ISSN 1870-4069

Selección y clasificación de genes cancerígenos utilizando un método híbrido filtro/wrapper

Luis Alberto Hernández Montiel, Carlos Edgardo Cruz Pérez, Juan Gabriel Ruiz Ruiz

Universidad del Istmo Campus Ixtepec, Ciudad Ixtepec, Oaxaca, México

luisahm@itamail.itapizaco.edu.mx, carloscruz@bianni.unistmo.edu.mx, jugaruizr@gmail.com

Resumen. En este artículo, se presenta un método híbrido aplicado en la selección y clasificación de genes obtenidos de microarreglos de ADN. El método realiza una limpieza del microarreglo utilizando un preprocesamiento basado en técnicas de filtrado de datos, después, se realiza una selección de genes dentro del subconjunto obtenido por el filtro, utilizando una búsqueda gravitacional combinada con un clasificador KNN. El método se ha construido para obtener un subconjunto de genes relevantes de alto desempeño, los resultados obtenidos se comparan con diferentes métodos reportados en la literatura, este método es implementado en cinco microarreglos de dominio público.

Palabras clave: microarreglos de ADN, filtrado de datos, selección, clasificación, búsqueda gravitacional.

Cancerous Genes Selection and Classification Using a Hybrid Filter/Wrapper Method

Abstract: In this paper, a hybrid method applied in selection and classification of genes obtained from DNA-microarray is presented. The method performs a microarray cleaning using a preprocessing based on data filtering techniques, then, a gene selection is performed within subset obtained from a filter, using a gravitational search combined a SVM-classifier. This method has been built for obtain a relevant genes subset with high performance, the results obtained are compared with other methods reported in literature. This method is tested using five public domain microarrays.

Keywords: DNA-microarray, data filtering, selection, classification, gravitational search.

1. Introducción

Actualmente existen tecnologías que apoyan la búsqueda de enfermedades degenerativas (como el cáncer) [1]. Una de las tecnologías más utilizadas son los

microarreglos de ADN, esta tecnología mide los niveles de expresión genética de miles de genes simultáneamente [2]. Diferentes estudios indican que los perfiles de expresión genética proporcionan información para distinguir un tipo de cáncer dentro de tejido morfológicamente similar y generar un mejor diagnóstico y proponer terapias para contraatacar la enfermedad [3]. La clasificación de tumores se asocia con el problema de selección de genes, el objetivo es extraer un conjunto de genes relevantes de un microarreglo. Sin embargo, esta no es una tarea fácil, el microarreglo contiene genes informativos y ruidosos [4]. Por ello es necesaria la identificación de biomarcadores para la elaboración de pruebas de diagnóstico. Para solucionar el problema de la selección y clasificación de genes se propone un método basado en filtros y wrapper. El método genera un pre-procesamiento basado en la puntuación que otorga una estrategia de filtro para hacer una limpieza del microarreglo, después se implementa una búsqueda gravitacional combinada con clasificador KNN para seleccionar los genes con mejor tasa de desempeño dentro del subconjunto obtenido por el filtro. Con este método, se buscan los genes más relevantes para el diagnóstico de cáncer dentro de cinco microarreglos de ADN de dominio público.

2. Estado del arte

La tecnología de microarreglos de ADN manipula grandes volúmenes de información genética perteneciente a varios tipos de enfermedades [2]. Esta información contiene un gran número de datos que requieren un largo tiempo de procesamiento. La alta dimensión del microarreglo genera un problema de precisión de análisis y complejidad computacional [5]. Los microarreglos no solo cuentan con una gran cantidad de atributos (genes) y un número limitado de muestras, también contienen dos o más número de clases (categorías) a las que pertenece cada uno de los atributos, además miles de los genes son redundantes o ruidosos [6].

Para solucionar este problema, diferentes autores han propuesto métodos basados en técnicas de minería de datos (como selección y extracción de características) y de aprendizaje máquina [7] que ayudan a obtener información relevante a través de la exploración del microarreglo.

El método más utilizado es la clasificación de características [8] con esta técnica se logra distinguir entre varias clases de muestras de tejido que contienen alguna enfermedad.

