ISSN 2007-9737

Exploración de microarreglos de ADN utilizando minería de datos y una búsqueda tabú

Luis Alberto Hernández Montiel[1], José-Antonio León-Borges[2], Luis David Huerta Hernández[1]

[1] Universidad del Istmo campus Ixtepec, Ciudad Ixtepec, Oaxaca, México

[2] Universidad de Quintana Roo unidad Cancún, Cancún, Quintana Roo, México

{luisahm@itamail.itapizaco.edu.mx, jleon}@uqroo.edu.mx, luisdh2@bianni.unistmo.edu.mx

Resumen. En este artículo, se presenta un método híbrido basado en técnicas de minería de datos y una búsqueda tabú aplicado en la selección y clasificación de genes de Microarreglos de ADN. El método está dividido en dos etapas, en la primera etapa se elimina toda la información no relevante de la base de datos utilizando cinco técnicas de filtrado de datos. Con los subconjuntos de genes obtenidos por esta etapa, se realiza una nueva etapa de selección de genes utilizando una búsqueda tabú, para el proceso de clasificación de los genes seleccionados, se utilizan los clasificadores SVM, LDA, KNN por separado. El método se ha implementado para obtener un subconjunto pequeño de genes de alto desempeño, los resultados obtenidos se comparan con otros métodos reportados en la literatura, este método se aplica en tres bases de datos de dominio público.

Palabras Clave. Microarreglos de ADN**,** normalización, filtrado de datos, selección, clasificación, búsqueda local.

Exploration of DNA Microarrays Using Data Mining and a Taboo Search

Abtract. In this article, we present a hybrid method based on data mining techniques and a taboo search applied in the selection and classification of DNA Microarray genes. The method is divided into two stages, in the first stage all non-relevant information in the database is eliminated using five data filtering techniques. With the subsets of genes obtained by this step, a new stage of gene selection is carried out using a taboo search, for the classification process of the selected genes, the SVM, LDA, KNN classifiers are used separately. The method has been implemented to obtain a small subset of high performance genes, the results obtained are compared with other methods reported in the literature, this method is applied in three databases of public domain.

Keywords. DNA microarrays, normalization, data filtering, selection, classification, local search.

1. Introducción

Definir las causas por las cuales se genera una enfermedad se ha vuelto una tarea que involucrar diferentes áreas del conocimiento como la medicina, la biología, las matemáticas aplicadas y la informática, generando nuevos estudios basados en grandes cúmulos de información médica que ayudan al descubrimiento de causas relevantes para el comportamiento de una enfermedad. Uno de los métodos más utilizados son los datos de expresión genética obtenidos de la tecnología de microarreglos de ADN [1, 2], ésta tecnología ayuda a comprender la dinámica celular y sus relaciones con estados patológicos [1]. Sin embargo, los datos de expresión genética son difíciles de estudiar, ya que tienen como característica principal una alta dimensión debido a que el número de genes existentes es considerablemente mayor (usualmente miles), en comparación con la cantidad de muestras analizadas (usualmente menos de 100) [3]. En este documento, para abordar la problemática de obtener información relevante de los microarreglos de ADN, se propone un método híbrido de selección y clasificación para realizar una reducción de la dimensión de las bases de datos.

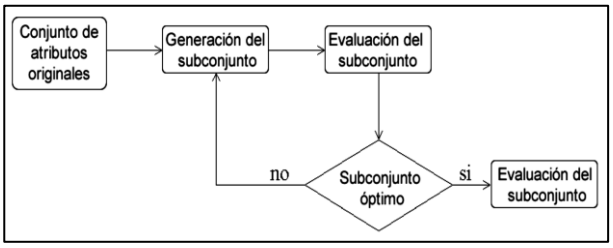

Fig. 1. Proceso general de selección y clasificación de características

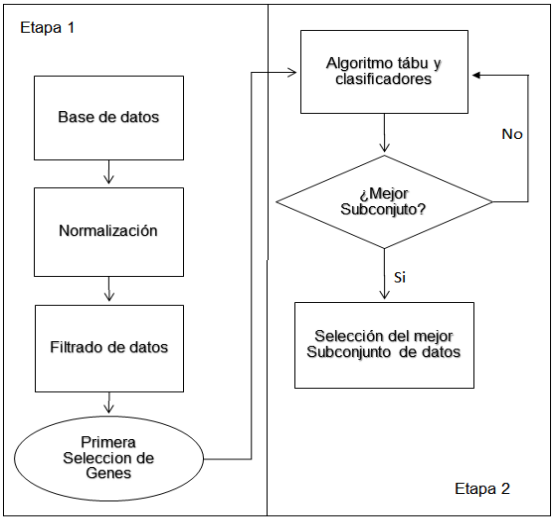

Fig. 2. Proceso de selección de características utilizando el algoritmo híbrido

Primero se realiza una limpieza del microarreglo utilizando un pre-procesamiento de los datos, eliminando los genes ruidosos y generando una primera reducción de la base genómica.

Después se genera una selección de un subconjunto de genes utilizando una heurística de búsqueda basada en la búsqueda tabú, para evaluar la calidad del subconjunto seleccionado, se utilizan tres métodos de clasificación, con la combinación de estas técnicas, se buscan los genes con información relevante dentro de tres bases de datos obtenidas de microarreglos de ADN de dominio público.

2. Selección y clasificación de datos genómicos utilizando minería de datos

La tecnología de microarreglos de ADN, se utilizan para la adquisición y almacenamiento de datos obtenidos directamente del genoma humano [2]. Permite manipular grandes cantidades de información genética, sin embargo, no toda la información contenida es útil, una base de datos obtenida de la tecnología de microarreglos de ADN, tiene un gran número de características que requieren un largo tiempo de procesamiento para ser analizada. Además, las bases de datos tienen dimensiones altas, miles de los genes son redundantes o contienen ruido [3].

