Programación Matemática y Software (2019) 11 (3): 12-31. ISSN: 2007-3283

Selección de Características de Microarreglos de ADN Utilizando una Búsqueda Cuckoo

Feature Selection from DNA-Microarray using a Cuckoo Search

Luis Alberto Hernández Montiel, Carlos Edgardo Cruz Pérez, Luis David Hernández

Universidad del Istmo, Campus Ixtepec (UNISTMO), Ciudad Ixtepec, Oaxaca, México, 70110

* Correo-e: luisd2h@bianni.unistmo.edu.mx

PALABRAS CLAVE:

Microarreglos de ADN, Preprocesamiento, Fusión de filtros, Selección, Clasificación.

RESUMEN

En este artículo, se propone un método híbrido para la selección y clasificación de datos de microarreglos de AND. Primero, el método combina los subconjuntos de genes relevantes obtenidos de cinco métodos de filtro, después, se implementa un algoritmo basado en una búsqueda cuckoo combinado con un clasificador MSV. El algoritmo híbrido explora dentro del subconjunto obtenido en la etapa anterior y selecciona los genes que alcanzan un alto desempeño al entrenar al clasificador. En los resultados experimentales, el algoritmo obtiene una tasa de clasificación alta seleccionado un número pequeño de genes, los resultados obtenidos son comparados con otros métodos reportados en la literatura.

KEYWORDS:

DNA-Microarray, Preprocessing, Filters Fusion, Selection, Classification.

ABSTRACT

In this paper, a method for selection and classification of DNA-microarray data is proposed. Firstly the method combines the relevant genes subsets obtained of four data filtering methods, second, an algorithm based on a cuckoo search combined with SVM-classifier is applicate. The hybrid algorithm, explores within of subset obtained in the previous stage and selects the genes that achieve a high performance when training the classifier. In the experimental results, the algorithm achieves a high classification rate selecting a smaller number of genes, the results obtained are compared with other methods reported in literature.

Recibido: 3 de agosto de 2018• **Aceptado:** 5 de diciembre de 2018 • **Publicado en línea:** 31 de octubre de 2019

I. INTRODUCCIÓN

Los microarreglos de ADN, son una tecnología de análisis de expresión genética, utilizada por científicos e investigadores para comprender mejor la dinámica celular y sus relaciones con diferentes estados patológicos. Permite el estudio de miles de genes simultáneamente para clasificar muestras de tejido entre muestras normales y muestras enfermas [1], [2], [3], [4], [5]. Sin embargo, la selección de un conjunto pequeño de genes que resulten relevantes para su clasificación no es una tarea fácil, al tener una alta dimensión, los datos seleccionados pueden ser erróneos, esto se debe a que no toda la información dentro del microarreglo es relevante, si se seleccionan genes ruidosos, el algoritmo de selección y clasificación puede obtener resultados poco confiables. Para ello, es necesario hacer una reducción significativa de la dimensión del microarreglo de ADN, esto nos permite descartar genes ruidosos y seleccionar solo genes con información relevante que ayuden en el diagnóstico de una enfermedad. Para abordar el problema de la alta dimensión, proponemos un método que utiliza cuatro filtros estadísticos y un algoritmo hibrido basado en una búsqueda cuckoo combinada con un clasificador MSV. El método está dividido en dos etapas, la primera consiste en hacer un pre-procesamiento de los microarreglos de ADN y en la siguiente etapa se crea un método de selección y clasificación utilizando el algoritmo hibrido. Con este método, se buscan los genes más relevantes dentro de cinco bases de datos públicas obtenidas a través de la tecnología de microarreglos de ADN.

II. EL PROBLEMA DE LA ALTA DIMENSIÓN

Los datos obtenidos de un microarreglo de ADN presentan un gran cúmulo de información, por lo que constituye un reto para hacer un análisis eficiente de su contenido. Su alta dimensión genera un problema en términos de precisión y de complejidad computacional [12]. La mayoría de las bases de datos, no solo cuentan con una gran cantidad de atributos (características) y un número limitado de muestras, también contienen dos o más número de clases (categorías) a las que pertenece cada uno de los atributos [1]. Para solucionar este problema, diferentes autores han propuesto la utilización de algoritmos de clasificación, por ejemplo Guyon et al. [1], y Golub et al [2], logran clasificar dos clases de muestras de tejido, utilizando un clasificador MSV y una clasificación molecular. En el trabajo de Hwang et al [3] y Wang et al [4], se propone la utilización de un algoritmo de aprendizaje máquina basados en métodos de filtros y wrappers, logrando eliminar la información menos relevante dentro del microarreglo de ADN y reducir la dimensión de la base de datos, obteniendo información confiable para su análisis. Otros autores aplican estrategias de computo bioinspirado para atacar el problema de alta dimensión, por ejemplo Kulkarni [5] utiliza una técnica basada en programación genética e información mutua, para generar una reducción efectiva del microarreglo que contiene información sobres cáncer de colon. Li [6] y Mohamad, et al [7], abordan el problema con un algoritmo genético y un algoritmo basado en una optimización por cumulo de partículas (PSO). La idea principal de estos métodos es utilizar las propiedades de los algoritmos generando una búsqueda aleatoria para obtener información relevante dentro de las bases genómicas y reducir eficientemente el tamaño de los microarreglos. Otras técnicas que se han implementado eficazmente para abordar este problema son los métodos de clusters, en el trabajo de XU [8] presenta un cluster basado en conjuntos corrugados difusos y el trabajo de Mishra [9] se basa en un técnica de clusters k-medias, el autor basa su estrategia para abordar el problema de alta dimensión, agrupando las características que tienen una similitud, de esta forma, al tener las características agrupadas, se utiliza algún método de puntuación discriminante para eliminar la información redundante y ruidosa de los microarreglos. Otra de las formas que se aborda este problema es la implementación de algoritmos híbridos. En el trabajo de Yang [10], se proponen un sistema hibrido que

consiste en la unión de métodos de filtrado de datos y un algoritmo wrapper, En el proceso de filtrado, genera una puntuación a cada gen candidato de los microarreglos de ADN. En el proceso wrapper, utiliza un algoritmo genético multi-objetivo para seleccionar los genes discriminativos utilizando la información proporcionada por el proceso de filtrado. El trabajo de Bonilla-Huerta [11] propone un algoritmo hibrido que genera una limpieza de los datos utilizando métodos de filtrados estadístico, después, genera una selección de un subconjunto de genes utilizando una combinación de un algoritmo genético, una búsqueda tabú para buscar dentro de la base genómica y una máquina de soporte vectorial que se encarga de medir la calidad del subconjunto seleccionado. A pesar del número de técnicas implementadas para solucionar el problema de la alta dimensión, no se ha llegado a una solución estable, debido a esto, siguen apareciendo trabajos con nuevas propuestas para solucionarlo y dar un estudio preciso de los genes seleccionado.

III. MATERIALES Y MÉTODOS

Los microarreglos contienen información relevante, mezclada con información ruidosa y redundante [12]. Debido a esto, se dificulta extraer información valiosa de un microarreglo, generando tiempos de procesamiento largos y resultados poco confiables. Para solucionar este problema se han implementado diferentes métodos (algunos descritos en la sección anterior) para obtener genes relevantes al explorar dentro de un microarreglo de ADN. En esta sección se describen los microarreglos utilizados en nuestro experimento, también los métodos con los cuales se ha abordado el problema planteado en la sección anterior.

3.1. Microarreglos de ADN

Con los avances de la tecnología de microarreglos de ADN, surgen más datos de expresión génica a disposición para ser analizados. Estos datos, se utilizan en la clasificación de muestras de tejidos, aunque los datos en bruto no son de gran utilidad. Su verdadero valor radica en extraer información relevante de ellos [13]. En este estudio se utilizan cinco bases de datos de dominio las cuales se describen a continuación.

