Una Nueva Metodología para la identificación de Patrones de Biomarcación aplicados al Estudio, Prevención y Tratamiento Temprano de Enfermedades Crónicas

Herrera R.*; Herrera L.**

*Escuela Politécnica Nacional, Facultad de Ingeniería Eléctrica y Electrónica, Quito, Ecuador

e-mail: roberto.herrera.lara@gmail.com

**Escuela Politécnica Nacional, Facultad de Ingeniería Eléctrica y Electrónica, Quito, Ecuador

Resumen: Hoy en día, la aplicación de métodos matemáticos y computacionales han permitido revolucionar las investigaciones relacionadas con la medicina, quimiometria, proteómica y genomica. Algoritmos matemáticos de minería de datos y aprendizaje de maquina están abriendo las pruebas a nuevas esperanzas en la lucha contra las enfermedades crónicas como el cáncer, la diabetes, el alzheimer, cirrosis, enfermedades cardiovasculares y enfermedades suprarrenales. El presente trabajo presenta una nuevo método de selección y validación de patrones de biomarcadores aplicados al tratamiento temprano de enfermedades crónicas. La metodología propuesta en este trabajo aborda el análisis de datos procedentes directamente de los instrumentos de medición, hasta la definición de los patrones de biomarcación. Ésta basa su funcionamiento en la combinación de *t–test* y *Mann–Whitney U test* en un filtro estadístico, el cual define zonas de interés en los espectros de mediciones, luego estas zonas de interés son agrupadas y reducidas a las características cercanas a la media de cada una de estas, eliminando así la información redundante. La contribución de este trabajo radica en la estructura de este filtro estadístico, el cual posee una enorme capacidad de extracción de información a través de cálculos sencillos, en comparación a complejos algoritmos presentados en trabajos similares. Finalmente las características seleccionadas por este filtro son validadas usando clasificadores de *Adaboost*, *TotalBoost* y *LPBoost* probados con validación cruzada y pruebas con muestras externas. Los resultados obtenidos reflejan un rendimiento superior al 95%, además gran robustez en contra del sobreentrenamiento(*Overfitting*) e infraentrenamiento(*Underfitting*).

Palabras claves: biomarcador, aprendizaje de máquina, minería de datos, espectrometría de masas, filtro estadístico

Abstract: Today, the application of mathematical and computational methods have revolutionized research related to medicine, chemometrics, proteomics and genomics. Mathematical algorithms for data mining and machine learning are opening the doors to new hopes in the fight against chronic diseases like cancer, diabetes, Alzheimer's disease, cirrhosis, cardiovascular disease and adrenal diseases. This paper presents a new method for selection and validation of biomarker patterns applied to early treatment of chronic diseases. The proposed methodology deals with the analysis of data obtained directly from measuring instruments to defining patterns. It is based on the combination of *t - test* and *Mann - Whitney U test* in a statistical filter. These tests define areas of interest in the spectra of measurements, and then these areas of interest are grouped and reduced the closest feature to the average of each group of these features, thereby eliminating redundant information. The contribution of this work lies in the structure of the statistical filter, which has an enormous capacity for extracting information through simple calculations compared to complex algorithms presented in similar articles. Finally, the features selected by this methodology are validated using *Adaboost*, *TotalBoost* and *LPTBoost* classifiers tested using cross-validation and testing with external samples. The obtained results reflect an efficiency greater than 95%, furthermore robustness against overfitting and underfitting.

Keywords: biomarker, machine learning, data mining, mass spectrometry, statistical filter

1. INTRODUCCIÓN

La Espectrometría de Masas (EM) es una técnica de adquisición de datos muy popular en investigaciones sobre enfermedades crónicas[1] tales como el cáncer [1–9], enfermedades suprarrenales [10, 11], diabetes [12–14], enfermedades cardiovasculares [15], cirrosis [16, 17] y alzheimer [18]. Su popularidad está basada en la elevada capacidad que esta técnica posee para extraer información, además de la facilidad que presenta para ser integrada a metodologías computacionales de análisis de conjuntos masivos de datos [19].

Las metodologías computacionales usadas en estas aplicaciones médicas han recibido un enorme interés por parte de los investigadores ya que a través de ellas es posible realizar análisis y experimentación en forma mucho más exhaustiva que con los métodos tradicionales. La combinación de la estadística, probabilidad y simulación computacional permiten abordar una cantidad de casos de análisis mucho más variada que la experimentación tradicional de laboratorio.

Mediante esta técnica se realizan mediciones denominadas espectros de masa, dichos espectros, para las aplicaciónes analizadas en este trabajo, proceden del analisis de muestras de fluidos biológicos(FB) como saliva o suero sanguíneo. Estos FB poseen una cantidad enorme de información relativa a la presencia de patologías en el cuerpo humano [28]. Esta información está representada por biomarcadores (BM), los cuales son substancias usadas en el análisis y diagnóstico de patologías médicas. Estas substancias tienen la capacidad de indicar la presencia de un estado patológico, así como también la respuesta a tratamientos químicos.

Los espectros de masa adquiridos están constituidos por un vector de valores de intensidades, donde se expresa la abundancia de los iones en fase gaseosa de la muestra analizada y un vector de las relaciones masa a radio m/z expresado en thomsons $[th]$. En la Figura 1 se muestra varias mediciones de un conjunto de datos *OvarianCD_PostQAQC.zip*, disponible en el Instituto Nacional del Cáncer de los Estados Unidos.

El proceso de análisis de datos de mediciones de EM para la identificación de BM consta básicamente del procesamiento de datos, selección y validación de patrones con alto grado de discriminación intergrupal. Una vez validados estos patrones, la siguiente etapa es traducirlos a expresiones químicas para que puedan ser utilizados como herramientas de diagnóstico y en la sintonización de péptidos y antigénicos para el tratamiento de las enfermedades mencionadas anteriormente [20–27, 31].

En este trabajo se presenta una nuevo método para la identificación de patrones de BM a través de la combinación de dos pruebas estadísticas *t–test* y *Mann–Whitney U test* en un filtro estadístico y la eliminación iterativa de características de las zonas de interes usando por agrupamiento a la media de cada zona reduciendo la información redundante. Estos resultados son validados usando tres clasificadores independientes *Adaboost*, *TotalBoost* y *LPBoost*. Estos clasificadores son especialmente robustos contra los efectos del infraentrenamiento y sobreentrenamiento, además de ser fácilmente adaptables a este tipo de aplicaciones. Con estas características se evita caer en la mala interpretación de resultados erroneos y falsos positivos o negativos.

La medición del rendimiento se hace a través de validación cruzada y pruebas externas usando muestras independientes que no hayan intervenido previamente en la modelización de los clasificadores. Esta metodología presenta resultados prometedores con un rendimiento superior al 95 % de aciertos en las pruebas realizadas. Este trabajo se limita a la identificación de los patrones de biomarcadores, dejando para desarrollos posteriores su traducción a compuestos químicos y proteínas.