El aprendizaje evolutivo [9] utiliza las propiedades de los algoritmos bioinspirados [10, 11] combinados con algún método de clasificación y así exploran el microarreglo buscando información relevante para el diagnóstico de alguna enfermedad.

Los algoritmos de aprendizaje máquina [12] basados en métodos de filtrado y/o wrapper, eliminan la información menos relevante dentro del microarreglo utilizando una puntuación que sirve como valor de pertinencia del gen, seleccionando solo información confiable para su análisis.

Las técnicas de clusters [13] agrupan los genes similares, después se utiliza algún método de puntuación discriminante que ayude a eliminar la información no relevante del microarreglo.

Actualmente se implementan algoritmos híbridos [14, 15] combinando diferentes técnicas de filtrado de datos y algoritmos wrapper que trabajan de forma paralela,

aplicados en la selección y extracción de genes relevantes para el diagnóstico de una enfermedad. A pesar del número de técnicas implementadas para abordar el problema de selección de genes, no se ha llegado a una solución concreta, cada modelo o método presentado selecciona genes que no se han reportado y abre el panorama a nuevos estudios, debido a esto surgen más trabajos con nuevas propuestas dando un estudio más confiable de los genes seleccionados.

3. Materiales y métodos

Los microarreglos de ADN contienen información relevante mezclada con información ruidosa y redundante [7] esto genera tiempo de procesamiento largo y dificulta para extraer información valiosa de ellos con resultados poco confiables. Para abordar este problema se propone un método híbrido combinando técnicas de filtrado de datos como primera etapa de selección. En la segunda etapa se implementa un modelo de selección y clasificación utilizando una búsqueda gravitacional combinada con un clasificador KNN, los materiales y métodos utilizados se describen a continuación. La figura 1 muestra la estructura general del método propuesto.

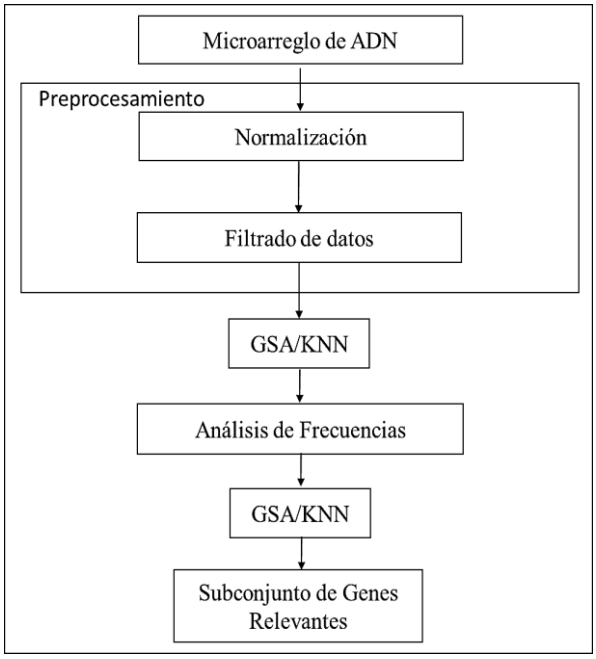

Fig. 1. Método general de selección y clasificación.

3.1. Microarreglos de ADN

Los datos obtenidos de un microarreglo de ADN se obtienen mediante una matriz donde las filas representan los genes y las columnas representan las muestras. Cada

celda dentro de la matriz es un valor de expresión genética que representa la intensidad del gen correspondiente a cada muestra [16]. Lo anterior se observa en la figura 2 donde ***x*** representa los datos genómicos, ***ng*** (número de gen, filas) los genes dentro de la matriz y ***nm*** (número de muestras, columnas) las muestras dentro de la matriz.

$$x_{ij} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots\ \dots\ \dots & x_{1n_m} \\ x_{21} & x_{22} & x_{23} & \dots\ \dots\ \dots & x_{2n_m} \\ x_{31} & x_{32} & x_{33} & \dots\ \dots\ \dots & x_{3n_m} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n_g 1} & . & . & . & x_{n_g n_m} \end{bmatrix}$$

Fig. 2. Matriz de datos de expresión genética.