La base de datos de **Leucemia** [2] contiene 7129 datos de expresión genética con 72 muestras de oligonucleótido, 25 de AML (Acute Myeloid Leukemia) y 47 ALL (Acute Lymphoblastic Leukemia). La base de datos de **Cáncer de Colon** [14] contiene 2000 datos de expresión genética con 62 muestras celulares, 40 son pruebas de tumor y 22 son pruebas normales. La base de datos de **CNS** [15] contiene con 7129 datos de expresión genética con 60 muestras de tumores embrionarios del sistema nervioso central, 21 son de survivors y 39 de failures. La base de datos de **Cáncer de Pulmón** [16] contiene 12533 datos de expresión genética con 181 muestras de tejido pulmonar, 31 son malignant pleural mesothelioma y 150 son de adenocarcinoma. La base de datos **DLBCL** [17] contiene 4026 datos de expresión genética con 47 muestras, 24 son del grupo B-like germinal y 23 son del grupo B-like activado.

3.2. Pre-procesamiento de Datos

Los datos obtenidos dentro de un microarreglo contienen factores que pueden generar un sobre entrenamiento del clasificador [18]. Esto se debe a que además de tener una dimensión alta, las bases de datos también incluyen información que no es relevante para su clasificación. Otro factor es que cada gen registrado dentro de la base de datos tiene una escala numérica diferente. Estos factores pueden generar un problema de precisión en su clasificación y de esta forma se puede llegar a seleccionar una característica poco relevante para un diagnóstico. En esta sección, se presenta la limpieza de los datos utilizando un pre-procesamiento dividido en dos pasos (ver figura 1), los cuales se explican a continuación.

A. Normalización de los datos

La normalización de las bases de datos se utiliza para transformar las datos a un rango entre cero y uno (0,1) y así hacer más fácil la clasificación y selección de los genes más pertinentes.

Es este estudio como primer paso se realiza una normalización de los datos basada en una técnica min-máx. [19]:

$$X' = \frac{X - Min(X)}{Max(X) - Min(X)} \quad (1)$$

Donde X es la base de datos original. Min(X) y Max(X) es el dato mínimo y máximo existente dentro de la base de datos. X' es la nueva base de datos normalizada.

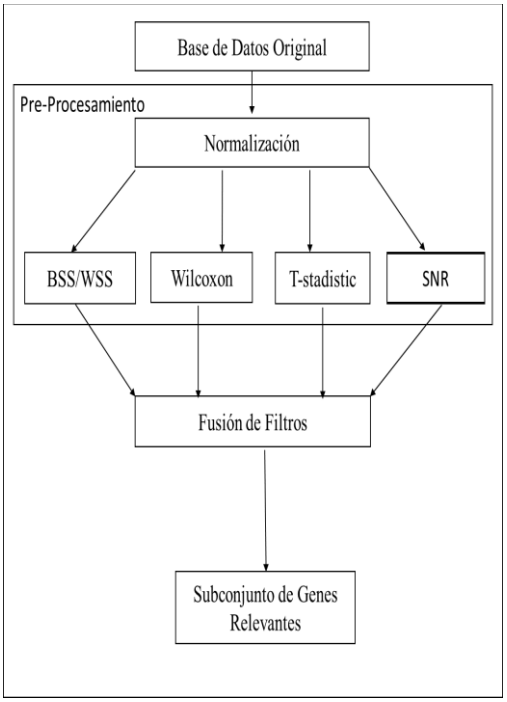

Figura 1. Resultados obtenidos por los clasificadores para la base de datos de leucemia

B. Filtrado de datos

Después de que la base de datos está normalizada, se genera una primera reducción de los microarreglos de ADN utilizando un pre-procesamiento de los datos. Esta etapa se realiza mediante la utilización de un método de filtro de datos. Los filtros seleccionan características basándose en un criterio de discriminación relativamente independiente de su clasificación, teniendo un conjunto de ejemplos y uno de características, los filtros toman cada variable de manera individual y calculan una medida de puntuación para utilizarla posteriormente como indicador discriminatorio de las variables, reduciendo la dimensionalidad del espacio de búsqueda descartando o filtrando características redundantes y/o irrelevantes. [20].

En nuestro experimento, se utilizan cuatro típicos métodos de filtrado de datos. La idea de ésta etapa, es que cada uno de estos métodos seleccionen un subconjunto de genes de una base genómica de manera independiente. El método de filtrado utiliza un valor de pertinencia (Ranking) que se le asigna a un gen en particular, de esta forma discriminan las características menos relevantes y así generan una primera reducción de un microarreglo de ADN. Los métodos utilizados en este experimento se describen a continuación.

- **BSS/WSS**

La selección de genes se basa en la razón de la suma de cuadrados entre grupos (BSS) y dentro de los grupos (WSS). Para el gen$_j$, la razón está dada por [21]:

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)\left(\bar{x}_{kj} - \bar{x}_{.j}\right)^2}{\sum_i \sum_k I(y_i = k)\left(x_{ij} - \bar{x}_{kj}\right)^2}, \quad (2)$$

Donde $\bar{x}_{.j}$ denota el nivel medio de la expresión del gen j a través de todas las muestras y $\bar{x}_{kj}$ denota el nivel medio de la expresión del gen j en todas las muestras para la pertenencia de la clase k.

- **Relación señal a ruido (SNR)**

Identifica los patrones de expresión genética con una diferencia máxima en la expresión media entre dos grupos y la variación mínima de expresión dentro de cada grupo. En este método, los genes son clasificados de acuerdo a sus niveles de expresión [9].

$$SNR = |(\mu_1 - \mu_2)/(\sigma_1 + \sigma_2)| \quad (3)$$

Donde $\mu 1$ y $\mu 2$ denotan los valores medios de expresión de la clase 1 y clase 2, respectivamente, $\sigma 1$ y $\sigma 2$ son las desviaciones estándar de las muestras en cada clase.

- **Wilcoxon test**

Para cada gen j, sólo se necesita el supuesto de que las observaciones $x_{ij} \ldots x_{nj}$ sean independientes. Si rank (x_{ij}) denota el rango de x_{ij} en la sucesión $x_{ij} \ldots x_{nj}$, la prueba estadística para el gen j está dada por [22]:

$$W_j = \sum_{i:Y_i=1} rank(x_{ij}) \quad (4)$$

Para probar la hipótesis se utiliza

$$\begin{aligned} H_0&: M_e(X_j|Y=1) = M_e(X_j|y=2) \\ &\quad vs \\ H_1&: M_e(X_j|Y=1) \neq M_e(X_j|y=2) \end{aligned} \quad (5)$$

Bajo H_0, W_j tiene una distribución de Wilcoxon con grados de libertad n1 y n2.
El valor descriptivo de la prueba (*p-value*) correspondiente para cada variable *j* puede ser usado como una medida de relevancia.

- **T-Statistic**

Cada muestra se etiqueta con {1, -1}. Para cada gen f_i la media μ_j^1 (μ_j^{-1}) y la desviación estándar δ_j^1 (δ_j^{-1}), se calculan utilizando sólo las muestras etiquetadas con 1 (-1). Entonces una puntuación Tf_j puede ser obtenida por [23]:

$$T(f_j) = \frac{\left|\mu_j^1 - \mu_j^{-1}\right|}{\sqrt{\left(\delta_j^1\right)^2/n_1 + \left(\delta_j^{-1}\right)^2/n_{-1}}} \quad (6)$$

Donde n_1 (n_{-1}), es el número de ejemplos etiquetados con 1 (-1). Son considerados como los genes más discriminatorios aquellos que obtengan la puntuación más alta.