Existen diferentes técnicas basada en minería de datos (como selección y extracción de características), y de aprendizaje máquina [4, 5], que ayudan a obtener información relevante a través de la exploración de los microarreglos de ADN. La selección de características (SC), ayuda a explorar datos de expresión genética que normalmente contienen un número grande de genes, pero un número pequeño de muestras.

La SC se puede ver como el proceso de encontrar un conjunto de genes que determinen

mejor las diferencias existentes en una muestra biológica [4]. Además, tiene los objetivos esenciales para reducir el ruido y redundancia de los datos, sirve para mejorar la exactitud de clasificación de una muestra, y los resultados ayudan a biólogos a que se enfoquen en los genes seleccionados para mejorar sus pruebas y validar sus hipótesis biológicas [6]. Para generar una selección de características efectiva se puede ocupar métodos basados en estadística y en aprendizaje máquina [4], los métodos estadísticos seleccionan características basándose en un criterio de discriminación que son relativamente independientes de su clasificación. Los métodos basados en aprendizaje máquina realizan la selección de variables usando como criterio de evaluación la estimación del error basadas en algún clasificador como las redes neuronales o clasificadores bayesianos [7, 8]. La figura 1 muestra un proceso de selección de características, éstas se evalúan en caso que sean características relevantes y se apartan del conjunto original, sino, se vuelve a generar un nuevo subconjunto para ser evaluado, así este proceso se repite un número de veces.

3. Algoritmo híbrido para la exploración de microarreglos de ADN

En este documento se aborda el problema de selección y clasificación efectiva de genes de microarreglos de ADN, utilizando un método híbrido combinando técnicas de selección de características basadas en filtrado de datos como primera etapa de selección. En la segunda etapa de selección y clasificación se utiliza un método híbrido basado en una búsqueda tabú combinada con tres diferentes clasificadores, la figura 2 muestra el método general de selección y clasificación de genes.

3.1. Estandarización de los datos

Las bases de datos obtenidas de la tecnología de microarreglos de ADN no son homogéneas, es decir, la forma en que se encuentra la información original tiene diferentes escalas numéricas y sigue diferentes distribuciones estadísticas.

La estandarización de datos se utiliza para transformar los datos con diferentes distribuciones a una escala igual para todos los datos. En este experimento, los datos son estandarizados utilizando una normalización Min-Max, la cual está definida por [9]:

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

donde X es la base de datos original. $Min(X)$ y $Max(X)$ son el dato mínimo y el dato máximo existentes en las bases de datos, X' es la nueva base de datos normalizada.

3.2. Filtrado de datos

Dentro de los microarreglos, existe información poco confiable, que puede sobre entrenar al clasificador y dar resultados erróneos, una de las formas de eliminar esta información es utilizando métodos de limpieza de ruido, al eliminar la información errónea o ruidosa se obtienen muestras bien etiquetadas para clasificar nuevos patrones dentro del microarreglo [10]. En este trabajo se utiliza un pre-procesamiento para eliminar el ruido del microarreglo de la siguiente forma: se genera una primera selección de información para cada base de datos, utilizando una puntuación generada por cinco métodos de filtrado estadístico independientes, esta puntuación sirve como indicador discriminatorio entre los genes para saber cuál de ellos contiene información más relevante [12].

Los cinco filtros que se utilizan en este estudio son: sumas de cuadrados entre los grupos y dentro de los grupos (BSS/WSS), información mutua, relación señal a ruido, prueba de wilcoxon y T-statistic. Estos filtros se utilizan por sus capacidades estadísticas, cada filtro prioriza un gen en particular y los demás filtros priorizan otros genes, se desea hacer un consenso de los genes mejores filtrados con cada uno de los métodos utilizados y con éste consenso trabajar por separado dentro del algoritmo propuesto.

3.2.1. BSS/WSS (BW)

La selección de genes se basa en la razón de las sumas de cuadrados entre los grupos (BSS) y

	Algoritmo 1.
1	**Genera** Solución inicial S;
2	**Evalúa** Solución inicial f(S);
3	**Genera** lista tabú LT;
4	Número de Iteraciones (NI)
5	**Mientras** (NI)
6	**genera** Vecindario N(S);
7	**Evalúa** Vecindario f(N(S))
8	**Actualiza** lista tabú LT=S
9	**Selecciona** mejor solución vecina (S' € N(s))
10	**Si** (S' > S y S' ≠ LT)
11	S=S'
12	**Fin.**
13	**Fin**

dentro de los grupos (WSS). Para el (gen) j, la razón está dada por [11]:

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

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 de gen *j* en todas las muestras para la pertenencia de la clase *k*.

3.2.2. Información mutua (MI)

Sean A y B dos genes aleatorios 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 [12]:

$$I(A;B) = \sum_{a_i} \sum_{b_j} P(a_i,b_j) \log \frac{P(a_i,b_j)}{P(a_i)P(b_j)}, \quad (3)$$

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.

3.2.3. Relación señal a ruido (SN)

Identifica los patrones de expresión genética con una diferencia máxima en la expresión media entre dos clases y la variación mínima de expresión dentro de cada clase. En este método, los genes son los primeros clasificados de acuerdo a sus niveles de expresión [13]:

$$SNR = |(\mu_1 - \mu_2)/(\sigma_1 + \sigma_2)|, \quad (4)$$

donde μ_1 y μ_2 denotan los valores medios de expresión de la clase 1 y clase 2, respectivamente, σ_1y σ_2 son las desviaciones estándar de las muestras en cada clase.

3.2.4. Prueba de Wilcoxon (WT)

Para cada gen *j*, se necesita el supuesto que las observaciones $x_{ij},\ldots,xn_j$ sean independientes. Si rank(x_{ij}) denota el rango de x_{ij} en la sucesión $x_{ij},\ldots,xn_j$, la prueba estadística para el gen *j* está dada por [14]:

$$W_j = \sum_{i:Y_i=1} rank(x_{ij}). \quad (5)$$

Para probar la hipótesis se utiliza

H0: mediana$\left(X_j \middle| Y = 1\right)$=mediana$\left(X_j \middle| Y = 2\right)$ vs

H1: mediana$\left(X_j \middle| Y = 1\right)$≠mediana$\left(X_j \middle| Y = 2\right)$

Bajo H0, W_j tiene una distribución de Wilcoxon con grados de libertad n_1 y n_2. El valor descriptivo de la prueba (p-value) correspondiente para cada gen j puede ser usado como una medida de relevancia.