En éste trabajo, se utilizan cinco bases obtenidas de la tecnología de microarreglos de ADN, descritas en la tabla 1.

Tabla 1. Características de los microarreglos de ADN.

Microarreglo de ADN	Genes	Muestras	Casos	Controles	Referencia
Leucemia	7128	72	25 ALL	47 AML	[8]
Colon	2000	62	22 Tumor	44 Normal	[3]
Pulmón	12533	181	31 MPM	150 ADCA	[17]
CNS	7129	60	21 survivors	39 failures	[18]
DLBCL	4026	47	23 B-like Activado	24 B-like germinal	[19]

3.2. Pre-procesamiento

La información original de un microarreglo se encuentra en diferentes escalas numéricas o diferentes distribuciones de probabilidad. Para solucionar este problema se genera una transformación de los datos a un rango entre cero y uno para facilitar su estudio. En este trabajo como primer paso del pre-procesamiento se realiza una normalización basada en una técnica min-máx. [20]:

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

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.

El segundo paso de la etapa de pre-procesamiento es realizar una primera selección de genes utilizando técnicas de filtrado. La idea es que cada método genere un ranking de los genes del microarreglo asignándole un valor de pertinencia que ayude a discriminar los genes relevantes de los no relevantes. Los métodos utilizados en este experimento se describen a continuación.

-**BSS/WSS**: El método selección genes basado en la razón de la suma de cuadrados entre clases (BSS) y dentro de las clases (WSS). Para el gen ***j***, la razón está dada por [21]:

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

donde $\acute{x}_{.j}$denota el nivel medio de la expresión del gen ***j*** a través de todas las muestras y $\acute{x}_{kj}$ 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) En este método se identifican los patrones de expresión genética con diferencia máxima en la expresión media entre dos grupos y la variación mínima de expresión dentro de cada grupo, los genes se clasifican de acuerdo a sus niveles de expresión [22].

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

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.

-Información mutua: Este método toma dos genes (A y B) de forma aleatoria con distribuciones de probabilidad diferentes y una distribución de probabilidad conjunta. La información mutua entre ambos genes I(A;B) se define como la entropía relativa entre la probabilidad conjunta y el producto de probabilidades [23].

$$I(A;B) = \sum_{a_i} \sum_{b_j} P(a_i,b_j) log \frac{P(a_i,b_j)}{P\ (a_i)P(b_j)}, \quad (4)$$

donde $P(a_i,b_j)$ es la probabilidad conjunta de los genes, $P(a_i)$ es la probabilidad del gen A y $P(b_j)$ es la probabilidad del gen B.

Para la segunda etapa del método propuesto, se ha implementado un algoritmo de selección de genes basado en una búsqueda gravitacional combinada con un clasificador KNN, el método se entrena con los conjuntos obtenidos por los filtros y se describe a continuación.

3.3. Búsqueda gravitacional

Es un algoritmo de búsqueda propuesto por Rashedi en 2009 [15, 24]. El GSA está basado en la ley de gravedad de Newton, donde la fuerza de gravedad entre dos cuerpos es directamente proporcional al producto de sus masas e inversamente proporcional al cuadrado de su distancia. Las soluciones en la población del GSA se llaman agentes que interactúan entre sí a través de la fuerza de gravedad. La calidad de cada agente se mide por su masa. Cada agente se considera como objeto y todos los objetos se mueven hacia otros objetos con masa más pesada debido a la fuerza de gravedad que generan, este paso representa un movimiento global (paso de exploración) del objeto, mientras que el agente con una masa pesada se mueve lentamente, lo que representa la etapa de explotación del algoritmo. La mejor solución encontrada por el algoritmo es el agente con la masa más pesada [25].

La constante de gravedad G en la iteración t que utiliza el algoritmo se calcula por [15]:

$$G(t) = G_0 e^{-\alpha t/T}, \tag{5}$$

donde G_0 y α son inicializados al comienzo de la búsqueda y sus valores se irán reduciendo durante cada iteración. T es el número total de iteraciones.