En la siguientes secciónes se describe algoritmicamente la metodología propuesta en este trabajo, se la pone a prueba usando dos conjuntos de datos, el primero *OvarianCD_PostQAQC.zip* del Programa de Proteínica Clínica del Centro para la Investigación del Cáncer perteneciente al Instituto Nacional del Cáncer de los Estados Unidos[2] y el segundo *Arcene* del *UCI Machine Learning Repository*. Se analizan y comparan los resultados obtenidos con trabajos anteriores. En la última sección se plantean posibles extensiones a la investigación presentada y en la parte final se adjunta la sección de conclusiones.

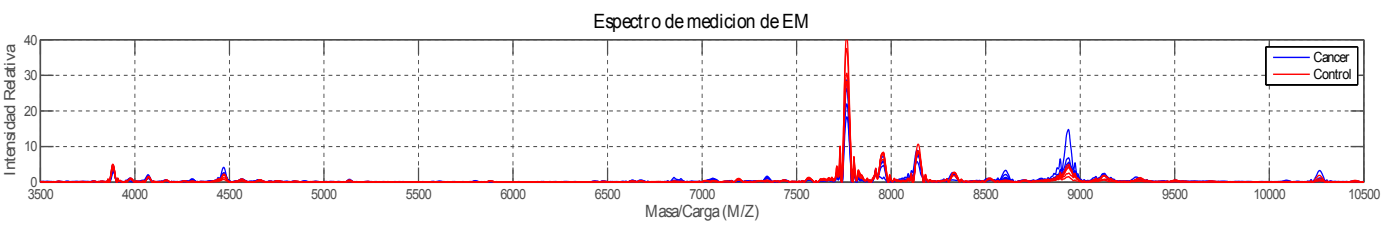


Figura 1. Ejemplo de espectros de mediciones del Conjunto *OvarianCD_PostQAQC.zip*

[1] La *World Health Organization* define una enfermedad crónica como aquella enfermedad de larga duración y progresión lenta.
[2] http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

2. DEFINICIÓN DE LA METODOLOGÍA PROPUESTA

La metodología propuesta en este trabajo parte del caso básico de análisis de datos para aplicaciónes de diagnóstico médico, un conjunto D formado por los grupos G_1 y G_2, donde G_1 representa un estado biológico normal y G_2 un estado biológico patológico. Del conjunto D se extraerán patrones que permitan diferenciar de entre G_1 y G_2 de forma fiable, estos patrones se validarán y una vez validados podrán ser usados como biomarcadores para la sintonización de proteínas que puedan ser usadas en el tratamiento de enfermedades. Este trabajo se limita a la definición de los patrones, dejando la sintetización de proteínas para trabajos futuros.

2.1 *Procesamiento de datos*

Las mediciones de los grupos $\{G_1, G_2\}$ del conjunto D vienen expresados como conjuntos de matrices de mediciones de la forma $M = \{((m/z)_1, i_1), ((m/z)_2, i_2), ((m/z)_3, i_3), \ldots, ((m/z)_n, i_n)\}$, donde los valores $\{(m/z)_k, i_k\} \in \mathbb{R}$ y $1 \leq k \leq n$. Debido a la heterogeneidad de las longitudes de las mediciones M, el primer paso a realizar es el procesamiento de estos datos. El procesamiento de los datos del conjunto D esta basado en [28–30] y consta de las siguientes etapas: remuestreo, corrección de la línea base, alineación de las mediciones, normalización y filtrado.

El remuestreo se basa en el concepto procedente del procesamiento de señales, donde dado una señal discreta en el tiempo, aplicando este concepto se puede reducir o aumentar su frecuencia de muestreo, en este caso, esto se entiende como aumentar o reducir la resolución de las medición y por tanto los elementos de cada vector de las mediciones. Para el procesamiento de datos se asume que las mediciones del conjunto de datos D vienen definidas por M_m, una matriz formada por dos vectores $\{I_m, (M/Z)_m\}$, dos señales discretas en el tiempo definida como $I_m = \{i_1, i_2, i_3, \ldots, i_n\}$ y $(M/Z)_m = \{(m/z_1), (m/z_2), (m/z_3), \ldots, (m/z_n)\}$ donde el índice n indica la resolución de la medición y m el número de la medición. Por regla general, en estas mediciones se cumple que la resolución de la medición $m-1 \neq m \neq m+1$. El remuestreo tiene como objetivo estas resoluciones a un valor común de homogeneización n_h fijando la dimensionalidad del conjunto de datos D en $m \times n_h$.

El algoritmo usado se basa en la combinación de un interpolador de señales, un filtro pasa bajos y un decimador. El interpolador eleva la resolución de las mediciones de un factor $n_1 > n$ luego de esto el filtro pasa bajos atenúa los efectos de aliasing e imaging producidos en la etapa de interpolación, finalmente el decimador reduce la resolución de un factor $n_2 < n_1$. La combinación de estas dos operaciones permite controlar el factor de homogeneización a un valor racional $\frac{n_1}{n_2}$ según sea necesario. En esta etapa el factor $\frac{n_1}{n_2}$ define a n_h. En ciertas aplicaciónes es necesario reducir la resolución radicalmente en factores de 10 a 1,sin embargo, en otras aplicaciónes, es necesario aumentar la resolución en ciertos segmentos del espectro para realizar análisis más puntuales.

Luego del procesamiento hay que eliminar un ruido típico presente en estos datos denominado efecto de la línea de base. Esta anomalía se produce debido a los contaminantes presentes en la muestra analizada y está presente en todas las mediciones en el segmento inicial de la medición. El algoritmo usado en esta etapa estima una frecuencia mínima de nivel de línea de base usando la frecuencia de las intensidades y el ruido de cada medición. Una vez estimada esta frecuencia de línea de base, mediante regresión de estos valores se obtiene un vector de valores de desplazamientos de línea de base de cada una de las intensidades de la medición procesada, el cual es finalmente substraído de las intensidades originales de la medición obteniendo un nuevo vector $I^{bc}_{m\times n}$ con el efecto de línea de base ya corregido(bc) sin alterar su resolución.

Con las mediciones ya homogeneizadas y eliminado de cada una de estas el efecto de línea base, el siguiente paso a realizar es la alineación de las medidas. La etapa de alineación de mediciones tiene como objetivo corregir los errores de calibración de los instrumentos de medida en el eje M/Z. Para esto se fijan picos de referencia de alineación $p = \{p_1, p_2, \ldots, p_\kappa\}$, donde κ para aplicaciónes prácticas toma valores de $3 \leq \kappa \leq 5 \forall \kappa \in Z^+$. Básicamente en esta etapa se reconstruyen nuevos vectores de intensidades $I^{alig}_{m\times n}$ tomando como referencia los picos de mayor intensidad de las mediciones. Esta reconstrucción está basada en uso de funciones de deformación temporal, mediante las cuales se desplazan los picos no alineados hacia picos de referencia adaptando su posición en el eje M/Z.

La etapa de normalización complementa a la alineación de mediciones en la corrección de los errores de calibración de los instrumentos, pero esta en cambio trabaja en el eje de las intensidades de abundancia. Esta etapa tiene como objetivo la reducción de las diferencias de las intensidades de las mediciones del conjunto D con respeto a un factor de normalización. El proceso consiste en identificar las máximas intensidades de cada una de las mediciones, a las cuales se le asignará un valor de normalización $norm_M$. Luego, todas las intensidades restantes de las mediciones se normalizarán con respecto a estas intensidades máximas de valor $norm_M$ obteniendo el factor de intensidad normalizada

$I_{i\times n}^{norm} = \frac{I_{i\times n}^{alig}}{max(I_{i\times n}^{alig})} \times norm_M$. La elección del factor de normalización dependerá de la naturaleza de la muestra analizada, su valor es esencial para una correcta traducción de los biomarcadores a proteínas.