C. Fusión de filtros

Con el paso anterior, se generan diferentes subconjuntos de una sola base de datos, esto se debe a que un microarreglo es pre-procesado por diferentes métodos de filtrado. Al utilizar diferentes filtros con distintas capacidades estadísticas, cada filtro prioriza un gen en particular otorgando diferentes posiciones para los genes, colocando los genes relevantes en una mejor posición que los no relevantes. En la actualidad, existen diferentes trabajos como el de Yang [10], Radivojac [24] y Kumari [25], que utilizan un pre-procesamiento similar al que se presenta en este trabajo. En estos trabajos, la etapa de pre-procesamiento genera una limpieza del microarreglo eliminando todos los genes ruidosos y redundantes, quedando solo subconjuntos con información relevante. Después, con el subconjunto resultante se emplea algún tipo de heurística de búsqueda para reducir la dimensión y seleccionar información relevante de ellos. En nuestro caso, basados en el método presentado en Bonilla Huerta [11], hemos combinado los diferentes subconjuntos de cada base de datos (como se muestra en la figura 1), obtenidos por el pre-procesamiento de datos, como resultados de este paso, se ha creado un subconjunto único de genes que servirá para entrenar el algoritmo BC/MSV.

Durante el pre-procesamiento se obtienen cuatro subconjuntos del mismo microarreglo, donde cada filtro ha priorizado un gen en particular. Se observa que un gen está

colocado en diferente posición dentro de la puntuación de cada subconjunto obtenido, esto se muestra en la tabla 1. Para generar un subconjunto único de genes relevantes, se propone una fusión de los cuatro subconjuntos obtenidos por los métodos de filtro.

Tabla I. Ranking obtenido por los métodos de filtro para la base de leucemia

Rank	BSS/WSS	Wilcoxon	T-estadístico	SNR
1	**1882**	4847	4847	**1882**
2	760	**1882**	2020	4847
3	4847	2020	**1882**	1745
4	1834	1745	6218	6041
5	5772	6041	1834	2020
⋮	⋮	⋮	⋮	⋮
50	1817	6919	3710	4324

La fusión de los filtros se hace de la siguiente manera:

El primer paso es buscar la posición que ocupa un gen específico dentro de la puntuación de cada filtro, como se muestra en la tabla 2.

Tabla 2. Ranking obtenido por los métodos de filtro para la base de leucemia

Gene	Posiciones				$f(x)$
	BS	WT	TS	SN	
1882	1	2	3	1	**7**
760	2	14	10	31	**57**
4847	3	1	1	2	**7**
2020	52	3	2	5	**62**

El paso dos es sumar las posiciones encontradas de la siguiente forma [11]:

$$f(x) = \sum_{i=1}^{n} posición(gen_i) \qquad (7)$$

Donde n es el número de filtro con los que se ha trabajado, $posición(gen_i)$ es el ranking dado por cada filtro.

El tercer paso es generar la fusión de las puntuaciones por cada filtro. El problema radica en que al sumar las posiciones encontradas para cada gen, el resultado se encuentra en diferentes escalas numéricas, lo cual dificulta crear un subconjunto basado en su puntuación (ver tabla 2). En nuestro caso, se ha generado un método basado en promediar la suma de las puntuaciones de la siguiente forma:

Primero se calcula la media de f(x) obtenida con la ecuación 7:

$$\bar{x} = \frac{\sum_{i=1}^{n} f(x_i)}{n} \qquad (8)$$

Donde $\bar{x}$ es la media, $\sum_{i=1}^{n} f(x_i)$ es la sumatoria de todas las posiciones dadas a cada gen (x_i), n es el número total de genes.

Después generamos una nueva puntuación para cada gen basada en la siguiente ecuación:

$$NR(gen) = 1 - \left(\frac{f(x)}{\bar{x}}\right) \qquad (9)$$

Donde $NR(gen)$ es la nueva puntuación que adopta el gen. $f(x)$ es la puntuación obtenida al sumar las posiciones de cada gen y $\bar{x}$ es la media obtenida de la ecuación 8.

Con los pasos anteriores se ha generado una fusión de los genes obtenidos por los cuatro filtros estadísticos utilizados en este experimento. Las nuevas puntuaciones dadas para cada gen se muestran en la tabla 3. El proceso de fusionar los filtros se ha realizado para las cinco bases de datos con las que se han trabajado.

Tabla 3. Nuevo ranking de los 10 primeros genes, obtenido por la fusión de los filtro para la base de leucemia

P	***Gene***	***NR***
1	4847	0.9989
2	1882	0.9989
3	1745	0.9952
4	1834	0.9943
5	6041	0.9934
6	3320	0.9926
7	5039	0.9918
8	760	0.9907
9	2121	0.9905
10	2020	0.9898

El nuevo subconjunto obtenido de esta fusión de filtros, será utilizado en la etapa de entrenamiento de un algoritmo hibrido basado en una búsqueda cuckoo combinado con un clasificador MSV.

3.3. Selección y Clasificación de Genes Utilizando una Búsqueda Cuckoo y un clasificador MSV

En esta nueva etapa se propone un método hibrido para reducir el tamaño del subconjunto de genes obtenido de la etapa anterior, para ello, se utiliza una búsqueda cuckoo binaria como técnica de selección de genes. Para la clasificación de los genes seleccionados por el algoritmo, se utiliza un clasificador basado en una máquina de soporte vectorial. El método propuesto se describe a continuación.

3.3.1. Búsqueda Cuckoo y Vuelos Lévy

A. Búsqueda Cuckoo

La búsqueda cuckoo [26], es una técnica basada en el comportamiento parasitario que tiene el ave cuckoo, específicamente las hembras, que llegan a tener alrededor de 12 huevos, estos huevos son depositados en nidos ajenos, con el fin de que sea otra ave la que se encargue del crecimiento del polluelo, evitando que el cuckoo se preocupe por la creación de un nido o la incubación del huevo y teniendo más tiempo para su reproducción [26]. Para poder depositar sus huevos, el ave cuckoo escoge un nido anfitrión con características específicas como ubicación del nido o que el pájaro receptor sea insectívoro además, eligen un nido donde el pájaro acaba de poner sus propios huevos [27]. Al depositar el huevo en un nido ajeno, el pájaro receptor puede tener un conflicto al descubrir un huevo extraño en su nido, así el huevo cuckoo tiene una alta probabilidad de ser expulsado del nido o que el pájaro pueda abandonar el nido y crear otro nuevo. Es por esto que algunas especies cuckoo han desarrollado una habilidad de mimetización logrando copiar características como tamaño, forma o color del huevo, similares a los del nido receptor y aumentar su probabilidad de no ser descubiertos [26], [27].

B. Vuelos Lévy

Varios estudios han demostrado que el comportamiento de vuelo de muchos animales e insectos tienen las características típicas de los vuelos Lévy. El vuelo lévy [26] es un tipo de paseo aleatorio que consiste en vuelos de manera lineal, usando una serie de trayectorias marcadas por repentinos giros bruscos de 90º. Lo que lleva a una escala intermitente de pasos aleatorios con un patrón de búsqueda libre de acuerdo a una distribución de probabilidad. La generación de los pasos aleatorios con los vuelos de Lévy consta de dos etapas: la elección de una dirección aleatoria y la generación de pasos que obedecen la distribución Lévy.