Dentro del algoritmo existen tres tipos de masas: Masa gravitacional activa Ma, masa gravitacional Pasiva Mp y masa inercial Mi. la fuerza de gravedad F_{ij} que actúa sobre las masas i y j obedeciendo la ley de gravedad de Newton definida por:

$$F = G\frac{M_{aj}\ x\ M_{pi}}{R^2}, \tag{6}$$

donde M_{aj}, M_{pi} son la masa activa y pasiva del objeto j, i, respectivamente.

De acuerdo con de la segunda ley de Newton, cuando se aplica una fuerza F a un objeto, el objeto se mueve con aceleración a dependiendo de la fuerza aplicada y la masa del objeto M, la aceleración del objeto i (agente) se calcula como sigue:

$$a_i = \frac{F_{ij}}{M_{ii}}, \tag{7}$$

donde M_{ii} es masa de inercial del agente i.

Los agentes actualizan sus velocidades y posiciones, como se muestra en las ecuaciones 8 y 9 respectivamente:

$$V_i(t+1) = rand_i * V_i(t) + a_i(t), \tag{8}$$

$$X_i(t+1) = rand_i * V_i(t) + a_i(t). \tag{9}$$

3.4. Clasificador KNN

El clasificador k-vecino más cercano (KNN por k-Nearest Neighbor) es un algoritmo de clasificación que basa su criterio de aprendizaje en la hipótesis de que los miembros de una población suelen compartir propiedades y características con los individuos que los rodean [25] de modo que es posible obtener información descriptiva de un individuo mediante la observación de sus vecinos más cercanos.

La regla de clasificación por KNN se describe a continuación: sea x^1, x^2,…, x^n una muestra con una función $f(x)$ de densidad desconocida. Se estima $f(x)$ a partir de un elemento central de la muestra x que crece hasta contener k elementos con una distancia euclidiana similar, donde el valor de k se define arbitrariamente. Estas observaciones son los k vecinos más cercanos a x. Se tiene entonces la siguiente condición [15, 26]:

$$\hat{f}(x) = \frac{\frac{k}{n}}{V_k(x)}, \tag{10}$$

donde $V_k \quad (x)$ es el volumen de un elipsoide centrado en $\boldsymbol{x}$, y de radio la distancia euclidiana de $\boldsymbol{x}$ al ***k-ésimo*** vecino más cercano.

3.5. Implementación de algoritmo híbrido GSA/KNN

En nuestro caso, el algoritmo híbrido se implementa de la siguiente forma [15].

Paso 1. Se establecen los valores iniciales, la constante de gravedad $\boldsymbol{G_0}$, $\boldsymbol{\alpha}$, $\boldsymbol{\varepsilon}$ y el número de iteraciones ***t***.

Paso 2. La población inicial se genera aleatoriamente siguiendo una distribución uniforme, se compone de N agentes asociados con cada gen dentro del microarreglo.

Paso 3. El clasificador KNN se introduce en la función de costo de algoritmo para evaluar los agentes de la población, para verificar el error del clasificador se utiliza un k-fold cross validation.

Paso 4. La constante de gravedad se actualiza como se muestra en la ecuación 5.

Paso 5. Cuando el agente ***j*** actúa sobre el agente ***i*** con fuerza, en un tiempo específico ***(t),*** la fuerza es calculada por:

$$F_{ij}^{d}(t) = G(t)\frac{M_{pi}(t)xM_{aj}(t)}{R_{ij}(t)+\epsilon}\left(x_j^d(t)-x_i^d(t)\right), \tag{11}$$

donde M_{aj} es la masa gravitacional activa del agente ***j***, M_{pi} es la masa gravitacional pasiva del agente ***i***, ***G(t)*** es constante de gravedad en el tiempo ***t***.

Paso 6. En cada iteración **t**, la fuerza total que actúa sobre el agente ***i*** se calcula por:

$$F_i^d(t) = \sum_{j\epsilon Kbest, j\neq i} rand_j F_{ij}^d(t), \tag{12}$$

donde ***Kbest*** es el conjunto de los primeros ***K*** agentes con la masa más grande.