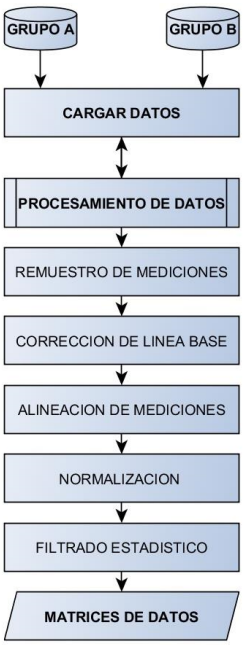
La etapa final del procesamiento de las mediciones es el filtrado del ruido de las mediciones. Si bien una etapa anterior realizó en parte un filtrado de ruido en la mediciones, todas las etapas anteriores al filtrado del ruido tienden a producir errores que se presentan como una nueva presencia de ruido en las mediciones. Desde un punto de vista práctico, este proceso de filtrado tiene como objetivo suavizar la curva del espectro eliminando la mayor cantidad de variaciones de carácter aleatorio de las mediciones. Esta etapa se lleva a cabo usando el filtro de Savitzky–Golay, el cual realiza una regresión polinomial de grado α con al menos $\alpha + 1$ puntos equidistantes, para determinar el valor de cada punto nuevo. El resultado de la aplicación de este filtro a las mediciones de EM son las mismas mediciones pero suavizadas, conservando su dimensionalidad.

La aproximación calculada a través de este filtro tiende a conservar las distribuciones originales de los datos filtrados, lo que significa que los puntos máximos y mínimos relativos o picos no se ven alterados. El proceso total del procesamiento de las matrices de medición se presenta en la Figura 2.

Algoritmo 1: Procesamiento de datos del conjunto $D = \{I_{m\times n},\ (M/Z)_{m\times n}\},\ n_1 \neq n_2 \neq \ldots \neq n$

Datos de entrada: $\{I_{m\times n},\ (M/Z)_{m\times n}, n_h, p, norm_M, \alpha\}$
Datos de salida: $D = \{I_{m\times n},\ M/Z\}$
Inicializar:
for $i = 1$ *until* m **do**
  $\{I_{i\times n},\ (M/Z)_{i\times n}\}$ = `remuestrear`$\{I_{i\times n},\ (M/Z)_{i\times n}, n_h\}$;
  $I_{i\times n}^{bc}$ = `corregirLineaBase`$\{I_{i\times n},\ M/Z\}$;
  $I_{i\times n}^{alig}$ = `alinear mediciones` $\left\{I_{i\times n}^{bc},\ M/Z,\ p\right\}$;
  $I_{i\times n}^{norm}$ = `normalización`$\left\{I_{i\times n}^{alig},\ M/Z,\ norm_M\right\}$;
  $I_{i\times n}^{filtr_{sg}}$ = `filtroRuido`$\left\{I_{i\times n}^{norm},\ M/Z,\ \alpha\right\}$;
  $I_{i\times n}^{proc} = I_{i\times n}^{filtr_{sg}}$;
  $(M/Z)_{i\times n}^{proc} = (M/Z)_{m\times n}$;
  if $i = m$ **then**
    $D = \left\{I_{i\times n}^{proc}, \texttt{mean}\{(M/Z)_{i\times n}\}\right\}$;
  end
end
Fin

(a)



(b)

Figura 2. (a)Descripción Algorítmica y (b)Diagrama de Flujo del Procesamiento de Datos de Mediciones de EM

2.2 *Selección de características discriminantes*

Una vez disponible el conjunto de datos D, el siguiente paso a realizar es de entre todas las mediciones, encontrar un subconjunto de valores $D^{red} = \{((m/z)_k, i_k\}, 1 < k < n$ en donde las intensidades y relaciones masa a radio k–esimas deben ser estadísticamente significativas en base a criterios que prueben cuan elevado es el grado de discriminación intergrupal que estas poseen.

Para esta etapa se ha diseñado una filtro estadístico que combina la prueba de t–Student y U–Mann–Whitney. El criterio de estas dos pruebas resulta ser complementario, ya que la primera asume que los grupos analizados poseen una distribución de probabilidad gaussiana, mientras que la prueba de *Mann–Whitney* asume que la distribución de probabilidad de los grupos analizados es la misma, pero no impone la condición de que esta sea gaussiana [35].

El proceso de filtrado se realiza sometiendo los datos de los grupos de analisis G_1 y G_2 del conjunto D por separado a cada una de estas pruebas estadísticas. De la aplicación de estas pruebas se obtienen p–valores que definen la probabilidad de variación en los i_k-ésimos valores de intensidades de las mediciones M_m, p–valores cuya probabilidad tienda a cero indicarán variaciones en cuyos valores de intensidad reside un alto poder de

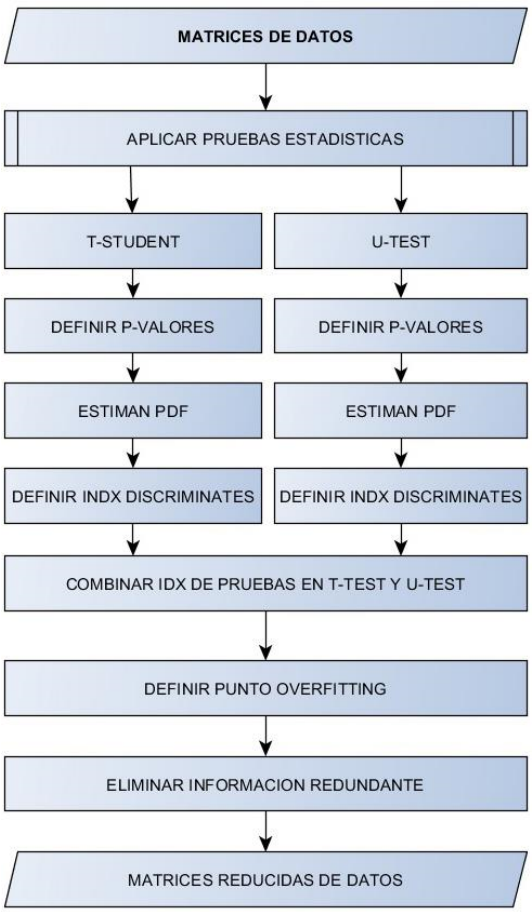
discriminación intergrupal. Es muy complicado definir cuantos de estos valores son necesarios para definir los patrones de biomarcadores, sin embargo es posible estimar cuantos de estos valores disminuyen al máximo el error de discriminación intergrupal en base al uso de algoritmos de clasificación supervisada.