La generación de una dirección aleatoria debe ser tomada de una distribución uniforme, mientras la generación de los pasos aleatorios es complicada, una de las formas más eficientes y sencillas de generar estos pasos, es utilizar el llamado algoritmo de Mantegna para que una distribución Lévy sea simétrica y estable. En el algoritmo Mantegna la longitud de los pasos aleatorios está dada por [26]:

$$s=\frac{u}{|v|^{1/\beta}} \qquad (10)$$

Donde u y v se extraen de las distribuciones normales. Es decir

$$u \sim N(0,\sigma_u^2) \quad , \quad v \sim N(0,\sigma_v^2) \qquad (11)$$

Donde σ_u y σ_v están dadas por:

$$\sigma_u=\left\{\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma[(1+\beta)/2]\beta 2^{(\beta-1)/2}}\right\}^{1/\beta}, \qquad \sigma_v=1 \qquad (12)$$

Esta distribución (para *s*) obedece la expectativa de la distribución Lévy cuando $|s|\geq|s_0|$ donde $|s_0|$ es el paso con una longitud más pequeña. En principio $|s_0| \gg 0$, pero en realidad $|s_0|$ puede ser tomado como un valor adecuado tal que $|s_0|$ puede tomas valores entre $0.1\ a\ 1$.

3.3.2. Implementación del Algoritmo Hibrido CS/MSV

Basados en el algoritmo cuckoo search (CS) desarrollado por Xin-She Yang y Suash Deb [26]. Hemos combinado el algoritmo de búsqueda cuckoo con el clasificador MSV, generando una selección y clasificación efectiva de los microarreglos de ADN. La búsqueda cuckoo funciona para problemas con datos continuos. Para abordar el problema de selección de características, hicimos una modificación al algoritmo original para que funcione en problemas con datos binarios.

Primero se fija un número disponibles de nidos receptores, después, se genera un nido binario de forma aleatoria que sigue una distribución uniforme. De esta forma se parte con un conjunto inicial de huevos representados por ceros y unos, que han sido depositados en el nido. Los genes del microarreglo se asocian a cada huevo depositado en el nido como se muestra en figura 2.

Huevo	H1	H2	H3	H4	H5	H6	⋯	Hn
	1	0	1	0	1	0	⋯	0
Gen	G1	G2	G3	G4	G5	G6	⋯	Gn

Figura 2. Representaciones de los huevos del nido y su asociación con los genes del microarreglo.

El siguiente paso es evaluar la calidad de los huevos del nido, para ello dentro de la función objetivo del algoritmo, se utiliza un clasificador MSV. El clasificador MSV discrimina los datos de entrada (huevos/genes) que tienen clases linealmente separables [28]. Crea un hiperplano óptimo en el espacio del vector de características (huevos/genes), de tal manera que maximice el margen de separación entre las características con etiquetas positivas y negativas [28]. En nuestro caso, este clasificador se utiliza dentro de la función objetivo del algoritmo, esto ayuda a medir la calidad los genes seleccionados por la búsqueda cuckoo, la clasificación de los genes se describe a continuación:

Dado un conjunto finito de muestras (genes candidatos) *m* con clases positivas y negativas definido por:

$$S=\{(x_i,y_i)|(x_i,y_i) \in \Re^n\ x\ \{\pm 1\},\ \ i=1,2,. \qquad (13)$$

Donde $x_i \in \Re^n$, $y_{i\text{i}} \in \{\pm 1\}$ indica una etiqueta de la muestra de x_i, y el hiperplano se define por [28]:

$$f(x)=\sum_{i=1}^{m} a_i y_i K(x_i,x)+b \qquad (14)$$

Donde $K(x_i,x)$ es la función del kernel y el signo de $f(x)$ determina a que clase pertenece. La construcción de un hiperplano óptimo es equivalente a encontrar todo el soporte de los vectores en a_i y un sesgo en b.

En este artículo, se utiliza el clasificador MSV para evaluar la calidad de los genes seleccionados por la búsqueda cuckoo. Para medir la estimación del error que genera el

clasificador MSV se utiliza una validación cruzada 10-fold, en donde se divide una muestra en *n* partes, tomando *n-1* partes aleatorias como conjunto de entrenamiento y una parte como conjunto de prueba [29]. Con esta validación aseguramos que los resultados obtenidos por el clasificador sean estables y no se genere un sobre-ajuste

Después de obtener la calidad del primer nido, lo siguiente es depositar los huevos cuckoo dentro del nido, para esto se siguen estas 3 reglas:

1.- Cada cuckoo coloca un huevo a la vez ($x_i^{(t+1)}$), depositándolo en el nido eligiendo su posición mediante la ejecución de un paseo aleatorio basado en el vuelo Lévy [27]:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Lévy(\lambda) \qquad (15)$$

Donde $\alpha > 0$ es la longitud del paso con el cual se genera el paseo aleatorio. Para que el paseo aleatorio pueda llegar a un nuevo estado/ubicación, este depende de la ubicación actual (el primer término de la ecuación anterior) y la probabilidad de transición (el segundo término). El símbolo $\oplus$ es una multiplicación a la entrada. El tamaño del paso se multiplica por los números aleatorios generados con la distribución de Levy, tal movimiento al azar se llama vuelo de Levy, este vuelo proporciona un paseo aleatorio basado en la distribución Levy definida por [26]:

$$Lévy \sim u = t^{-\lambda} \quad , \quad (1 < \lambda < 3) \qquad (16)$$

Donde el vuelo Lévy emplea una longitud de paso al azar que se extrae de una distribución de Lévy dada por las ecuaciones (11) y (12). Esta distribución tiene una varianza infinita con una media infinita. Aquí se forma un proceso de paseo aleatorio con una distribución marcada por una longitud de paso. Algunos de los huevos cuckoo son generados por un paseo Lévy alrededor de la mejor solución obtenida hasta el momento.

Utilizando un vuelo Lévy, se depositan los huevos cuckoo dentro del nido. Al generar el vuelo Lévy, los nuevos huevos depositados en el nido cambian a valores continuos, ya que la nueva posición es obtenida mediante la distribución Lévy.

En nuestro caso, al ser un problema binario, utilizamos las ecuaciones (17) y (18) para modificar los nuevos valores obtenidos por el vuelo Lévy y transformar los huevos a un valor de 0 o 1 [30].

$$S\left(x_i^j(t)\right) = \frac{1}{1+e^{-x_i^j(t)}} \qquad (17)$$

$$+x_i^j(t+1) \quad = \quad \begin{cases} 1 & si\ x_i^j(t+1) > \sigma \\ 0 & otro\ caso \end{cases} \qquad (18)$$

En donde $\sigma \sim U(0,1)$ y $x_i^j(t)$ denota el nuevo valor del huevo en el paso t. De esta forma se han depositado los huevos cuckoo dentro del nido.

2.- Se mide la calidad del nido con los huevos cuckoo a través de la función objetivo y el clasificador MSV.

3.- Se utiliza una probabilidad definida por $p_a \in [0,1]$, Para descubrir y eliminar huevos cuckoo del nido, con los huevos sobrevivientes se construye un nuevo nido que pasa a la siguiente generación. Un ciclo se ha completado, el algoritmo se detiene por un número fijo de iteraciones o cuando la función objetivo se ha estabilizado.

Siguiendo estos pasos, se ha creado un algoritmo que logra seleccionar un subconjunto óptimo de características informativas. Reduciendo el tamaño del microarreglo de ADN y seleccionando genes con información relevante que permite distinguir dos tipos de cáncer. El algoritmo que se ha creado se describe a continuación

ALGORITMO 1
Inicio

1 ***Genera*** *N – Población inicial de Nidos binarios*
2 ***Función Objetivo*** *– Clasificador SVM*
3 *K – número de iteraciones*

```
4   F(x) – Calidad del nido N
5   Pa – % de huevos descubiertos
6     Mientras (k < Kmax) hacer
7     N' - Inserta los huevos cuckoo via Vuelos lévy
8     N' – Binario(N')
9     F(x') – Calidad del nido con huevo cuckoo
10      Para i hasta N'
11      Selecciona un nido de N'(i)
12        Si (f(x')>f(x))
13          Reemplaza N(i) por N'(i)
14        Fin.
15      Fin
16    N' = N*pa – Fracción de huevos descubiertos,
17    Evalúa calidad del nuevo nido f(x')
18      Para i hasta N'
19      Selecciona un nido de N
20        Si (f(x')>f(x))
21          Reemplaza N(i) por N(j)
22        Fin.
23      Fin
24    Rankind de mejores soluciones
25    Fin
26  Selecciona el mejor nido
```

IV. RESULTADOS E INTERPRETACIÓN BIOLÓGICA

El protocolo experimental se realizó en una pc DELL vostro con procesador i5 y memoria RAM de 4gb. El algoritmo fue implementado en Matlab versión 7.12. Los parámetros más confiables con los que ha trabajado el algoritmo cuckoo se muestran en la tabla 4.