3.2.5. T-statistic (TT)

Cada muestra se etiqueta con {1, -1}. Para cada gen f_j 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 T(f_j) pueden ser obtenidas por [15]:

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

donde n_1 (n_{-1}), es el número de ejemplos etiquetados con 1 (-1). Son consideradas como los genes más discriminatorios aquellas con la puntuación alta.

3.3. Subselección y clasificación utilizando una búsqueda tabú

Después de generar una primera selección de genes con los métodos de filtro (etapa 1), se hará una nueva selección utilizando una búsqueda tabú combinada con diferentes clasificadores, el método será entrenado con los datos obtenidos por los filtros y se describe a continuación.

3.3.1. Búsqueda tabú

La búsqueda tabú (BT), es una técnica que utiliza una memoria, con el objetivo de guiar un procedimiento de búsqueda local para resolver problemas de optimización combinatoria con un alto grado de dificultad, explorando el espacio de soluciones más allá del óptimo local [16]. Se puede obtener un algoritmo BT básico mediante la utilización de una lista tabú. En cada iteración, la solución actual ***S*** es reemplazada por la mejor solución vecina ***S'*** que no esté prohibida por la lista tabú. *S' Є N(s)* de tal manera *∀ s'' Є N(s), f(S'') ≤ f(s')* y *S'* donde es el conjunto de soluciones prohibidas por la lista tabú.

Note que el vecino seleccionado ***S'*** puede o no puede ser mejor que ***S***. El algoritmo BT se detiene cuando un número fijo de iteraciones se alcanza o cuando todos los movimientos se han convertido en tabú. La principal función de la lista tabú es prevenir que se cicle la búsqueda [17]. EL código simple de una búsqueda tabú se describe a continuación.

El algoritmo muestra un proceso general de una búsqueda tabú, actualizando la lista tabú y reemplazando la solución anterior por la mejor solución encontrada dentro del vecindario.

3.3.2. Máquina de vectores de soporte

El clasificador máquina de vectores de soporte (SVM por *Support Vector Machine*), es una técnica que opera de acuerdo a un paradigma de aprendizaje supervisado, aprendiendo de una relación funcional entre los atributos (o características), de entrada y salida por medio de apariciones de ejemplos etiquetados, se utiliza para analizar datos y reconocer patrones, para metodologías estadísticas y análisis de regresión, el algoritmo de entrenamiento SVM, construye un modelo que predice si un nuevo ejemplo sigue dentro de una categoría o de otra [13].

Discriminan datos de clases linealmente separables, dibujando un hiperplano óptimo en el espacio del vector de características, de tal manera que maximice el margen de separación entre los ejemplos positivos y negativos [18]. Los clasificadores SVM funcionan de la siguiente forma. Dado un conjunto de muestras m etiquetados S= {(xi, yi) | (xi, yi) ϵ Rn x {±1}, i=1, 2 ,..., m}, donde xi ϵ Rn, yi ϵ {+ 1} es una etiqueta de la muestra de xi, el hiperplano se define por [19]:

$$f(x) = \sum_{i=1}^{m} a_i y_i K(x_i, x) + b, \tag{7}$$

donde $K(x_i, x)$, es la función del núcleo 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.

3.3.3. K-vecino más cercano

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 [14], 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 [20]:

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

Fig. 3. Matriz de datos de expresión genética

Tabla 1. Descripción de los Microarreglos de ADN

Bases de datos	Genes	Muestras	Clases
Leucemia [22]	7129	72	2
Cáncer de Colon [23]	2000	62	2
Cáncer de Pulmón [24]	12533	181	2

$$\hat{f}(x) = \frac{k/n}{V_k(x)}, \tag{8}$$

donde $V_k(x)$, 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.

3.3.4. Análisis lineal discriminante (LDA)

El clasificador basado en Análisis Lineal Discriminante (LDA por *Lineal Discriminant Analysis*), es una técnica de aprendizaje supervisado para clasificar datos. La idea central de LDA es obtener una proyección de los datos en un espacio de menor (o incluso igual), dimensión que los datos entrantes, con el fin de que la separabilidad de las clases sea la mayor posible [21]. LDA se acerca al problema de clasificación mediante la búsqueda de una matriz de transformación que ayude a preservar la mayor parte de la información que se utilice para discriminar entre diferentes clases.

Lo anterior se logra mediante la reestructuración de los datos de alta dimensión proyectándolos en un espacio de pocas dimensiones. Para alcanzar la matriz de transformación óptima, dos matrices S_B (dispersión entre las clases) y S_W (dispersión dentro de las clases) deben ser calculadas de acuerdo a las siguientes ecuaciones [21]:

$$S_w = \sum_k \sum_{x_i \in c_k} (x_i - \mu_k)(x_i - \mu_k)^t, \tag{9}$$

$$S_B = \sum_k n_k (\mu_k - \mu)(\mu_k - \mu)^t, \tag{10}$$

donde n_k es el número de ejemplos de entrenamiento para la clase k, c_k es el conjunto de índices de los ejemplos de entrenamiento pertenecientes a la clase k, x_i es el valor de expresión genética del gen i, μ_k es la media de la clase k y μ es la media resultante de las dos clases.