Los algoritmos de clasificación supervisada han sido ampliamente utilizados en las aplicaciónes relativas al análisis de datos de espectrometría de masas [36–39]. Sin embargo algoritmos basados en la teoría de bayes, vectores de soporte de máquina, k-vecinos cercanos, e incluso las redes neuronales exigen un procesamiento adicional para cambiar las características originales del conjunto D a las exigidas como parámetros de entrada de estos algoritmos. Este procesamiento adicional limita la cantidad de información que puede ser extraída de los conjuntos, ya que los grupos analizados deben ser estadísticamente suficientes, la matriz del conjunto de datos debe ser cuadrada, no negativa, e invertible. Estas limitaciones son solucionadas a través de los algoritmos basados en *Boosting–Learning* sobre árboles de decisión binarios.

Algoritmo 2: Selección de características discriminantes $D = \{I_{m\times n},\ M/Z\}$

Datos de entrada: $D = \{I_{m\times n},\ M/Z\}$
Datos de salida: $D^{red} = \{I_{m\times red},\ M/Z\}$
Inicializar:
Sea: $G_1 = \{I_{m_1\times n},\ M/Z\}$ y $G_2 = \{I_{m_2\times n},\ M/Z\}$, $m = m_1 + m_2$;
Aplicar: $t = \dfrac{\mu_{G_1} - \mu_{G_2}}{s \times \sqrt{\dfrac{1}{m_1} + \dfrac{1}{m_2}}}$ y definir $p_{t-student}$-valor , $s = \dfrac{(m_1 - 1){S_A}^2 + (m_2 - 1){S_B}^2}{m_1 + m_2 - 2}$;
Aplicar: $U_1 = m_1 \times m_2 + \dfrac{m1_\times (m_1 + 1)}{2} - R_1$, $U_2 = m_1 \times m_2 + \dfrac{m_2 \times (m_2 + 1)}{2} - R_2$ y
$U = min(U_1, U_2)$; $R_1 = \sum_{i=0}^{n} \left(I_{m_1\times 1} + I_{m_1\times i+1}\right)$, $R_2 = \sum_{i=0}^{n} \left(I_{m_2\times 1} + I_{m_2\times i+1}\right)$;
Definir índices de los elementos de los vectores: $indx_t := p_{t-student}$ y $indx_U := p_U$ y definir p_U–valor;
Ordenar ascendetemente: $\{indx_t, p_{t-student}\}$ y $\{indx_U, p_U\}$;
Estimar función de distribución acumulada (fda) : $fda(p^{ordenado}_{t-student})$ y $fda(p^{ordenado}_{U})$;
Definir elementos de $p^{ordenado}_{t-student}$ y $p^{ordenado}_{U}$ cuya probabilidad tienda a cero: η_t y η_U;
Prueba de t-Student:
Definir conjuntos de entrenamiento y pruebas: $D_{entrenamiento-\eta_t)} = \{I_{m_\epsilon\times n}(indx_t(\eta_t)),\ M/Z\}$;
$D_{prueba-\eta_t} = \{I_{m_\tau\times n}(indx_t(\eta_t)),\ M/Z\}$;
$m_\epsilon =$ mediciones de entrenamiento y ;
$m_\tau =$ mediciones de prueba;
for $i = 1$ *until* η_t **do**
  $intx_{prueba} = indx_t(i))$;
  Modelar classificador:;
  $classificadorAdaboost = modelarClasificador(\{I_{m_\epsilon\times n}(indx_t(i)),\ M/Z\})$;
  Medir error de clasificación:;
  $errorClassificacion = errorClassificador(\{I_{m_\tau\times n}(indx_t(i)),\ M/Z\})$;
  $errorClassificacion(i) = errorClassificacion$;
  if $errorClassificacion(i-1) < errorClassificacion(i)$ **then**
    Guardar punto de inflexión:;
    $banderaSalida = 1$;;
    $ptoInflexion_t = i$;
    if $banderaSalida = 1$ **then**
      Terminar:
    end
  end
end
Definir pares $\{((m/z)_k\}$ con índices $indx_t(1\ hasta\ ptoInflexion_t)$;

(a)



(b)

Figura 3. (a)Descripción Algorítmica y (b)Diagrama de Flujo del Proceso de Selección de Características Discriminantes

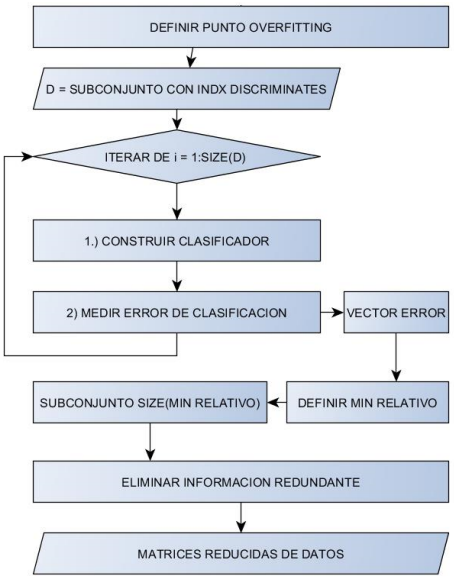
En cada una de las pruebas se producen dos vectores de p–valores independientes, p_1 y p_2, de dimensión n. A continuación, con cada uno de estos vectores se estiman Funciones de Probabilidad, sobre las cuales se fija un valor aproximado η de pares de $\{((m/z)_k, i_k\}$, cuya valor p tienda a ser cero. Luego para los p–valores de las dos pruebas estadisticas se modelan dos clasificadores binarios usando el algoritmo de `Adaboost`$M.1$ en base a los grupos de análisis G_1 y G_2, de donde se obtienen de manera aleatoria las mediciones que formarán parte de los conjuntos de entrenamiento y prueba del clasificador modelado. En cada una de las pruebas, una vez establecida la estructura del clasificador, iterativamente se va cambiando la dimensión de los conjuntos de entrenamiento y prueba desde 1 hasta η, a fin de encontrar un punto de inflexión en ς pares de $\{((m/z)_k, i_k\}$ donde el error de clasificación sea mínimo antes de la presencia de infraentrenamiento o sobreentrenamiento en los

clasificadores. La resolución del paso de iteración dependerá de la capacidad computacional disponible, sin embargo, se puede partir de una resolución determinada e ir aumentándola, el punto de inflexión del error de clasificación no cambia, de esta forma se puede conseguir un número preciso de pares $\{((m/z)_k, i_k\}$ con un alto grado de discriminación intergrupal.
Aquí nace un nuevo subconjunto, fijado por el punto de inflexión del rendimiento de los clasificadores. En este nuevo subconjunto se desecharán las características que producen errores de clasificación, normalmente originadas por problemas de ruido que no pudo ser filtrado en etapas anteriores. En este punto se han definido zonas de biomarcación donde los pares de $\{((m/z)_k, i_k\}$ tienden a ser redundantes, para eliminar esta redundancia adicionalmente a los filtros se realizó un agrupamiento de datos en función del promedio de los índices de $\{((m/z)_k\}$ definidos por el punto de inflexión en ς pares. El agrupamiento se realiza buscando un número de grupos en función de las medias de cada una de las zonas marcadas anteriormente. En la última parte se toman los dos índices más cercanos a la media de las zonas marcadas desechando los demás. De esta forma se elimina información redundante y se fortalece el modelamiento de los clasificadores con un número de pares $\{((m/z)_k\}$ de un alto grado de discriminación intergrupal. Finalmente se combinan los índices de los pares $\{((m/z)_k, i_k\}$ de cada una de las pruebas estadísticas, obteniendo asi el conjunto reducido de datos $D^{red} = \{I_{m\times k},\ M/Z\}$. Este procedimiento se muestran en las Figuras 3 y 4.