Paso 7. La masa inercial es calculada por:

$$m_i(t) = \frac{fit_i - worst(t)}{best(t) - worst(t)}, \tag{13}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)}. \tag{14}$$

Paso 8 La aceleración de agente ***i*** se calcula por la ecuación 7, la velocidad y la posición de agente ***i*** se calculan con las ecuaciones 8 y 9. Se incrementa el número de iteraciones hasta cumplir con los criterios de paro, se produce el mejor subconjunto de agentes (genes) con la mejor aptitud.

4. Protocolo y resultados experimentales

El método propuesto se entrena con cinco microarreglos de ADN descritos en la tabla 1. Se observa que el método es capaz de seleccionar genes con información que

ayuda con el diagnóstico de un tipo de cáncer. En esta sección se muestran los parámetros utilizados y los resultados obtenidos por el método propuesto.

4.1. Parámetros y resultados

El protocolo experimental se realizó en un pc DELL vostro con procesador i5 y memoria RAM de 4gb. El algoritmo fue implementado en Matlab versión 7.12. Los parámetros más confiables se muestran en la tabla 2.

Tabla 2. Parámetros utilizados por el algoritmo híbrido.

PARAMETROS	
Número de agentes	50
Dimensión	300
Número de iteraciones	500

El protocolo experimental se dividió en dos etapas, en la primera se utilizan tres métodos de filtrado de datos que funcionan como una fase de preselección descartando genes ruidosos o redundantes y obtenido como resultado subconjuntos con información relevante, de este proceso se han obtenido un subconjunto de genes. En la siguiente etapa se presenta un algoritmo basado en una búsqueda gravitacional, este algoritmo logra explorar y explotar el espacio de búsqueda dentro del subconjunto obtenido en la etapa de filtrado seleccionando genes para entrenar un clasificador basado en un vecino más cercano (KNN). El clasificador KNN se ha introducido en la función de costo del algoritmo con el objetivo de saber sí el gen seleccionado es relevante para distinguir algún tipo de cáncer. La combinación de estas dos técnicas permite seleccionar genes utilizando la tasa de clasificación como método de discriminación y eliminando los genes que no logren entrenar al clasificador, seleccionando genes que obtienen una tasa de clasificación alta. De esta forma se logra reducir el tamaño del microarreglo y se genera un subconjunto de genes que contiene información relevante para el diagnóstico de una enfermedad.

Dentro de cada microarreglo el algoritmo GSA/KNN ha seleccionado un pequeño conjunto de genes informativos, para saber si estos genes son relevantes para el diagnóstico de cáncer, se revisa sí han sido reportados en la literatura obteniendo una interpretación biológica confiable de cada gen.

Tabla 3. Tasa de clasificación obtenida por el método propuesto.

Microarreglos	**Etiqueta**	**Nombre del Gen**	**Reportado**
Leucemia	X95735	Zyxin	[27], [8]

En el microarreglo de leucemia el algoritmo selecciona solo un gen que debido a su nivel de expresión logra identificar dos tipos de leucemia aguda y así ser clasificado en la clase Leucemia Mieloide Aguda o Leucemia Linfoblastica Aguda. La tabla 3 muestra la descripción del gen seleccionado por el algoritmo GSA/KNN.

Tabla 4. Tasa de clasificación obtenida por el método propuesto.

Microarreglo	Etiqueta	Nombre del Gen	Reportado
Cáncer de Colon	R87126	*MYOSIN HEAVY CHAIN, ONMUSCLE (Gallus gallus)*	[28], [29].
	M76378	*Human cysteine-rich protein (CRP) gene, exons 5 and 6*	

La tabla 5 muestra la descripción del gen seleccionado por el algoritmo para el microarreglo de cáncer de pulmón, logrando separar la clase Malignant Pleural Mesothelioma (MPM) de la clase Adenocarcinoma (ADCA).

Tabla 5. Descripción de los genes seleccionados para Cáncer de pulmón.