Entonces LDA está preparada para clasificar nuevas muestras después de que encontró un valor óptimo para el vector w tal que $w^t S_B w$ es maximizada mientras $w^t S_w w$ es minimizada como se muestra en la siguiente ecuación [21]:

$$F(w) = \frac{w^t S_B w}{w^t S_w w}. \tag{11}$$

3.3.5. Procedimiento general

En nuestro caso el algoritmo es implementado de la siguiente manera:

- La búsqueda tabú se implementa de manera binaria, donde la solución inicial ***S*** se genera de forma aleatoria con una distribución uniforme.
- En la función de costo de la búsqueda, se utiliza uno de los tres clasificadores descritos anteriormente, esto servirá para evaluar (***f(s)***) la calidad de los genes seleccionados por la solución inicial, el resultado obtenido por el clasificador es validado utilizando el método *10-fold cross-validation*.
- Se genera el vecindario a partir de la solución inicial ***S***, el cual se evalúa con la función de costo. Se busca dentro del vecindario la mejor solución ***S'*** y se verifica si se encuentra prohibida por la lista tabú, si ***S'*** es tabú, se toma la segunda mejor solución del vecindario, si ***S'*** no es tabú, ***S'*** se toma como la mejor solución del vecindario.

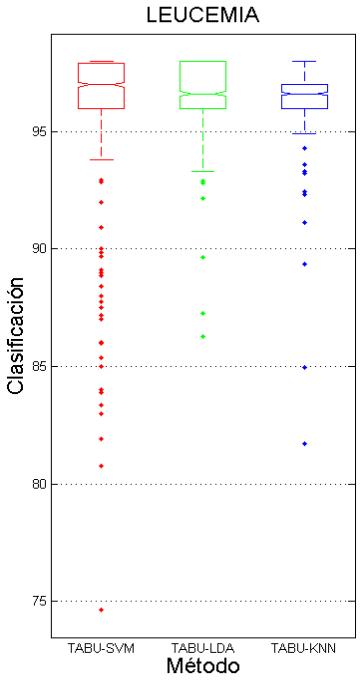

Fig. 4. Tasa de clasificación obtenida por el algoritmo híbrido para la base de datos de leucemia

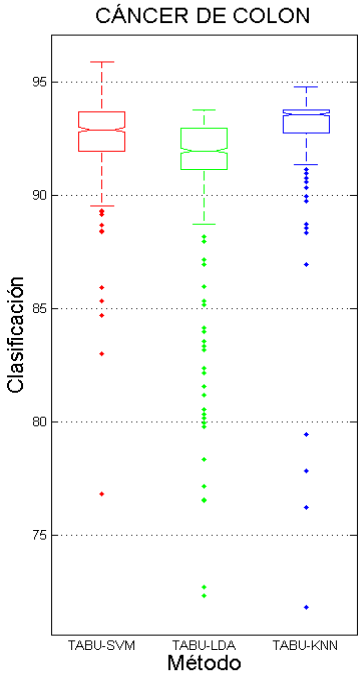

Fig. 5. Tasa de clasificación obtenida por el algoritmo híbrido para la base de datos de cáncer de colon

- La solución ***S'*** reemplaza a la solución inicial (***S=S'***) y se genera un nuevo vecindario a partir de la solución ***S***,
- Este proceso se repite *n* veces hasta que ***S'*** hasta que se cumplan un número de iteraciones.
- La lista tabú es implementada de la siguiente manera.
- Cada vez que un movimiento mv(i,j) se lleva a cabo, un gen es descartado y un gen es seleccionado, el gen seleccionado es guardado en la lista tabú por las siguientes ***k*** iteraciones.

Por consecuencia, este gen no se puede volver a seleccionar durante el proceso.

- El valor de ***k*** es el tiempo de permanencia que el gen estará dentro de la lista tabú y varía desde k_{min} a k_{max}.
- La lista tabú prohíbe un nuevo gen seleccionado, este gen se puede retirar de la lista tabú (criterio de aspiración), en la siguiente iteración sí el coeficiente de clasificación del nuevo gen seleccionado es muy bajo.

Tabla 2. Parámetros utilizados por la búsqueda tabú

Parámetros	
Solución inicial	300
Tamaño de la lista tabú	7
Criterios de aspiración	si
Número de iteraciones	1000

4. Experimentos y resultados

El método propuesto se entrena con diferentes conjuntos de datos genómicos, se logra observar que el método es capaz de seleccionar genes con información relevante.

En esta sección se muestran los resultados obtenidos por el método propuesto y se genera un estudio de comparación de los resultados obtenidos con diferentes literaturas.

Tabla 3. Tasas de clasificación obtenidas por el algoritmo híbrido

AUTORES	Leucemia	Colon	Pulmón
	%(Ng)	%(Ng)	%(Ng)
Luo et al. [25]	71.39 (5)	80.07(7)	--
Yu et al. [26]	96.8(10)	88.6(10)	94.7(10)
Cho et al. [29]	95.9(25)	87.7(25)	--
Hernández et al. [17]	92.52(6)	87.00(8)	--
Filippone et al. [27]	94.7(13)	80.6(21)	--
Li et al. [28]	95.1(21)	88.7(16)	--
Bonilla et al. [33]	**99.5**(**3**)	90.5(**3**)	96.0**(3)**
Tan et al. [34]	91.1	95.1	93.2
Yue et al.[35]	83.8(100)	85.4(100)	--
Pang et al. [31]	94.1(35)	83.8(23)	91.2(34)
Li et al. [32]	97.1(20)	83.5(20)	--
Zhang et al. [30]	**100**(30)	90.3(30)	**100**(30)
González [54]	***99.62****(3)*	*89.19 (5)*	***99.89****(7)*
Tabú-SVM	98.00(4)	**95.90**(3)	97.94(**3**)
Tabú-LDA	98.00(**2**)	93.77(**2**)	97.17(**3**)
Tabú-KNN	98.00(**2**)	94.77(3)	97.72(**3**)

4.1. Microarreglos de ADN

Los microarreglos de ADN, son una herramienta que permite realizar diversos análisis genéticos basados en la miniaturización de procesos biológicos, su funcionamiento se basa en la capacidad que tienen las moléculas complementarias de ADN de hibridar entre sí, utilizando pequeñas cantidades de ADN correspondientes a diversos genes cuya expresión se desea medir [1]. Los microarreglos de ADN tiene forma de una matriz de datos donde las filas representa los genes y las columnas representan las muestras.