Algoritmo 3: Selección de características discriminantes $D = \{I_{m\times n},\ M/Z\}$ (continuación...)

1 Definir pares $\{((m/z)_k\}$ con índices $indx_U(1\ hasta\ pto_{Inflexion_U})$;
2 **Prueba de** U**–Mann–Whitney:** – idem **Prueba de t-Student:**
3 **Eliminar pares** $\{((m/z)_k\}$ **redundantes en** $indx_t(1\ hasta\ pto_{Inflexion_t})$**:;**
4 Agrupar pares $\{((m/z)_k\}$ en función de los índices definidos desde 1 hasta el $pto_{Inflexion_t}$;
5 Calcular la media de los grupos;
6 Guardar par superior e inferior $\{((m/z)_k\}$ cercano a la media de cada grupo;
7 Definir los índices de los elementos $\{((m/z)_k\}$ guardados: $indRed_t$;
8 **Eliminar pares** $\{((m/z)_k\}$ **redundantes en** $indx_U(1\ hasta\ pto_{Inflexion_U})$ **:;**
9 idem $indx_t(1\ hasta\ ptoInflexion_t)$;
10 **características discriminates:** $indx_T = indRed_t + indRed_U$;
11 **Conjunto de datos reducido en** $indx_T$**:;**
12 $D^{red} = \{I_{m\times n}(indx_T),\ M/Z(indx_T)\},\ \{I_{m\times red},\ M/Z\} = \{I_{m\times n}(indx_T),\ M/Z(indx_T)\}$;
13 $D^{red} = \{I_{m\times red},\ M/Z\}$;
14 **Fin**

(a)



(b)

Figura 4. (a)Descripción Algorítmica y (b)Diagrama de Flujo del Proceso de Selección de Características Discriminantes (cont...)

2.3 *Validación de características discriminantes*

La validación de los patrones detectados en la sección anterior se hacen usando tres clasificadores independientes, Adaboost, TotalBoost y LPBoost, donde en cada uno de estos se medirán los errores de clasificación usando validación cruzada y adicionalmente pruebas con muestras externas. Esta etapa consiste básicamente en la modelación de los clasificadores usando el conjunto D^{red}, medir su rendimiento, eliminar los predictores deficientes, luego volver a evaluar el rendimiento de los clasificadores. El modelamiento de los clasificadores y la eliminación de predictores deficientes mejoran el rendimiento de los clasificadores modelados. El rendimiento final es comparado entre los tres clasificadores a fin de medir la efectividad de los pares $\{((m/z)_k, i_k\}$ seleccionados en la sección anterior.

Adaboost se define básicamente como una metodología de aprendizaje mediante la cual se toma un algoritmo de clasificación sencillo y se la aplica iterativamente un número determinado de veces en forma secuencial, donde en cada iteración se mejora el error de clasificación, logrando rendimientos superiores a la aplicación de complejos algoritmos de clasificación. El algoritmo usado en esta etapa es la variante `Adaboost.M1` usada en de conjuntos de datos de dos grupos. En este artículo, se aplica este algoritmo en función de las m disponibles, para lo cual, primeramente hay que separar el conjunto total D^{red} en 3 subconjuntos de for-

ma aleatoria, entrenamiento (m_T), pruebas en validacion cruzada(m_{VC}) y pruebas externas(m_{VE}). Una vez realizada esta separación, se asume las (m_T) mediciones posee $\{I_1, I_2, \ldots, I_{m_T}\}$ vectores de intensidad asociadas a un predictor p, dichas intensidades están etiquetadas con un vector $y = \{+1, -1\}$, donde $+1$ etiqueta las mediciones del grupo G_1 y -1 las mediciones del grupo G_2. El clasificador $h((m_T)$ se obtendrá al entrenar un clasificador simple, en este artículo, un árbol de desición de dos ramales o binario. El error de clasificación de $h((m_T)$ estara definido por $\epsilon = \frac{1}{m_T}\sum_{i=1}^{m_T}\mathcal{I}(y_i \neq h((m_T))$. La función $\mathcal{I}(y_i \neq h((m_T))$ es 1 si se hubo acierto en la clasificación y 0 si hubo error. Este proceso se repite β veces, donde el clasificador final $\mathcal{H}$ será la combinación de todos los clasificadores $h((m_T)_\beta$ en función de un vector de ponderación $\mathcal{B}_\beta$. El clasificador final estará definido por $\mathcal{H}(h((m_T)) = sign\left(\sum_{i=1}^{\beta}\mathcal{B}_i h(m_T)_i\right)$, en este caso la función signo define si el elemento clasificado pertenece al G_1 o al G_2 mediante el mapeo en $y = \{+1, -1\}$ [33]. TotalBoost y LPBoost son dos variantes de Adaboost, que no necesitan el parámetro β para su entrenamiento, ya que buscan una solución óptima y limitan el número de interaciones en el entrenamiento de manera automática. Estos dos algoritmos son ideales para conjuntos limitados de datos, lo que los hace útiles en las aplicaciónes estudiadas en este artículo, donde es muy dificil disponer de enormes cantidades de mediciones de EM. LPBoost en comparación a Adaboost y Totalboost converge mas rápido hacia una solución final [40].

En la evaluación del rendimiento de esta etapa se usan dos métodos independientes, la validación cruzada y pruebas de clasificación usando muestras externas.El usar dos metodologías de evaluación del rendimiento en los clasificadores modelados robustece los resultados obtenidos. Una tendencia común a disminuir el error de clasificación en los tres clasificadores, indica que los datos escogidos tienen un alto grado de discriminación inter–grupal, el caso contrario en cualquiera de estas dos pruebas sera muestra suficiente sobre – entrenamiento o de infra – entrenamiento en el comportamiento de los clasificadores. Además de evidenciar los efectos del infraentrenamiento y sobreentrenamiento, la combinación de estas dos pruebas de rendimiento limitan encontrar falsos positivos o falsos negativos en las muestras analizadas, factor que resulta trágico en este tipo de aplicaciones. En la Figura 5 se presenta la aplicación de estos algoritmos en la metodología desarrollada en este artículo.

3. PRUEBAS DE LA METODOLOGÍA PROPUESTA

Las pruebas de la metodología descrita en este artículo fueron realizadas en Matlab usando el *Bioinformatics Toolbox*, *Statistics Toolbox* y el y el *Optimization Toolbox*, en una plataforma de hardware con Windows 7, velocidad de procesamiento de 2,4$[GHz]$ y 12$[GB]$ de memoria RAM.