Microarreglo	Etiqueta	Nombre del Gen	Reportado
Cáncer de Pulmón	X99270	three prime repair exonuclease 2	[--]

La tabla 6 y la tabla 7 muestran la descripción de los genes seleccionados por el algoritmo para el microarreglo DLBCL y el microarreglo CNS.

Tabla 6. Descripción de los genes seleccionados para el microarreglo DLBCL.

Microarreglo	Etiqueta	Nombre del Gen	Reportado
DLBCL	GENE3327X	Unknown UG Hs.169565 ESTs, Moderately similar to ALU SUBFAMILY SB WARNING ENTRY [H.sapiens]; Clone=825217	[30]
	GENE3261X	*Unknown; Clone=1353015	
	GENE3330X	*Unknown; Clone=825199	
	GENE3329X	Unknown UG Hs.224323 ESTs, Moderately similar to alternatively spliced product using exon 13A [H.sapiens]; Clone=1338448	
	GENE477X	*Putative oncogene protein similar to C. elegans ZC395.7 gene product; Clone=590942	

Tabla 7. Descripción de los genes seleccionados para el microarreglo CNS.

Microarreglo	Etiqueta	Nombre del Gen	Reportado
CNS	L17131	*High mobility group protein (HMG-I(Y)) gene exons 1-8*	[19]
	U69126	*FUSE binding protein 2 (FBP2) mRNA, partial cds*	
	X03689	*mRNA fragment for elongation factor TU (N-terminus)*	
	M64347	*FGFR3 Fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism)*	
	X93510	*37 kDa LIM domain protein*	

En cáncer de colon, el algoritmo selecciona dos genes con información relevante para identificar células con cáncer de colon, separando las células de tejidos cancerosos de los tejidos normales. La tabla 4 muestra la descripción de los genes seleccionados.

Una forma de verificar sí el algoritmo propuesto es competente, es comparando las tasas de clasificación obtenidas por el algoritmo GSA/KNN con diferentes tasas de clasificación reportadas por diferentes autores que han utilizado un modelo similar al propuesto en este trabajo. La tabla 8 muestra el estudio de comparación de las tasas de clasificación obtenidas con otros métodos reportados en la literatura. Al comparar los resultados, se verificar que en algunos casos el algoritmo ha superado las tasas de clasificación y/o reduciendo el número de genes a utilizar en comparación con los métodos reportados en la literatura.

Tabla 8. Estudio de comparación.

AUTOR	Leucemia	Colon	DLBCL	Pulmón	CNS
	%(G)	%(G)	%(G)	%(G)	%(G)
Hernández et al [31]	92.52% (6)	87.00%(8)	--	--	95.44% (12)
Filippone et al [32]	94.7% (13)	80.6(21)	--	--	--
Gunavathi [33]	100% (3)	95.00%(7)	--	100%(5)	87.5%(5)
Bonilla-Huerta [34]	97.5%(3)	90.5% (3)	96.00% (3)	93.8% (3)	94.30% (4)
Li et al [35]	95.1%(21)	88.7%(16)	--	--	--
Yu et al. [36]	--	--	--	92.86% (71)	--
Pang et al. [37]	94.1(35)	83.8(23)	--	91.2(34)	--
GSA/KNN	98% (**1**)	94.77% (**2**)	**97.72% (1)**	86.43% (5)	86.33% (5)

T**abla 9**. Tiempo de ejecución de los algoritmos implementados.

Algoritmo	Leucemia	Colon	DLBCL	Pulmón	CNS
AG/KNN	2452.189090 segundos.	2343.115777 segundos.	2243.731505 segundos.	2627.635081 segundos.	2449.328465 segundos.
BT/KNN	2580.345980 segundos.	6255.776846 segundos.	10696.475925 segundos.	4021.763980 segundos.	12450.260821 segundos.
BC/KNN	837.955481 segundos.	801.900155 segundos.	795.473768 segundos.	885.674457 segundos.	816.134201 segundos.
GSA/KNN	**430.297611** segundos.	**438.097286** segundos.	**437.329687** segundos.	**458.215342** segundos.	**438.984319** segundos.