Cada celda dentro de ésta matriz, es un valor de expresión genética que representa la intensidad del gen correspondiente a cada una de las muestras. Lo anterior se observar en la figura 3 donde x representa el dato genómico, n_g (número de gen, filas) los genes dentro de la matriz y n_m (número de muestras, columnas) las muestras dentro de la matriz.

En éste trabajo, se utilizan tres microarreglos de ADN descritos en la tabla 1.

4.2. Parámetros

El algoritmo híbrido ha sido implementado en Matlab (Versión 7.11.0). Los parámetros más confiables con los cuales fue entrenada la búsqueda tabú para las tres bases de datos se muestran en la tabla 2.

4.3. Resultados

En el protocolo experimental, los cinco métodos de filtrado de datos funcionan como una etapa de pre-selección generando una reducción significativa de las tres bases genómicas, descartando los genes ruidosos y genes redundantes y obteniendo como resultado los nuevos subconjuntos con información relevante,

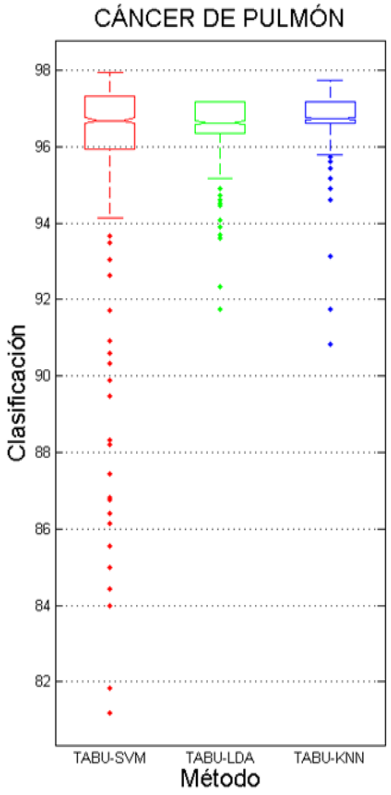

Fig. 6. Tasa de clasificación obtenida por el algoritmo híbrido para la base de datos de cáncer de pulmón

que se utilizan en la siguiente etapa para entrenar la búsqueda tabú y los clasificadores. En esta nueva etapa se genera una selección de genes dentro los nuevos subconjuntos, eliminando los genes menos informativos y seleccionando los genes que logran entrenar mejor al clasificador. Así se obtiene un subconjunto mínimo de genes con una tasa de desempeño alta.

La figura 4 (leucemia), 5 (cáncer de colon) y 6 (cáncer de pulmón), muestran la comparación de las tasas de clasificación que se han obtenido al entrenar el algoritmo tabú combinado con cada uno de los clasificadores (SVM, LDA, KNN).

La tabla 3, muestra la mejor tasa de clasificación obtenida por el método propuesto y la comparación con diferentes métodos reportados en la literatura. La tabla se divide de la siguiente forma: en la primera 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 (Ng), que fueron obtenidos por los métodos propuestos para las tres bases de datos. Los resultados obtenidos se han comparado con los autores mostrados en la tabla 3. Cabe mencionar que algunos autores presentan un modelo basado en algún tipo de metaheurística como un algoritmo genético o se basan en búsquedas locales y otros clasifican basándose en técnicas de aprendizaje máquina.

En tabla 3 al comparar las tasas de clasificación obtenidas por nuestro método para las tres bases de datos, se aprecia que el método es muy competitivo en relación con algunos de los autores con los que se han comparado.

Por ejemplo, Zhang et al. [30] obtuvo el 100% de clasificación seleccionando un subconjunto de 30 genes para Leucemia, Bonilla-Huerta et al. [33], obtuvo 90.5% de clasificación seleccionando un subconjunto de 3 genes para Cáncer de Colon, Zhang et al. [30], obtuvo el 100% seleccionando un subconjunto de 30 genes para Cáncer de Pulmón.

Los mejores resultados obtenidos para las bases de datos son:

Para Leucemia la tasa de clasificación más alta fue de 98% seleccionando un subconjunto de 2 genes, en Cáncer de Colon la tasa de clasificación más alta fue de 95.90% seleccionando un subconjunto de 3 genes y Cáncer de Pulmón la tasa de clasificación más alta fue 97.94% seleccionando un subconjunto de 3 genes.

En cada una de las tres bases de datos existen genes informativos que entrenan al clasificador de manera eficiente, una forma de verificar si los genes seleccionados pueden ayudar en el diagnóstico de una enfermedad es revisando en la literatura si estos genes han sido reportados por algún autor, de modo que podemos encontrar una interpretación biológica de los genes seleccionados.

La base de datos de leucemia y de cáncer de colon, han sido estudiadas ampliamente, permite encontrar la mayoría de genes relevantes reportados por diferentes autores. Por otra parte, la base de datos de cáncer de pulmón no ha sido estudiado ampliamente, consecuentemente surgen dudas para la comparación de los resultados con los genes reportados. A continuación, se muestran los genes

seleccionados y reportados por diferentes autores obteniendo así una interpretación biológica más confiable:

En nuestro trabajo encontramos el gen 4847 (Zyxin) como el más relevante dentro de la base de datos de leucemia, este gen fue seleccionado por los tres métodos utilizados Tabú-(SVM, LDA, KNN). El gen ha sido reportado por [15, 22, 36, [37], [38], indicando que tiene un rol importante dentro de la clasificación de leucemia, debido a su nivel de expresión logra identificar dos tipos de leucemia aguda y así ser clasificado o etiquetado en la clase Leucemia Mieloide Aguda o Leucemia Linfoblastica Aguda. Otros genes encontrados para leucemia son 1882 (CST3 Cystatin C amyloid angiopathy and cerebral hemorrhage) 2020 (FAH Fumarylacetoacetate) y el gen 760 (CYSTATIN A). Los dos primeros fueron encontrados por dos de los tres métodos propuestos Tabú-(SVM, KNN) y Tabú-(SVM, LDA) respectivamente y el último seleccionado por el método (Tabú-SVM). Los tres genes han sido reportados en [15, 29, 37, 38, 39, 40, 41]. Para la base de datos de cáncer de colon cada uno de los métodos ha logrado identificar de dos a tres genes relevantes que ayudan a la clasificación de muestras con tejidos de tumores y muestras de tejidos normales.