Algoritmo 4: Validación de los datos del conjunto $D_{red} = \{I_{m\times red},\ M/Z\}$

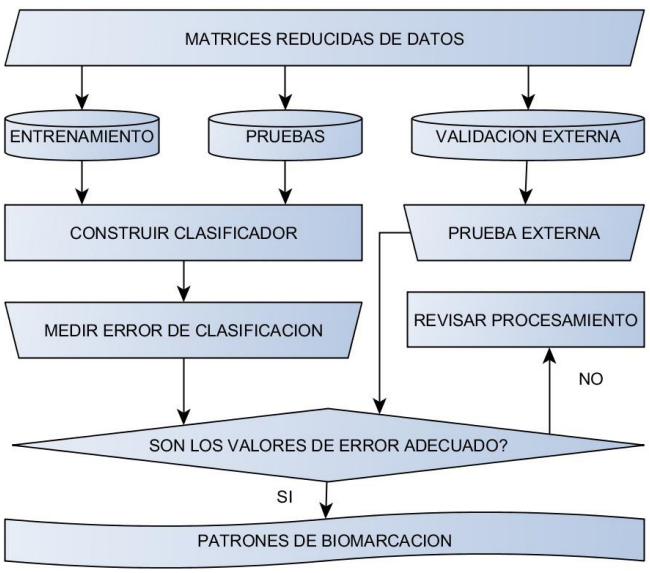
Datos de entrada: $D^{red} = \{I_{m\times red},\ M/Z\}, \beta$
Datos de salida: $errorEntrenamiento, errorPruebaExterna$
Inicializar:
  Dividir $D^{red} = \{I_{m\times red},\ M/Z\}$ en 3 subconjuntos;
    entrenamiento D_{m_T};
    pruebas en validación cruzada$D_{m_{VC}}$;
    y pruebas externas$D_{m_{VE}}$;
  Modelar clasificadores:;
  $classificadorAdaboost = \mathcal{H}_{Adaboost.M1}(D_{m_T}, \beta)$;
  $classificadorTotalboost = \mathcal{H}_{Totalboost}(D_{m_T})$;
  $classificadorLPboost = \mathcal{H}_{LPboost}(D_{m_T})$;

  Evaluar error de clasificación en cada clasificador usando validación cruzada:
  $errorAdaboost_{VC} = \bullet_{Adaboost.M1}(D_{m_{VC}})$;
  $errorTotalboost_{VC} = \bullet_{Totalboost}(D_{m_{VC}})$;
  $errorLPboost_{VC} = \bullet_{LPboost}(D_{m_{VC}})$;

  Evaluar error de clasificación en cada clasificador usando muestras externas:;
  $errorAdaboost_{VE} = \bullet_{Adaboost.M1}(D_{m_{VE}})$;
  $errorTotalboost_{VE} = \bullet_{Totalboost}(D_{m_{VE}})$;
  $errorLPboost_{VE} = \bullet_{LPboost}(D_{m_{VE}})$;

  Guardar errores de clasificación:;
  $errorEntrenamiento = \{errorAdaboost_{VC}, errorTotalboost_{VC}, errorLPboost_{VC}\}$;
  $errorPruebaExterna = \{errorAdaboost_{VE}, errorTotalboost_{VE}, errorLPboost_{VE}\}$;
Fin

(a)



(b)

Figura 5. (a)Descripción Algorítmica y (b)Diagrama de Flujo de la Validación de los Patrones de Biomarcación

Los resultados obtenidos se muestran a continuación divididos en la prueba de validación cruzada y pruebas de muestras externas respectivamente. En cada una de las curvas de error de clasificación mostradas, se puede ver claramente que los resultados muestran resultados prometedores sobre la metodología, cuyo rendimiento es superior al 95 % en simulaciones validadas en forma cruzada.

3.1 *Resultados usando validación cruzada*

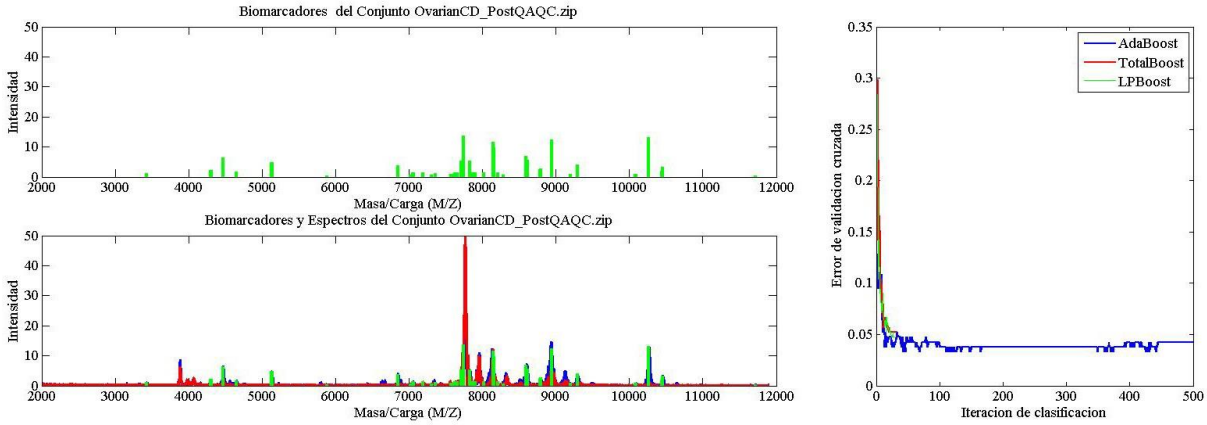


Figura 6. Patrones de Biomarcadores(54 pares detectados) – Conjunto OvarianCD_PostQAQC.zip

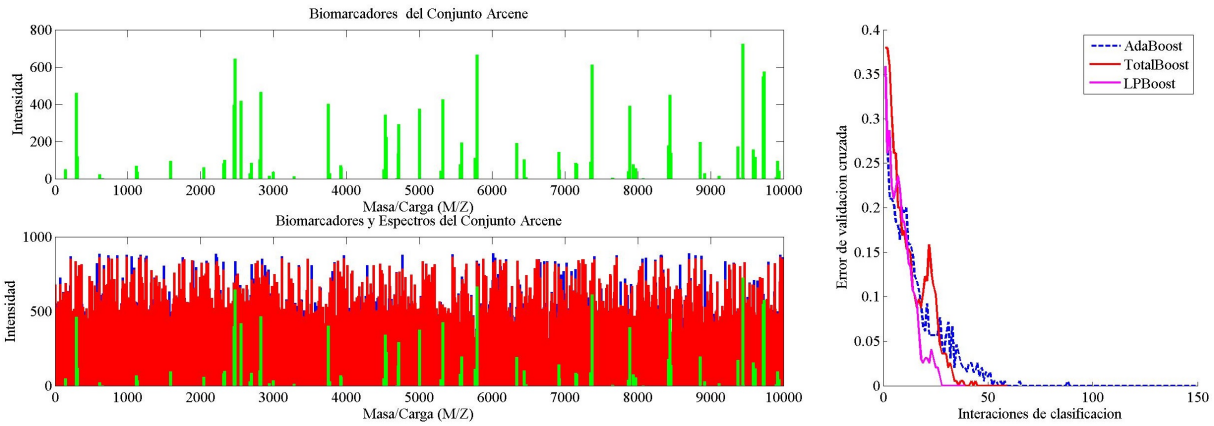


Figura 7. Patrones de Biomarcadores(93 pares detectados) – Conjunto Arcene

3.2 *Resultados usando muestras externas*

En las pruebas realizadas con muestras externas se realizaron clasificaciones de datos con los modelos generados obteniendose los siguientes resultados:

Grupos	Clasificación Correcta	Clasificación erronea
Normal	28	1
Cancer	29	2

Cuadro 1. Conjunto OvarianCD_PostQAQC5.zip

Grupos	Clasificación Correcta	Clasificación erronea
Grupo 1	305	5
Grupo 2	387	3

Cuadro 2. Conjunto Arcene

El error de clasificación en las pruebas realizadas con cada uno de los conjuntos analizados se calcula sumando el número de aciertos incorrectos y dividiendo esto para el número de aciertos correctos. Para OvarianCD_PostQAQC5.zip y Arcene el error de clasificación obtenido fue de 5,263 % y 1,156 % respectivamente.