Tabla 4. Parámetros utilizados por el algoritmo hibrido

PARÁMETROS	
Numero de nidos	50
Longitud del nido	50
% de huevos descubiertos	0.25
Numero de iteraciones	500

Con los parámetros mostrados en la tabla 4, el método propuesto ha obtenido resultados confiables que se describen a continuación

6.1. Resultados

El protocolo experimental se dividió en dos etapas, en la primera se utilizan cuatro métodos de filtrado de datos que funcionan como una etapa de pre-procesamiento generando una reducción significativa de las cinco bases genómicas. En esta etapa se descartan los genes ruidosos y redundantes y se obtiene como resultado los nuevos subconjuntos con información relevante, de este proceso se han obtenido cuatro subconjuntos de una sola base de datos, estos subconjuntos se han fusionado utilizando la puntuación que cada gen obtiene por el método de filtrado, este paso tiene la finalidad de crear un subconjunto único de información relevante.

En la siguiente etapa se ha construido un algoritmo hibrido basado en una búsqueda cuckoo, este algoritmo logra explorar y eliminar características, gracias a sus propiedades de optimización, para saber si el gen seleccionado por la búsqueda es relevante para un diagnóstico, un clasificador basado en una máquina de soporte vectorial es introducido en la función objetivo del algoritmo, esta combinación permite obtener una tasa de clasificación del o los genes que el algoritmo ha seleccionado, dejando o eliminando genes que no logren entrenar de manera efectiva al clasificador y seleccionado genes que tiene una tasa alta de clasificación, logrando reducir el tamaño del microarreglo y generando un subconjuto de genes relevantes.

La figura 3 muestra las tasas de clasificación obtenidas por el algoritmo hibrido propuesto. En la imagen se observa que el algoritmo se estabilizo mejorando su tasa de clasificación para cada una de las bases de datos. Al combinar la búsqueda cuckoo con un clasificador MSV se ha logrado alcanzar una tasa de clasificación alta.

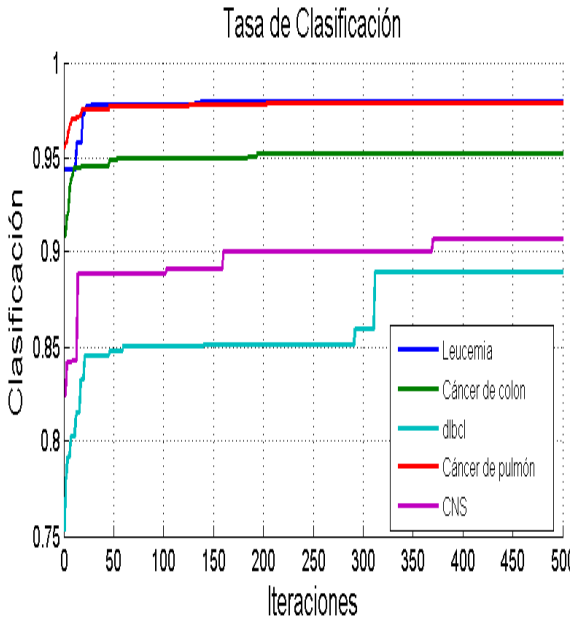

Figura 3. Resultados obtenidos por los clasificadores para la base de datos de leucemia.

El algoritmo ha obtenido los siguientes resultados, para la base de datos de leucemia alcanzo una taza de clasificación del 97.93%. Con la base de datos de cáncer de colon, el algoritmo obtuvo un 95.21% de clasificación, en la base de datos de DLBCL el algoritmo obtuvo 88.99% de clasificación, para la base aprendizaje máquina.

Tabla 5. Comparación de las tasas de clasificación obtenidas por el algoritmo CS/SVM

AUTOR	Leu	Colon	DLBCL	Pulmón	CNS
	%(G)	%(G)	%(G)	%(G)	%(G)
[29]	71.39 (5)	80.07(7)	--	--	--
[32]	96.8(10)	88.6(10)	--	94.7(10)	--
[34]	95.9(25)	87.7(25)	--	--	--
[40]	--	--	--	92.86(71)	--
[31]	92.52(6)	87.00(8)	--	--	95.44(12)
[33]	94.7(13)	80.6(21)	--	--	--
[6]	95.1(21)	88.7(16)	--	--	--
[11]	99.5(3)	90.5(3)	93.8(3)	96.0(3)	94.3(4)
[38]	91.1	95.1	--	93.2	--
[39]	83.8(100)	85.4(100)	--	--	--
[36]	94.1(35)	83.8(23)	--	91.2(34)	--
[37]	97.1(20)	83.5(20)	--	--	--
[35]	100(30)	90.3(30)	--	100(30)	--
[42]	100(3)	95.1(6)	--	98.3(6)	--
[44]	95.5	90.4	--	--	--
[43]	100(3)	96.7	--	--	--
[45]	97.3	87.10	97.40	--	--
[41]	98.6	93.5	93.6	86.2	--
[27]	100(3)	95(7)	--	100(5)	87.5(5)
Nuestro modelo	97.93(4)	95.21(7)	88.99(7)	97.88(5)	90.71(5)

de datos de cáncer de pulmón, se obtuvo una clasificación de 97.88% y para la base de datos de CNS se obtuvo 90.71%.

Para probar sí el desempeño de nuestro algoritmo es aceptable, se ha creado un estudio de comparación que se muestra en la tabla cinco. En el estudio se observa la mejor tasa de clasificación obtenida por el método propuesto y la comparación con diferentes desempeños de algunos métodos reportados en la literatura. La tabla se divide de la siguiente forma: en la primer columna, se muestran los autores con los que se han comparado los resultados obtenidos, el resto de las columnas muestran las tasas de clasificación (%) y el número de genes seleccionados (G) para las cinco bases de datos

Los resultados obtenidos se han comparado con los autores mostrados de la tabla 5. Cabe mencionar que algunos autores presentan un modelo basado en algún tipo de meta heurística como un algoritmo bioinspirado o se basan en búsquedas locales y otros generan su clasificación basados en técnicas de

Cada tasa de clasificación representada en la figura 3, se obtuvo mediante la utilización de un subconjunto de genes de alto desempeño, que ha entrenado mejor al clasificador. En cada una de las cinco bases de datos existen genes informativos, estos genes son seleccionados por el algoritmo propuesto, obteniendo un subconjunto de genes pequeño que ha logrado entrenar al clasificador de manera eficiente.

El algoritmo BC/MSV ha seleccionado un subconjunto de genes relevante para cada base de datos. Una forma de verificar si cada gen seleccionado pueden ayudar en el diagnóstico de una enfermedad, es revisando si han sido reportados en la literatura, de esta forma podemos encontrar una interpretación biológica confiable de los genes seleccionados. La base de datos de leucemia y de cáncer de colon, han sido estudiadas ampliamente, esto permite encontrar la mayoría de genes relevantes reportados por diferentes autores. En comparación con las bases de datos de cáncer de pulmón, DLBCL y CNS, el estudio de estas bases de datos su estudio es poco frecuente, consecuentemente surgen dudas para la comparación de los resultados con los genes reportados.