Los genes más relevantes que han encontrado son: el gen 245 (Human cysteine-rich protein (CRP) gene, exons 5 and 6) con el método Tabú-(LDA, KNN) reportado en [42, 43, 44] y el gen 765 (Human cysteine-rich protein CRP gene, exons 5 and 6) con el método Tabú-(SVM, KNN) reportado en [37, 42, 45, 46, 47]. Estos dos genes han logrado separar mejor la clase de tejidos de tumores de la clase de tejidos normales, y se pueden utilizar en la identificación células con cáncer de colon. El resto de genes seleccionados para cada método son: para el método Tabú-SVM el gen 249 (Human desmin gene, complete cds), y el gen 897 (3' UTR 2ª 183264 Complement Factor D Precursor (Homo sapiens)). El método tabú-KNN ha seleccionado el gen 267 (Human cysteine-rich protein CRP gene, exons 5 and 6). El método Tabú-LDA también ha seleccionado el gen 493 (Myosin Heavy Chain, Nonmuscle Gallus gallus), estos genes han sido reportados en [33, 37, 46, 48, 49, 50, 51, 52].

En cáncer de pulmón, un total de tres genes relevantes han sido seleccionados por los tres métodos propuestos Tabú-(SVM, LDA, KNN), los genes encontrados son: el gen 3844 (Interferon, alpha-inducible protein clone IFI-6-16), el gen 8537 (Replication protein A1, 70kDa) y el gen 11841 (leucine-rich PPR-motif containing). Estos genes han logrado entrenar mejor los clasificadores, a diferencia de los demás genes utilizados en el estudio. La selección de estos genes se debe a la separación por el clasificador de 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). Estos genes han sido reportados en [33, 36, 53].

5. Conclusiones

En este trabajo, se presentó un método híbrido basado en una búsqueda local y técnicas de minería de datos, implementado en la selección y clasificación de un conjunto de genes importantes explorando dentro de tres bases de datos de dominio público (Leucemia, Cáncer de pulmón, y Cáncer de Colon). El método propuesto tiene una etapa de pre-selección de genes mediante la utilización de cinco técnicas de filtrado de datos, estos filtros utilizan una puntuación o categoría que sirve para discriminar los genes contenidos en la base de datos, así se eliminan los genes no relevantes (ruidosos o redundantes) y son seleccionados los genes con información pertinente.

Con lo anterior se ha generado una primera reducción efectiva de la dimensión de las bases de datos. Para realizar la selección dentro de los subconjuntos obtenidos por las técnicas de filtrado, se ha creado un algoritmo híbrido basado en una búsqueda tabú como método de selección de genes combinada con tres técnicas de clasificación (SVM, LDA, KNN). Utilizando las propiedades de memoria de la búsqueda tabú, se ha logrado crear un algoritmo guiado que recuerda los genes que han sido utilizados en un proceso (iteración) anterior. Basándose en la tasa de clasificación del gen recordado permite que el algoritmo prohíba genes que han sido utilizados durante su ejecución y trabaje con nuevos genes logrando explorar a profundidad la base de datos

y consecuentemente obtener la mejor tasa de clasificación.

Cada técnica utilizada en éste trabajo ha seleccionado un subconjunto de genes con una tasa de clasificación alta. Para saber la relevancia que tiene cada gen seleccionado, se utiliza la frecuencia de selección del gen por cada método propuesto, de esta manera, se logra observar que un gen en particular al ser seleccionado logra entrenar el clasificador obteniendo una tasa de clasificación aceptable.

El método propuesto determina una tasa de clasificación alta, obtenida con un subconjunto de genes pequeño para las tres bases de datos. Para evaluar la eficiencia del método, se genera 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 de los métodos reportados. Además de las tasas de clasificación, se desea conocer si los genes han sido 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 Tabú con uno de los tres clasificadores (SVM, KNN, LDA) dentro del proceso de minería de datos. En trabajos futuros, se pretende probar y compara otros algoritmos de selección de características, también utilizar otros métodos de clasificación, la meta es minimizar el número de genes a utilizar y maximizar la tasa clasificación.

Referencias

1. **Guyon, I. & Elisseeff, A. (2003).** An Introduction to Variable and Feature Selection. *J. of Machine Learning Research*, pp. 1157–1182.
2. **Moreno, V. & Solé, X. (2000).** *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, Vol. 122, No. 1, pp. 73–79.
3. **Huang, Q., Tao, D., Li, X., & Liew, A. (2012).** Parallelized Evolutionary Learning for Detection of Biclusters in Gene Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, No. 2, pp. 560–570. DOI: 10.1109/TCBB.2011.53.
4. **Zhang, Y., Ding, C., & Li, T. (2008).** Gene Selection Algorithm by Combining ReliefF and mRMR. *IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School Boston,* Vol. 9, No. 2, pp. 1–10. DOI: 10.1186/1471-2164-9-S2-S27.
5. **Altamiranda, J., Aguilar, J., & Hernández, L. (2015).** Pattern Recognition System Based on Data Mining for Analysis of Chemical Substances in Brain. *Computación y Sistemas,* Vol. 19, No. 1, pp. 89–107. DOI: 10.13053/CyS-19-1-1409.
6. **Yang, P., Zhou, B.B., Zhang, Z., & Zomaya, A. (2010).** A Multi-Filter Enhanced Genetic Ensemble System for Gene Selection and Sample Classification of Microarray Data. *The Eighth Asia Pacific Bioinformatics Conference Bangalore,* Vol. 11, No. 1, pp. 18–21. DOI: 10.1186/1471-2105-11-S1-S5.
7. **Ladha, L. & Deepa, T. (2011).** Feature Selection Methods and Algorithms. *International Journal on Computer Science and Engineering (IJCSE),* pp. 1787–1797.
8. **Sepúlveda, R., Montiel, O., Díaz, G., Gutiérrez, D., & Castillo, O. (2015).** Classification of Encephalographic Signals using Artificial Neural Networks. *Computation y Sistemas,* Vol. 19, No. 1, pp. 69–88. DOI: 10.13053/CyS-19-1-1570.
9. **Martínez, W.L., Martínez, A.R., & Solka, J.L. (2005).** Exploratory Data Analysis with MATLAB®. *A CRC Press Company.*
10. **Pascual-González, D., Vázquez-Mesa, F., & Toro-Pozo, J.L. (2014).** Noise Detection and Learning Based on Current Information. *Computación y Sistemas,* Vol. 18, No. 1, pp. 153–167. DOI: 10.13053/CyS-18-1-2014-025.
11. **Dudoit, S., Fridlyand, J., & Speed, T.P. (2002)**. Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. *Journal of the American Statistical Association*, Vol. 97, No. 457, pp. 77–87. DOI: 10.1198/016214502753479248.
12. **Zaffalon, M. & Hutter, M. (2002).** Robust Feature Selection by Mutual Information Distributions. *18th International Conference on Uncertainty in Artificial Intelligence*, pp.577–584.
13. **Mishra, D. & Sahu, B. (2011).** Feature Selection for Cancer Classification: A Signal-to-noise Ratio