4. ANÁLISIS DE RESULTADOS

Los resultados obtenidos al analizar los conjuntos *OvarianCD_PostQAQC.zip* y *Arcene* usando la metodología descrita en este artículo reflejan la efectividad de esta al alcanzar rendimientos superiores al 95 % de efectividad. Las curvas de medición de error usando validación cruzada reflejan además que los clasificadores modelados usando los datos del conjunto D^{red} poseen una gran resistencia a los efectos del infraentrenamiento y sobreentrenamiento, lo que disminuye la probabilidad de detección de falsos positivos o negativos.

5. TRABAJOS FUTUROS

Debido a la versatilidad de la EM como técnica de adquisición de datos, asi como también de la metodología propuesta en este trabajo, se plantean las siguientes lineas de investigacion basadas en este artículo:

1. Complementar la metodología descrita en este trabajo con la traducción de los biomarcadores a proteínas y antígenos.

2. Análisis de la composición química de plantas medicinales usadas en el tratamiento de enfermedades crónicas. [41,42]. La EM permite obtener información muy precisa acerca de la composición química de las muestras analizadas, en base a bibliotecas de compuestos quimicos disponibles en internet se puede aplicar la metodología propuesta en este trabajo para realizar busquedas exaustivas de estos compuestos en plantas medicinales endémicas de Ecuador [43].

3. Implementar la metodología usada en plataformas de supercomputación en donde se pueda controlar la dimensión del conjunto de datos analizados, pudiendo elevar la resolución de las mediciones de estos sin sacrificar el tiempo de procesamiento, memoria, ni otros recursos computacionales.

6. CONCLUSIONES

- La metodología propuesta en este trabajo es sencilla y requiere una supervisión mínima. No es necesario definir arbitrariamente un número de posibles características del conjunto analizado, ya que la metodología define por si sola el número de posibles características de los patrones de biomarcadores.

- La combinación de la *t–test* y *Mann–Whitney U test* en el filtro estadístico propuesto en este artículo resulta poderosa a la hora de extraer información. Ambas pruebas estadísticas definen un número menor de zonas de interés en comparación a cuando se las combina. El número de zonas de interés definido por la combinación de estas técnicas ofrece una mayor cantidad de información sobre posibles patrones de biomarcadores, los cuales presentan mayor resistencia a los feñomenos de infraentrenamiento y sobreentrenamiento en el modelamiento de sistemas de clasificación de datos.

- La sencillez de la metodología presentada en este trabajo presenta cualidades para ser optimizada e implementada en plataformas basadas en tarjetas procesadoras gráficas. Los cálculos realizados son básicamente operaciones numéricas sobre matrices masivas.

REFERENCIAS

[1] BAOLIN Wu,ABBOTT Tom , FISHMAN David , McMURRAY Walter, MOR Gil, STONE Kathryn, WARD David, WILLIAMS Kenneth y ZHAO Hongyu, *Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data*, Bioinformatics Journal, Print ISSN 1367-4803.Online ISSN 1460-2059

[2] WRIGHT Michael, HAN David, AEBERSOLDRuedi, *Mass spectrometry-based expression profiling of clinical prostate cancer*, Molecular & Cellular Proteomics Journal, Print ISSN 1535-9476, Online ISSN 1535-9484.

[3] PAULO A. Joao, KADIYALA Vivek, BANKS A. Peter, CONWELL L. Darwinl, STTEN Hanno, *Mass Spectrometry-based Quantitative Proteomic Profiling of Human Pancreatic and Hepatic Stellate Cell Lines*, Genomics, Proteomics & Bioinformatics Journal, ISSN: 1672-0229.

[4] CHO William C. S.,YIP Timothy T. C.,YIP Christine ,YIP Victor ,THULASIRAMAN Vanitha, NGAN Roger K. C., YIP Tai-Tung, LAU Wai-Hon,AU Joseph S. K., LAW Stephen C. K., CHENG Wai-Wai , MA Victor W. S., y LIM Cadmon K. P., *Identification of Serum Amyloid A Protein As a Potentially Useful Biomarker to Monitor Relapse of Nasopharyngeal Cancer by Serum Proteomic Profiling*, Clinical Cancer Research (CCR) Journal, Print ISSN: 1078-0432; Online ISSN: 1557-3265.

[5] ZHANG Z, BAST RC Jr, YU Y, LI J, SOKOLL LJ, RAI AJ, ROSENZWEIG JM, CAMERON B, WANG YY, MENG XY, BERCHUCK A, VAN Haaften-Day C, HACKER

NF, HW Bruijn DE, VAN der Zee AG, IJ Jacobs , ET Fung,DW Chan , *Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer*, Cancer Research (CanRes) Journal, Print ISSN: 0008-5472; Online ISSN: 1538-7445.

[6] Dr. PETRICOIN Emanuel F. PhD., ARDEKANI Ali M. PhD., HITT Ben A. PhD., LEVIANE Peter J. , FUSARO Vincent A., STEINBERG Seth M. PhD., MILLS Gordon B. MD., SIMONE Charles MD., FISHMAN David A. MD., KOHN Elise C. MD., LIOTTA Lance A. MD., *Use of proteomic patterns in serum to identify ovarian cancer*, The Lancet Journal (Vol. 359, Issue 9306, Pages 572-577) , ISSN: 0140-6736.

[7] WADSWORTH J. Trad , MD.;SOMERS Kenneth D. PhD.; BRENDAN C. Stack, Jr, MD.;CAZARES Lisa, BS.; GUNJAN Malik, PhD.; BAO Ling Adam, PhD.;WRIGHT George L. Jr, PhD.; O. John Semmes, PhD., *Identification of Patients With Head and Neck Cancer Using Serum Protein Profiles*, JAMA Otolaryngology–Head & Neck Surgery Journal, Print: ISSN 2168-6181, Online: ISSN 2168-619X.

[8] O. J. Semmes, L. H. Cazares, M. D. Ward, L. Qi, M. Moody, E. Maloney, J. Morris, M. W. Trosset, M. Hisada, S. Gygi y S. Jacobson, *Discrete serum protein signatures discriminate between human retrovirus-associated hematologic and neurologic disease*, Leukemia Journal, ISSN: 0887-6924, EISSN: 1476-5551.