Tabla 6. Genes eleccionados para la bases de datos de leucemia

Gen	Gene annotation	Referencia
M27891	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	[47], [23], [11], [2], [48], [49], [34]. [50], [46], [51]
M16038	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	[6], [33]
D88422	CYSTATIN A*	[6], [41]
X07743	PLECKSTRIN	[52], [53], [54]

En el caso de leucemia cuatro genes son identificados como relevantes, debido a su nivel de expresión, estos genes tienen un rol importante dentro de la clasificación de dos tipos de leucemia aguda y así son etiquetados en la clase Leucemia Mieloide Aguda o Leucemia Linfoblastica Aguda. Los genes seleccionados se muestran en la tabla 6

Para la base de datos de cáncer de colon, el algoritmo identifica siete genes relevantes. Estos genes permiten separar la clase de tejidos tumorales de la clase de tejidos normales. Estos genes se pueden utilizar en la identificación de células con cáncer de colon. La descripción biológica de los genes seleccionados se muestra en la tabla 7.

Tabla 7. Genes eleccionados para la bases de datos de cáncer de colon

Gen	Gene annotation	Referencia
R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)	[55], [56], [52], [57], [58], [59], [60], [38], [61]
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.	[62], [63], [64], [41]
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.	[48], [62], [55], [52], [61], [65]
M63391	Human desmin gene, complete cds.	[56], [57], [58], [59], [60]
Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor.	[60], [47], [66], [52], [49], [67], [68], [69]
H43887	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)	[56], [57], [58], [59], [60]
J02854	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element TAR1 repetitive element	[70]

La tabla 8, muestra los resultados obtenidos por el algoritmo BC/MSV para el microarreglo DLBCL. El algoritmo selecciono siete datos de expresión genética que separan las clases B-like germinal y las clases B-like activado.

Tabla 8. Genes eleccionados para la bases de datos DLBCL

Gen	Gene annotation	Referencia
GENE3939X	(Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene); Clone=1355435)	[71], [72]
GENE3965X	Deoxycytidylate deaminase; Clone=489681	[72]
GENE3968X	Deoxycytidylate deaminase; Clone=1302032	[71], [72]
GENE3966X	Deoxycytidylate deaminase; Clone=489681	[73], [72]
GENE3985X	T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2; Clone=665903	[72]
GENE1252X	Cyclin D2/KIAK0002=3' end of KIAK0002 cDNA; Clone=1357360	[72], [41]
GENE3988X	Potassium voltage-gated channel, shaker-related subfamily, member 3; Clone=1337856	[72]

Un total de cinco genes relevantes han sido seleccionados de la base de cáncer de pulmón. La selección de estos genes se debe a que el clasificador logro separar correctamente la información contenida en la base de datos, esto significa que el clasificador ha logrado separar la clase Malignant Pleural Mesothelioma (MPM) de la clase Adenocarcinoma (ADCA). La tabla 9 muestra la descripción biológica de los genes seleccionados.

Tabla 9. Genes eleccionados para la bases de datos de cáncer de pulmón

Gen	Gene annotation	Referencia
AF039945	synaptojanin 2	[47], [11]
M92439	leucine-rich PPR-motif containing	[60], [11], [41], [74]
W28516	hypothetical protein MGC11308	--
D42087	RAB21, member RAS oncogene family	--
AB015228	aldehyde dehydrogenase 1 family, member A2	[75]

La tabla 10 muestra la descripción biológica de los cinco genes que el algoritmo

selecciono de la base de datos de CNS, estos genes son capaces de separar las clases de survivors y failures.

Tabla 10. Genes eleccionados para la bases de datos CNS

Gen	Gene annotation	Referencia
D63880	KIAA0159 gene	[15]
X74801	T-COMPLEX PROTEIN 1, GAMMA SUBUNIT	[76]
D64154	Mr 110,000 antigen	[11]
U70439	PHAPI2b protein	--
U69126	FUSE binding protein 2 (FBP2) mRNA, partial CDS	--

V. CONCLUSIONES

En este trabajo, se presentó un método híbrido basado en una búsqueda cuckoo combinada con una máquina de soporte vectorial, implementado en la selección y clasificación de un conjunto de genes importantes, explorando dentro de cinco bases de datos de dominio público (Leucemia, Cáncer de pulmón, DLBCL, Cáncer de Colon y CNS). El método propuesto tiene una etapa de preselección de genes mediante la utilización de cuatro técnicas de filtrado de datos, estos filtros utilizan una puntuación que sirve para discriminar los genes contenidos en la base de datos, así se eliminan los genes no relevantes (genes ruidosos o redundantes) y selecciona los genes con información pertinente. En esta etapa se ha generado una primera reducción efectiva de la dimensión de las bases de datos. Se ha creado un método de fusión de los subconjuntos obtenidos por el pre-procesamiento, en este paso, se utilizan las posiciones que los métodos de filtrado han colocado a cada gen, con la finalidad de tener un solo subconjunto de información relevante. Para explorar dentro del subconjunto obtenido al fusionar los subconjuntos de cada filtro, se ha creado un algoritmo híbrido basado en una búsqueda cuckoo como método de selección de genes combinada con una máquina de soporte vectorial como método de clasificación. Utilizando las propiedades del vuelo lévy, se ha logrado crear un algoritmo que explora el microarreglo exhaustivamente, al generar paseos que logran colocar un huevo (gen) en una dirección diferente, de esta manera se utiliza la mayoría de genes propuestos para el estudio (p-value); utilizando las propiedades de la búsqueda cuckoo, se logran eliminar de los nidos, los huevos (genes) que han sido descubiertos. Combinando esta característica con el método de clasificación, se eliminan genes que no han logrado entrenar correctamente al clasificador, dejando solo genes que han obtenido una tasa de clasificación alta.

El método propuesto determina una tasa de clasificación alta, obtenida con un subconjunto de genes pequeño para las cinco bases de datos. Para evaluar la eficiencia del método, se generó un estudio de comparación de los resultados obtenidos con otros métodos reportados en la literatura, esto permite verificar si el método es competitivo.

Se observa que en algunos casos se ha logrado superar las tasas de clasificación y se han obtenido un subconjunto de genes pequeño en comparación con los métodos reportados en la literatura. Además de las tasas de clasificación, se desea conocer los nombres de los genes reportados en la literatura, esto permite tener una mejor interpretación biológica de los genes que ha seleccionado el algoritmo. También se ha minimizado el número de genes a utilizar y en algunos casos igualado la exactitud de la clasificación utilizando la búsqueda cuckoo con el clasificador MSV, dentro del proceso de selección y clasificación de datos.

IV. METAS Y TRABAJOS FUTUROS

En trabajos futuros se modificara el algoritmo utilizado en este trabajo, se propondrá un paseo que utilice un sistema de memoria para recordar los genes utilizados. También se utilizaran otros clasificadores y aumentara el número de genes a utilizar. La meta es aumentar al máximo la exactitud de la clasificación y por otro lado, minimizar el número de genes a utilizar.