Approach. *International Journal of Scientific & Engineering Research,* Vol. 2, No. 4, pp. 1–7**.**

14. **Porras-Cerrón, J.C. (2005).** Componentes Principales Supervisados Para Clasificación de Datos De Expresión Genética. *Tesis de Maestro en Ciencias, Universidad De Puerto Rico Mayagüez.*
15. **Tan, A.H. & Pan, H. (2005).** Predictive neural networks for gene expression data analysis. *Neural Networks,* Vol. 18, No. 3, pp. 297–306. DOI: 10.1016/j.neunet.2005.01.003.
16. **Glover, F. & Melián, B. (1989).** Tabu Search. *Revista Iberoamericana de Inteligencia Artificial.* Vol. 1, No. 3, pp. 190–206. DOI: 10.1287/ijoc.1.3.190.
17. **Hernández-Hernández, J.C., Duval, B.J., & Hao, J.K. (2008).** SVM-based local search for gene selection and classification of microarray data. *Comunicativos in Computer and Information Science*, Vol. 13, pp. 499–508. DOI: 10.1007/978-3-540-70600-7_39.
18. **Maji, P. & Paul, S. (2012).** Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification. *IEEE Transactions on Knowledge and Data Engineering*, pp. 225–252, DOI: 10.1007/978-3-319-05630-2_9.
19. **Wang, S., Chen, H., Li, R., & Zhang, D. (2006).** Gene Selection with Rough Sets for the Molecular Diagnosing of Tumor Base on Support Vector Machines. *International Computer Symposium*, pp. 1368–1373.
20. **Sugunal, N., & Thanushkodi, K. (2010).** An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. (*IJCSI) International Journal of Computer Science Issues*, Vol. 7, No. 4, pp. 18–21.
21. **Salem, D.A., Seoud, A., Ahmed, R., & Ali, H.A. (2011).** Mgs-cm: a multiple scoring gene selection technique for cancer classification using microarrays. *International Journal of Computer Applications*, Vol. 36, No. 6, pp. 30–37.
22. **Golub, T., Slonim, D., Tamayo, P.C., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., & Downi, J.R., (1999).** Molecular Classification of Cancer: Classes Discovery and Class Prediction by Gene Expression Monitoring. *Science*, Vol. 286, No. 5439, pp. 531–537. DOI: 10.1126/science.286.5439.531.
23. **Alon, U., Barkai, N., Notterman, D.A., Gish, K. Ybarra, S., Mack, D., & Levine, A.J. (1999).** Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Nat. Acad. Sci.*, pp. 6745–6750. DOI: 10.1073/pnas.96.12.6745.
24. **Gordon, G.J., Jensen, R.V., Hsiao, L.L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., & Bueno, R. (2002).** Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. *Cancer Res,* Vol. 62, No. 17, pp. 4963–4967.
25. **Luo, L.K., Huang, D.F., Ye, L.J., Zhou, Q.F., Shao, G.F., & Peng, H. (2011).** Improving the Computational Efficiency of Recursive Cluster Elimination for Gene Selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8, No. 1, pp. 122–129. DOI: 10.1109/TCBB.2010.44.
26. **Yu, L., Han, Y., & Berens, M.E. (2012).** Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, No. 1, pp. 262–272. DOI: 10.1109/TCBB.2011.47.
27. **Filippone, M., Masulli, F., & Rovetta, S. (2011).** Simulated Annealing for Supervised Gene Selection. *Soft Computing*, pp. 1471–1482.
28. **Li, S., Wu, X., & Tan, M. (2008).** Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Comput.,* pp. 1039–1048. *DOI: 10.1007/s00500-007-0272-x.*
29. **Cho, S.B. & Won, H.H. (2007).** Cancer classification using ensemble of neural networks with multiple significant gene subsets. *In Applied Intelligence,* Vol. 26, No. 3, pp. 243–250. DOI: 10.1007/s10489-006-0020-4.
30. **Zhang, L., Li, Z., & Chen, H. (2007).** An effective gene selection method based on relevance analysis and discernibility matrix. *PAKDD, (*LNCS), Vol. 4426, pp. 1088–1095. DOI: 10.1007/978-3-540-71701-0_123.
31. **Pang, S., Havukkala, I., & Hu, Y. (2007).** Classification consistency analysis for bootstrapping gene selection*. Neural Computing and Applications*, Vol. 16, No. 6, pp. 527–539. DOI: 10.1007/s00521-007-0110-1.
32. **Li, G.Z., Zeng, X.Q., Yang, J.Y., & Yang, Q. (2007).** Partial least squares based dimension reduction with gene selection for tumor classification. *Proceedings of IEEE 7th International Symposium on Bioinformatics and Bioengineering,* pp. 1439–1444. DOI: 10.1109/BIBE.2007.4375763.
33. **Bonilla-Huerta, E., Hernández-Montiel, A., Morales-Caporal, R., & Arjona-López, M. (2015).** 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, Vol. 13, No. 1, pp. 12–26. DOI: 10.1109/TCBB.2015.2474384.