Cada algoritmo propuesto para solucionar el problema de selección y clasificación de genes genera un costo computacional que se puede medir en tiempo-máquina por segundos, en este trabajo se hace un estudio de comparación del tiempo de respuesta del método propuesto y diferentes algoritmos que se han utilizado para solucionar este problema. Los algoritmos con los que se ha comparado son un Algoritmo Genético (AG), una Búsqueda Tabú (BT), una Búsqueda Cuckoo (BC) combinados con un clasificador KNN, la tabla 9 muestra el tiempo de ejecución en segundos de cada algoritmo implementado trabajando bajo las mismas condiciones presentadas en la sección 4.1. Se observa que el algoritmo propuesto GSA/KNN tiene el menor tiempo de ejecución en comparación de los algoritmos comparados.

5. Conclusiones y trabajos futuros

En este trabajo se presentó un método híbrido basado en técnicas de filtro y un algoritmo wrapper implementado en la selección y clasificación de un conjunto de

genes relevantes, explorando dentro de cinco microarreglos de ADN de dominio público (Leucemia, Cáncer de Colon, Cáncer de Pulmón, DLBCL y CNS). El método propuesto tiene una etapa de preselección de genes, utilizando tres técnicas de filtrado de datos, estos filtros generan una puntuación que sirve para eliminar los genes no relevantes (genes ruidosos o redundantes) y selecciona subconjuntos de genes con información pertinente.

Para realizar la selección dentro del subconjunto obtenido por el proceso de filtrado, se ha implementado un método wrapper basado en una búsqueda gravitacional como método de selección combinada con clasificador KNN como método de clasificación. Utilizando las propiedades de la búsqueda se ha logrado implementar un algoritmo que explora el microarreglo exhaustivamente, colocando agentes (genes) que son atraídos por la fuerza de gravedad de agentes con masas más grandes (mejores soluciones) logrando recorrer cada lugar en un espacio dimensional (espacio de soluciones), de esta manera se utiliza la mayoría de genes propuestos para el estudio (p-value) eliminando agentes (genes) que tienen una masa pequeña (malas soluciones). Combinando esta característica con el método de clasificación se seleccionan genes que han obtenido una tasa de clasificación alta.

Referencias

1. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, pp. 1157–1182 (2003)
2. Moreno, V., Solé, X.: Uso de Chips de ADN (Microarrays) en Medicina: Fundamentos Técnicos y Procedimientos Básicos para el Análisis Estadístico de Resultados. Unidad de Bioestadística y Bioinformática, Instituto Catalán de Oncología, Barcelona España (2000)
3. Alon, U., Barkai, N., Notterman, D.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. USA., 96, pp. 6745–6750 (1999)
4. Ben-Dor, A., Bruhn, L., Friedman, N.: Tissue classification with gene expression profiles. Journal of Computational Biology, 7(3-4), pp. 559–583 (2000)
5. Rubido, R. P.: Una revisión a algoritmos de selección de atributos que tratan la redundancia en datos de microarreglos. Revista Cubana de Ciencias Informáticas, pp. 16–30 (2013)
6. Hwang, T., Sun, C. H., Yun, T., Yi, G. S.: Figs: A Filter-Based Gene Selection Workbench for Microarray Data. BMC Bioinformatics (2010)
7. Zhang, Y., Ding, C., Li, T.: Gene Selection Algorithm by Combining ReliefF and mRMR. IEEE 7th International Conference on Bioinformatics and Bioengineering, MA, USA (2008)
8. Golub, T., Slonim, D., Tamayo, P.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, pp. 531–537 (1999)
9. Kulkarni, A., Kumar, B. S. C.: Colon cancer prediction with genetics profiles using evolutionary techniques. Expert Systems with Applications, pp. 2752–2757 (2011)
10. Rakkiannan, T., Palanisamy, B.: Hybridization of Genetic Algorithm with Parallel Implementation of Simulated Annealing for Job Shop Scheduling. American Journal of Applied Sciences, pp. 1694–1705 (2012)
11. Li, S., Wu, X., Tan, M.: Gene Selection using Hybrid Particle Swarm Optimization and Genetic Algorithm. Soft Comput, pp. 1039–1048 (2008)