REFERENCIAS

1. I. Guyon, A. Elisseeff.: "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, pp 1157-1182, 003.
2. T. Golub, D. Slonim, P. Tamayo et al.: "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, pp. 531– 537, 1999.
3. T. Hwang, C. H. Sun, T. Yun, and G. S. Yi.: "Figs: A Filter-Based Gene Selection Workbench for Microarray Data", BMC Bioinformatics, 2010.
4. Y. Wang, I. V. Tetko, M. A Hall, E Frank, et al.: "Gene selection from microarray data for cancer classification--a machine learning approach". Comput Biol Chem, pp. 37-46, 2005.
5. A. Kulkarni, B.S.C. N. Kumar, V. Ravi, U. S. Murthy. "Colon cancer prediction with genetics profiles using evolutionary techniques", Expert Systems with Applications, pp. 2752–2757, 2011.
6. S. Li, X. Wu, M. Tan.: "Gene Selection using Hybrid Particle Swarm Optimization and Genetic Algorithm", Soft Comput, pp. 1039–1048, 2008.
7. M. S. Mohamad, et al.: "A Hybrid of Genetic Algorithm and Support Vector Machine for Features Selection and Classification of Gene Expression Microarray", International Journal of Computational Intelligence and Applications, pp. 91–107, 2005.
8. F. XU, L. WEI, W. WANG.: "Fuzzy Rough Feature Selection Based on Normalized Conditional Mutual Information", Journal of Computational Information Systems, pp. 2519–2529, 2012.
9. D. Mishra, B. Sahu,: "Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach", International Journal of Scientific & Engineering Research. (2011).
10. P. Yang, B. B Zhou, Z. Zhang, A. Zomaya.: "A Multi-filter Enhanced Genetic Ensemble System for Gene Selection and Sample Classification of Microarray Data", the Eighth Asia Pacific Bioinformatics Conference Bangalore, pp. 18-21, 2010.
11. E. Bonilla-Huerta, et al.: "Hybrid Framework using Multiple-Filters and an Embedded Approach for an Efficient Selection and Classification of Microarray Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015.
12. R. P. Rubido. "Una revisión a algoritmos de selección de atributos que tratan la redundancia en datos microarreglos". Revista Cubana de Ciencias Informáticas, pp. 16 - 30. 2013
13. Q. Huang, D. Tao, X. Li, W. C. Liew. "Parallelized Evolutionary Learning for Detection of Biclusters in Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012.
14. U. Alon, N. Barkai, D. Notterman et al.: "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays", Proc. Nat. Acad. Sci. USA, pp. 6745–6750, 1999.
15. S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, and T. Golub.: "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature, pp. 436–442, 2002.
16. G.J. Gordon et al.: "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests

Using Gene Expression Ratios in Lung Cancer and Mesothelioma", Cancer Res., 2002.

17. A. A. Alizadeh, B.M. Eisen, R.E. Davis et al.: "Distinct Types of Diffuse Large (B)–Cell Lymphoma Identified by Gene Expression Profiling", Nature, pp. 503–511, 2000.
18. TG Dietterich. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization". Machine Learning. 2000; 40:139–158.
19. W. L Martínez, A. R. Martinez: "Exploratory Data Analysis with MATLAB®". A CRC Press Company. Boca Ratón London New York Washington, D.C. (2005).
20. L. Ladha et al.: "Feature Selection Methods and Algorithms", International Journal on Computer Science and Engineering (IJCSE), pp. 1787-1797, 2011.
21. S. Dudoit, J. Fridlyand, T. Speed. "Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data", Journal of the American Statistical Association, pp. 77–87, 2002.
22. J. C. Porras-Cerrón. "Componentes Principales Supervisados Para Clasificación De Datos De Expresión Genética", Tesis de Maestro en Ciencias, Universidad De Puerto Rico Mayagüez, 2005.
23. A. H. Tan, H. Pan.: "Predictive Neural Networks for Gene Expression Data Analysis", Neural Networks, pp. 297–306, 2005.
24. P. Radivojac, Z. Obradovic, A. K. Dunker, S. Vucetic, "Feture selection filters based on the permutation test", Proc. ECML, pp. 334-346, 2004.
25. B. Kumari, T. Swarnkar, "Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review", International Journal of Computer Science and Information Technologies, Vol. 2 (3) , 2011, 1048-1053. 2012
26. X.S. Yang and S. Deb, 'Cuckoo search via Levy flights', *World Congress on Nature & Biologically Inspired Computing NaBIC'09*, 9–11 December, Coimbatore, India, pp.210–214. 2009
27. C. Gunavathi, and K. Premalatha, 'Cuckoo search optimization for feature selection in cancer classification: a new approach', Int. J. Data Mining and Bioinformatics, Vol. 13, No. 3, pp.248–265. (2015)
28. S. Wang, H. Chen, R. Li, D. Zhang. "Gene Selection with Rough Sets for the Molecular Diagnosing of Tumor Base on Support Vector Machines", International Computer Symposium, pp. 1368-1373, 2006.
29. L. K. Lou, D. F. Huang, L. J. Ye, Q. F. Zhou, G. F. Sheo, F Peng. "Improving the Computational Efficieny of Cluster Elimination for Gene Selection". IEEE/ACM Trans. Comput. Bioinform. 8(1): 122-129. 2011.
30. G. Kulshestha A. Agarwal A. Mittal A. Sahoo Hybrid cuckoo search algorithm for simultaneous feature and classifier selection, International Conference on Cognitive Computing and Information Processing (CCIP), IEEE, pp. 1 – 6, 2015,
31. J. C., Hernández-Hernández, B. J., Duval, K. Hao "SVM-based local search for gene selection and classification of microarray data". Comunicativos in Computer and Information Science, Vol. 13. pp. 499–508. 2008.

32. Yu, L., Han, Y. and Berens M. E.: "Stable gene selection from microarray data via sample weighting", IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 262-272, 2012.
33. M., Filippone, F., Masulli, S. Rovetta, "Simulated Annealing for Supervised Gene Selection", Soft Computing, pp. 1471–1482, 2011.
34. S.-B. Cho, and H.-H Won: Cancer classification using ensemble of neural networks with multiple significant gene subsets. In Applied Intelligence, 26(3):243–250, 2007.
35. L., Zhang, Z., Li, and H. Chen, "An effective gene selection method based on relevance analysis and discernibility matrix." In PAKDD, volume 4426 of Lecture Notes in Computer Science, pages 1088– 1095, 2007.
36. S., Pang, I., Havukkala, Y Hu and N. Kasabov. "Classification consistency analysis for bootstrapping gene selection". In Neural Computing and Applications, 16:527,539, (2007).
37. G-Z., Li, X-Q Zeng, J.Y Yang, and M. Q Yang. "Partial least squares based dimension reduction with gene selection for tumor classification". In Proceedings of IEEE 7th International Symposium on Bioinformatics and Bioengineering, pages 1439–1444, 2007.
38. A. C. Tan, and D.Gilbert: "Ensemble machine learning on gene expression data for cancer classification". In Applied Bioinformatics, 2(2):75–83, 2003.
39. F., Yue, K., Wang, and W. Zuo, "Informative gene selection and tumor classification by null space LDA for microarray data". In ESCAPE'07, volume 4614 of Lecture Notes in Computer Science, pages 435–446. Springer, 2007.
40. G. Yu , Y. Feng, D. J. Miller, J. Xuan, E. P. Hoffman, R. Clarke, B. Davidson, I. M. Shih, Y. Wang.: "Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases", Journal of Machine Learning Research, pp. 2141-2167, 2010.
41. L., Sun, D., Miao, H. Zhang "Gene Selection and Cancer Classification: A Rough Sets Based Approach". Transactions on Rough Sets XII. LNCS Springer, Heidelberg, vol. 6190, pp. 106–116. 2010
42. Y. Leung and Y. Hung. "A Multiple-filter-multiple-wrapper Approach to Gene Selection and Microarray Data Classification". In IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(1):108117, 2010.
43. S-L. Wang, X. Li, S. Zhang, J. Gui and D-S. Huang. "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction". In Computers in Biology and Medicine. 40,179-189, 2010.
44. S-W. Zhang, D-S. Huang and S. L. Wang. "A method of tumor classification based on wavelet packet transforms and neighborhood rough set". In Computers in Biology and Medicine, 40, 420–437, 2010.
45. C. H. Zheng, L. Zhang, V. T. Ng, S. C. Shiu and D. S. Huang. "Metasample Based sparse representation for tumor classification". In IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(5):1273–1282. 2011.
46. 52 S-L. Wang, L. Sun, and J. Fang. "Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification". In BMC Bioinformatics, 13(178): 1–26, 2013.

