichia Coli IAI1, GenBank, NC_011741, 4.7Mb			
Ancho de la ventana móvil: 245bp			
Método	# 'GACT'	Significación	Localización
Scan circular	12	p = 0.00	246343-246376
Scan Circular Borroso (g=2)	12	Significativo	246343-246376
Scan Circular Borroso (g=4)	12	Significativo	246343-246376

La columna correspondiente a la significación muestra, para el Scan circular, la existencia de conglomerados de sitios Dam. Para los métodos borrosos la respuesta es significativo o no significativo, en dependencia de la pertenencia a ambos conjuntos borrosos. Obsérvese que en ambos casos la respuesta es positiva utilizando principio del máximo para ofrecer una respuesta dura [7], corroborando la existencia de las aglomeraciones buscadas.

8. CONCLUSIONES

En el presente trabajo se combinan varios métodos matemáticos para brindar una solución eficiente al problema de la detección de conglomerados en secuencias binarias o de tiempo.
Se utiliza un algoritmo bioinspirado de optimización para determinar valores adecuados de los parámetros de los métodos Scan.

Los algoritmos desarrollados están implementados en el software Optimus, que es sencillo y ofrece un ambiente amigable.

A manera de ejemplo aparece una aplicación Bioinformática en el campo del análisis de secuencias.

RECEIVED OCTOBER 2010
REVISED SEPTEMBER 2011

REFERENCIAS

[1] BUCKLEY, J. y JOWERS, L. (2007): **Monte Carlo Methods in Fuzzy Optimization**, Heidelberg.(editorial y ciudad)

[2] GLAZ, J. (1993): **Approximations for the tail probabilities and moments of the Scan statistics.** Statistics in medicine 12: 1845-1852.

[3]. JACQUEZ, G., WALLER, L., GRIMSON, R. y WATENBERG, D. (1996): The analysis of Disease Clusters, Part II: Introduction to techniques. **Infection Control and Hospital Epid.** 17, 385-97.

[4] KARLIN, S. Y BRENDEL, V. (1992): Chance and Statistical Significance in Protein and DNA Sequence Analysis. **Science** 39-49. 257, 39-49.

[5] KENNEDY, J. (1997): The particle swarm: social adaptation of knowledge. **IEEE International Conference on Evolutionary Computation**. Indianápolis, USA.

[6] KENNEDY, J., EBERHART, R. y SHI, Y. (2001): **Swarm Intelligence**. . Morgan Kaufmann Series in Artificial Intelligence. Georgia.

[7] MARTÍN DEL BRÍO, B. y SÁNCHEZ, A. (2005): **Redes Neuronales y Sistemas Difusos**. Alfaomega, México.

[8]. NAUS, J. I. (1982): Approximations for distributions of Scan statistics. **Journal of the American Statistical Association** 77, 177-183.

[9] RODRÍGUEZ, L., CASAS, G., GRAU, R. y GÓMEZ, O. (2009): Approximations for the distribution of Fuzzy Scan Statistics. **Investigación Operacional**, 30, 131-139.
.http://www.journaldatabase.org/articles/approximations_for_distribution_fuzzy.html.

[10] RODRÍGUEZ, L., CASAS, G., GRAU, R. y MARTÍNEZ, Y. (2008): Fuzzy Scan Method to Detect Clusters. **International Journal of Biomedical Sciences**, 3, 111 -115.
©http://www.waset.org/journals/ijbls/v3/v3-2-16.pdf Spring 2007

[11] RODRÍGUEZ, L., CASAS, G., GRAU, R. y PUPO, M. (2007): Generalización de dos métodos de detección de conglomerados. Aplicaciones en Bioinformática. **Revista de Matemática: Teoría y Aplicaciones**, http://www.latindex.ucr.ac.cr/mat004-03.php. 15 (1): 27 - 40.

[12] WANG, X., YANG, J., TENG , X., XIA, W. y JENSEN, R. (2007): Feature selection based on rough sets and particle swarm optimization. **Pattern Recognition Letters** 28, 459-471.