12. Yu, G., Feng, Y., Miller, D. J., Xuan, J.: Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases. Journal of Machine Learning Research, pp. 2141–2167 (2010)
13. Do, K.: Applications of gene shaving and mixture models to cluster microarray gene expression data. Cancer Informatics, 2, pp. 25–43 (2007)
14. Hernandez-Montiel, L. A.: Hybrid Algorithm Applied on Gene Selection and Classification from Different Diseases. IEEE Latin America Transactions, pp. 930–935 (2016)
15. Xiang, J.: A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method. Applied Soft Computing, pp. 293–307 (2015)
16. Huang, Q., Tao, D., Li, X., Liew, W. C.: Parallelized Evolutionary Learning for Detection of Biclusters in Gene Expression Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2012)
17. Gordon, G. J.: Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. Cancer Res. (2002)
18. Pomeroy, S. L., Tamayo, P.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, pp. 436–442 (2002)
19. Alizadeh, A. A., Eisen, B. M., Davis, R. E.: Distinct Types of Diffuse Large (B)–Cell Lymphoma Identified by Gene Expression Profiling. Nature, pp. 503–511 (2000)
20. Martínez, W. L., Martinez, A. R.: Exploratory Data Analysis with MATLAB®. A CRC Press Company. Boca Ratón London New York Washington, D.C (2005)
21. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. J. American Statistical Association, pp. 77–87 (2002)
22. Mishra, D., Sahu, B.: Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach. International Journal of Scientific & Engineering Research (2011)
23. Zaffalon, M., Hutter, M.: Robust Feature Selection by Mutual Information Distributions. 18th International Conference on Uncertainty in Artificial Intelligence, pp. 577–584 (2002)
24. Rashedi, E., Nezamabadi-pour, H., Saryazdi, S.: GSA: A Gravitational Search Algorithm. Information Sciences, Vol. 179, No. 13, pp. 2232–2248 (2009)
25. Sajedi, H., Razavi, S. F.: DGSA: discrete gravitational search algorithm for solving knapsack problem. Oper Res Int J, pp. 1–29 (2016)
26. Sugunal, N., Thanushkodi, K.: An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. IJCSI International Journal of Computer Science Issues, pp. 18–21 (2010)
27. Wang, X., Gotoh, O.: Cancer classification using single genes. G. Inf., pp. 176–188 (2009)
28. Zhang, H., Song, X., Wang, H., Zhang, X.: Miclique: An Algorithm to Identify Differentially Co-expressed Disease Gene Subset from Microarray Data. Journal of Biomedicine and Biotechnology (2009)
29. Li, S., Wu, X., Hu, X.: Gene selection using genetic algorithm and support vectors machines. Soft Comput, pp. 693–698 (2008)
30. Aguilar-Ruiz, J. S., Azuaje, F., Riquelme, J. C.: Data Mining Approaches to Diffuse Large B–Cell Lymphoma Gene Expression Data Interpretation. DaWaK, Springer-Verlag Berlin Heidelberg, LNCS, 3181, pp. 279–288 (2004)
31. Hernández-Hernández, J. C., Duval, B. J., Hao, K.: 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. Filippone, M., Masulli, F., Rovetta, S.: Simulated Annealing for Supervised Gene Selection. Soft Computing, pp. 1471–1482 (2011)

33. Gunavathi, C., Premalatha, K.: Cuckoo search optimization for feature selection in cancer classification: a new approach. Int. J. Data Mining and Bioinformatics, pp. 248–265 (2015)
34. Bonilla-Huerta, E.: 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)
35. Li, S., Wu, X., Tan, M.: Gene Selection using Hybrid Particle Swarm Optimization and Genetic Algorithm. Soft Comput., pp. 1039–1048 (2008)
36. Yu, G.: Matched Gene Selection and Committee Classifier for Molecular Classification of Heterogeneous Diseases. Journal of Machine Learning Research, pp. 2141–2167 (2010)
37. Pang, S., Havukkala, L., Hu, Y., Kasabov, N.: Classification consistency analysis for bootstrapping gene selection. In Neural Computing and Applications, 16, pp. 527–539 (2007)