47. X. Wang, O. Gotoh.: "Cancer Classification using Single Genes", Genome Informatics, pp. 176-188, 2009.
48. L.F. Wessels, M.J.T. Reinders, T. Van-Welsem, P.M. Nederlof and Y. Wang. "Representation and classification for high-throughput data". In SPIE., 4626:226–237, 2002.
49. W. Chu, Z. Ghahramani, F. Falciani and D.L. Wild. "Biomarker discovery in microarray gene expression with gaussian process". In Bioinformatics, 21(16):3385–3393, 2005.
50. K. Deb and R. Reddy. "Reliable classification of two-class cancer data using evolutionary algorithms". In BioSystems, 72(1):111–129, 2003.
51. P. Yang, B. Zhou, Z. Zhang and A.Y. Zomaya. "A multi-filter enhanced genetic ensemble system for gene selection and sample classificationof microarray data". In BMC Bioinformatics, 11(55):1– 12, 2010.
52. Ben-Dor, L. Bruhn, et al. "Tissue classification with gene expression profiles". In Journal of Computational Biology, 7(3-4):559–583, 2000.
53. Y. Wang, F.S. Makedon, J.C. Ford and J. Pearlman. HykGene: "A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data". In Bioinformatics, 21(8):1530–1537, 2005.
54. S.A. Vinterbo, E-Y. Kim and L. Ohno-Machao. "Small, fuzzy and interpretable gene expression based classifiers". In Bioinformatics, 21(9):1964–1970, 2005.
55. V. Roth, "The Generalized LASSO: a wrapper approach to gene selection for microarray data". University of Bonn, Computer Science III, Roemerstr. Bonn Germany. August, 2002.
56. H. Zhang, X. Song, and H. Wang, y X. Zhang. "Miclique: An Algorithm to Identify Differentially Co-expressed Disease Gene Subset from Microarray Data". Journal of Biomedicine and Biotechnology, 2009.
57. L. Li, T. A. Darden, C. R. Weinberg, A. J. Levine y L. G. Pedersen.: "Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/K-Nearest Neighbor Method", Combinatorial Chemistry & High Throughput Screening, pp. 727-739, 2001.
58. S. Li, X. Wu, X. Hu.: "Gene selection using genetic algorithm and support vectors machines", Soft Comput, pp. 693-698, 2008.
59. F. Tan, X. Fu, Y. Zhang, y A. G. Bourgeois.: "Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data". IEEE Congress on Evolutionary Computation, pp. 2529-2534, 2006.
60. X. Wang, O. Gotoh. "Inference of cancer-specific gene regulatory networks using soft computing rules". Gene Regul Syst Biol. pp. 19–34, 2010.
61. T. M. Huang, V. Kecman.: "Gene Extraction for Cancer Diagnosis by Support Vector Machines–an Improvement". Artificial Intelligence in Medicine. pp. 185-194, 2005.
62. B. Krishnapuram, L. Carin, A. J. Hartemink.: "Joint Classifier and Feature Optimization for Comprehensive Cancer Diagnosis Using Gene Expression Data". J. Comput. Biol., To Appear, 2004.
63. R. Maglietta, A. D'Addabbo, A. Piepoli, BF. Perri et al. "Selection of relevant genes in cancer diagnosis based on their prediction accuracy". Artif Intell Med 40:29–44. 2007.
64. A. SUNDARAM, N. L.VENKATA, & R. S. PARTHASARATHY. "Hybrid SPR algorithm to select predictive genes for

effectual cancer classification". Turkish Journal of Electrical Engineering & Computer Sciences, 21(2). 2013

65. J. J. Chen, C.A. Tsai, S.L. Tzeng y C.H. Chen: "Gene Selection with Multiple Ordering Criteria". BMC Bioinformatics, 8:74, 2007.
66. J-M. Arevalillo and H. Navarro. "A new approach for detecting bivariate interactions in high-dimensional data using quadratic discriminant analysis". In BIOKDD10, pages 1–7, 2010.
67. Z. Guan and H. Zhao. "A semiparametric approach for marker gene selection based on gene expression data". In Bioinformatics, 24(4):529–536, 2005.
68. Y. Tang, Y. Zhang, and Z. Huang. "Development of two-stage SVMRFE gene selection strategy for microarray expression data analysis". In IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(3):365–381, 2007.
69. H-H. Li, Y-Z. Liang et al. "Recipe for Uncovering Predictive Genes Using Support Vector Machines Based on Model Population Analysis". In IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(6):1633-1641.2011.
70. K. Do: "Applications of gene shaving and mixture models to cluster microarray gene expression data". Cancer Informatics, 2: 25–43. 2007
71. J. S. Aguilar-Ruiz, F. Azuaje, and J. C. Riquelme Santos: "Data Mining Approaches to Diffuse Large B-Cell Lymphoma Gene Expression Data Interpretation", Lecture Notes in Computer Science 3181, Springer. Pp. 279-288. 2004
72. S, Baek. H, Moon. H. Ahn, et al. "Identifying high-dimensional biomarkers for personalized medicine via variable importance ranking", J Biopharm Stat, vol. 18 pg. 853-68. 2008
73. Y. Wang, IV Tetko, M. A Hall, E Frank, et al, "Gene selection from microarray data for cancer classification--a machine learning approach". Comput Biol Chem, pp. 37-46, 2005.
74. I. K., Yoon, H. K., Kim, Y. K., Kim, Song et. al. "Exploration of replicative senescence-associated genes in human dermal fibroblasts by cDNA microarray technology". Experimental gerontology, 39(9), 1369-1378. 2004.
75. VB Mahajan, C Wei, PJ McDonnell. "Microarray analysis of corneal fibroblast gene expression after interleukin-1 treatment. Invest Ophthalmol". Vis Sci; 43: 2143-2151. 2002
76. K. Iwao-Koizumi, R. Matoba, N. Ueno, J. K. Seung., et al, "Prediction of Docetaxel Response in Human Breast Cancer by Gene Expression Profiling". Journal of Clinical Oncology, pp. 422-431, 2005.

Acerca de los autores

L**uis Alberto Hernández Montiel**, se recibe como Licenciado en informática en abril del 2011 por el Instituto Tecnológico de Apizaco, obtiene el grado de maestro en sistemas computacionales en noviembre del 2013 por el Instituto Tecnológico de Apizaco, Apizaco Tlaxcala México. Actualmente es profesor-investigador de la Licenciatura en Informática en la universidad del istmo campus Ixtepec, Oaxaca, México. Sus áreas de interés son: Algoritmos Evolutivos, Metaheurísticas, Optimización y Bioinformática.

Carlos Edgardo Cruz Pérez. Obtuvo el título de Ingeniero en Electrónica por el Instituto Tecnológico de Oaxaca, obtiene el grado de Maestro en Ingeniera en Tecnologías de la Información otorgada por la Universidad Anáhuac. Tiene la especialidad en seguridad de la información otorgada por el INACIPE y certificaciones como UBWA y Perito en Informática Forense. Actualmente es Profesor de Tiempo Completo en la universidad del Istmo (UNISTMO) Campus Ixtepec, donde ha dirigido proyectos de evaluación y desempeño de redes de computadoras y sistemas expertos enfocados al diagnóstico de diabetes. Sus líneas de investigación son: Ruteo en redes de computadora, VoIP, Seguridad y Cifrado de la información.

Luis David Hernández Huerta, se graduó como Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Tehuacán en México en 2000. Recibió una maestría en ciencias computacionales por el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) México en 2007, es especialista en Soft Computing. Trabajo en el instituto de investigación y desarrollo de la armada mexicana. En la actualidad, es profesor-investigador en la Universidad del Istmo campues Ixtepec, Oaxaca, México. El área de investigación es acerca de la aplicación de algoritmos evolutivos, lógica difusa, redes neuronales y procesamiento del habla.