34. **Tan, A.C. & Gilbert, D. (2003).** Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, Vol. 2, No. 3, pp. 1–9.
35. **Yue, F., Wang, K., & Zuo, W. (2007).** Informative gene selection and tumor classification by null space lda for microarray data. (*ESCAPE)*, Vol. 4614 of *Lecture Notes in Computer Science*, Springer, pp. 435–446. DOI: 10.1007/978-3-540-74450-4_39.
36. **Wang, X. & Gotoh, O. (2009).** Cancer classification using single genes. *Genome Informatics*, Vol. 23, pp. 176–188. DOI: 10.1142/9781848165632_0017.
37. **Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., & Yakhini, Z. (2000).** Tissue classification with gene expression profiles. *Journal of Computational Biology*, Vol. 7, No. 3–4. pp. 559–583. DOI: 10.1089/106652700750050943.
38. **Wessels, L.F., Reinders, M.J.T., Van-Welsem, T., Nederlof, P.M., & Wang**, **Y. (2002).** Representation and classification for high-throughput data. *SPIE*. Vol. 4626, pp. 226–237.
39. **Cho, S.B. & Won, H.H. (2003).** Machine learning in DNA microarray analysis for cancer classification. (*PCAPB)*, Vol. 19, pp. 189–198.
40. **Chu, W., Ghahramani, Z., Falciani, F., & Will, D.L. (2005).** Biomarker discovery in microarray gene expression with gaussian process. *Bioinformatics*, Vol. 21, No. 16, pp. 3385–3393. DOI:10.1093/ bioinformatics/bti526.
41. **Wang, S.L., Sun, L., & Fang, J. (2012).** Minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. *(BMC) Bioinformatics,* Vol. 13, No. 178, pp. 1–26. DOI: 10.1186/1471-2105-13-178.
42. **Krishnapuram, B., Carin, L., & Hartemink**, **A.J. (2004).** Joint Classifier and Feature Optimization for Comprehensive Cancer Diagnosis Using Gene Expression Data. *Journal of Computational Biology,* Vol. 11, No. 2–3. DOI: 10.1089/1066527041410463.
43. **Maglietta, R., D'Addabbo, A., Piepoli, A., Perri, B.F., Liuni, S., Pesole, G., & Ancona, N. (2007)** Selection of relevant genes in cancer diagnosis based on their prediction accuracy. *Artificial Intelligence in Medicine,* Vol. 40, No. 1, pp. 29–44. DOI: 10.1016/j.artmed.2006.06.002.
44. **Sundaram, A., Venkata, N.L., & Parthasarathy, R.S. (2013).** Hybrid SPR algorithm to select predictive genes for effectual cancer classification. *Turkish Journal of Electrical Engineering & Computer Sciences*, Vol. 21, No. 2.
45. **Roth, V. (2002).** The Generalized LASSO: A Wrapper Approach to Gene Selection for Microarray Data. *Proc. Conference on Automated Deduction,* pp. 252–255
46. **Huang, T. & Kecman, M.V. (2005).** Gene Extraction for Cancer Diagnosis by Support Vector Machines–an Improvement. *Artificial Intelligence in Medicine.* pp. 185–194.
47. **Chen, J.J., Tsai, C.A., Tzeng, S.L., & Chen, C.H. (2007).** Gene Selection with Multiple Ordering Criteria. (*BMC) Bioinformatics*, Vol. 8, No. 74, pp. 1–17, DOI: 10.1186/1471-2105-8-74.
48. **Zhang, H., Song, X., Wang, H., & Zhang, X. (2009).** Miclique: An Algorithm to Identify Differentially Co-Expressed Disease Gene Subset from Microarray Data. *Journal of Biomedicine and Biotechnology*, pp. 1–9. DOI: 10.1155/2009/642524**.**
49. **Li, L., Darden, T.A., Weingberg, C.R., Levine, A.J., & Pedersen, L.G. (2001).** Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/K-Nearest Neighbor Method. *Combinatorial Chemistry & High Throughput Screening,* Vol 4, No. 8, pp. 727–739. DOI: 10.2174/1386207013330733.
50. **Li, S., Wu, X., & Hu, X. (2008).** Gene selection using genetic algorithm and support vectors machines. *Soft Comput*, Vol. 12, No. 7, pp. 693–698. DOI: 10.1007/s00500-007-0251-2.
51. **Tan, F., Fu, X., Zhang, Y., & Bourgeois, A.G. (2006).** Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data. *IEEE Congress on Evolutionary Computation,* pp. 2529–2534. DOI: 10.1109/CEC.2006.1688623.
52. **Wang, X. & Gotoh**, **O. (2010).** Inference of cancer-specific gene regulatory networks using soft computing rules. *Gene Regulation and System Biology,* Vol. 4, pp. 19–34.
53. **Yoon, I.K., Kim, H.K., Kim, Y.K., Song, I.H., Kim, W., Kim, S., Baek, S.H., Kim, J.H., & Kim, J.R. (2004).** Exploration of replicative senescence-associated genes in human dermal fibroblasts by cDNA microarray technology. *Experimental gerontology,* Vol. 39, No. 9, pp. 1369–1378. DOI: 10.1016/j.exger.2004.07.002.
54. **González-Navarro, F.F. & Belanche-Muñoz, L.A. (2014).** Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy. *Computación y Sistemas,* Vol. 18, No. 2, pp. 275–293. DOI: 10.13053/CyS-18-2-2014-032.

Article received on 18/01/2018; accepted on 02/11/2017.
Corresponding author Luis Alberto Hernández Montiel.