[9] FERRARI Lorenza , SERAGLIA Roberta , ROSSI Carlo Riccardo , BERTAZZO Antonella ,LISE Mario , ALLEGRI Graziella y TRALDI Pietro ,*Protein profiles in sera of patients with malignant cutaneous melanoma*Rapid Communications in Mass Spectrometry Journal, ISSN: 1097-0231.

[10] WOODING Kerry M. y AUCHUS Richard J., *Mass spectrometry theory and application to adrenal diseases*, Molecular and Cellular Endocrinology Journal, ISSN: 0303-7207.

[11] McDONALD Jeffrey G., MATTHEW Susan ,AUCHUS Richard J., *Steroid profiling by gas chromatography-mass spectrometry and high performance liquid chromatography-mass spectrometry for adrenal diseases*, Hormones and Cancer Journal, ISSN: 1868-8497 (print version) ISSN: 1868-8500 (electronic version).

[12] LAPOLLA Annunziata, MOLIN Laura , and TRALDI Pietroi, *Protein Glycation in Diabetes as Determined by Mass Spectrometry*, International Journal of Endocrinology, ISSN: 1687-8337.

[13] LAPOLLA Annunziata,FEDELE1 D. y TRALDI Pietroi, *Diabetes and mass spectrometry*, Diabetes/Metabolism Research and Reviews Journal, ISSN: 1520-7560.

[14] LI Xiang , LUO Xiangxia , LU Xin, DUAN Junguo y XU Guowang , *Metabolomics study of diabetic retinopathy using gas chromatography–mass spectrometry: a comparison of stages and subtypes diagnosed by Western and Chinese medicine*, Molecular BioSystems Journal, ISSN: 1742-206X (print).

[15] FERNANDEZ Llama P., *Aportaciones de la proteómica al estudio de las enfermedades cardiovasculares*, Revista Hipertensión y Riesgo Vascular, ISSN: 1889-1837.

[16] CAO Yuan, HE Kun , CHENG Ming , SI Hai-Yani,ZHANG He-Lin, SONG Wei , LI Ai-Ling, HU Cheng-Jin , y WANG Na, *Two Classifiers Based on Serum Peptide Pattern for Prediction of HBV-Induced Liver Cirrhosis Using MALDI-TOF MS*, BioMed Research International Journal, ISSN: 2314-6133.

[17] A. K. Batta, R. Arora, G. Salen, G. S. Tint, D. Eskreis y S. Katz, *Characterization of serum and urinary bile acids in patients with primary biliary cirrhosis by gas-liquid chromatography-mass spectrometry: effect of ursodeoxycholic acid treatment*, Journal of Lipid Research, ISSN 0022-2275.

[18] MUSUNURI Sravanii , WETTERHALL Magnusl ,INGELSSON Martin , LANNFELT Lars , ARTEMENKO Konstantin ,BERGQUIST Jonas , Kúltima Kim , and SHEVCHENKO Ganna, *Quantification of the Brain Proteome in Alzheimer's Disease Using Multiplexed Mass Spectrometry*, Journal of Proteome Research, ISSN: 1535-3893.

[19] MATTHIESEN Rune and MUTENDA Kudzai E., *Introduction to Proteomics*, pp. 1-37, Mass spectrometry data analysis in proteomics / edited by Rune Matthiesen, ISBN-13: 978-1-58829-563-7.

[20] FUSHIKI Tadayoshii, FUJISAWA Hironori y EGUCHI Shinto, *Identification of biomarkers from mass spectrometry data using a çommon"peak approach*, BMC Bioinformatics Journal, ISSN 1471-2105.

[21] PHAM P., *A Novel Algorithm for Multi-class Cancer Diagnosis on MALDI-TOF Mass Spectra*, Bioinformatics and Biomedicin IEEE Journal, pages 398-401, ISBN 978-1-4577-1799-4, 12-15 Nov. 2011.

[22] JELONEK Karol , ROS Malgorzata ,PIETROWSKA Monika, WIDLAK Piotr, *Cancer biomarkers and mass spectrometry-based analyses of phospholipids in body fluids*, Clinical Lipidology Journal, ISSN 1746-0875, pages 137-150, 2013/2.

[23] PIETROSWSKA M., JELONEK K., MICHALAK M., ROS M.,RODZIEWICZ P.,CHMIELEWSKA K ,POLAMSKI K ,POLANSKA J,KLOSOK A Gdowicz,GIGLOK M,SUWINSKI R,TARNAWSKI R , DZIADZIUSZKO R, RZYMAN W ,WIDLAK P, *Identification of serum proteome components associated with progression of non-small cell lung cancer*, Acta biochimica Polonica Journal, 2014/5.

[24] G. A. GOWDA Nagana , ZHANG Shucha,GU Haiwei , ASIAGO Vincent , SHANAIAH Narasimhamurthy, y RAFTERY Daniel, *Metabolomics-Based Methods for Early Disease Diagnostics - A Review*, Expert Review of Molecular Diagnostics Journal, Sep 2008; 8(5): 617–633, ISSN 1473-7159.

[25] TARAWNEH Sandra K. Al,BORDER Michael B.,DIBBLE Christopher F., y BENCHARIT Sompop,*Defining Salivary Biomarkers Using Mass Spectrometry-Based Proteomics - A systematic review*, OMICS A Journal of Integrative Biology, ISSN: 1536-2310.

[26] Dr. LEE Yu Hsiang, Phd. y Dr.WONG David T., DMD., DMSC. *Saliva - An emerging biofluid for early detection of diseases*, Am J Dent 2009;22:241-8.

[27] KHADIR Abdelkrim and TISS Ali, *Proteomics Approaches towards Early Detection and Diagnosis of Cancer*, Carcinogenesis & Mutagenesis Journal, ISSN: 2157-2518.

[28] GIL Alterovitz, RAMONI Marco F., *Systems Bioinformatics: An Engineering Case-based Approach*, cap. 4, Editorial: Artech House; Edición: Har/Cdr (1 de marzo de 2007), ISBN-10: 1857431820.

[29] EIDHAMMER Ingvar, FLIKKA Kristian, MARTENS Lennart, MIKALSEN Svein-Ole, *Computational Methods for Mass Spectrometry Proteomics* , Wiley & Sons Publications, January 2008, ISBN: 978-0-470-51297-5.

[30] EIDHAMMER Ingvar, BARSNES Harald, EGIL EIDE Geir, MARTENS Lennart, *Computational and Statistical Methods for Protein Quantification by Mass Spectrometry*, Wiley & Sons Publications, February 2013, ISBN: 978-1-119-96400-1.

[31] TESSITORE Alessandra, GAGGIANO Agata, CICCIARELLI Germanai, VERZELLA Daniela, CAPECE Daria, FISCHIETTI Mariafausta, ZAZZERONI Francesca, y ALESSE Edoardo, *Serum Biomarkers Identification by Mass Spectrometry in High-Mortality Tumors*, International Journal of Proteomics, Volume 2013 (2013), Article ID 125858, 15 pages, ISSN 1874-3919.

[32] SAEYS Yvan, INZA Inaki y LARRANAGA Pedro, *A review of feature selection techniques in bioinformatics*, Bioinformatics Journal, ISSN 1460-2059, 2007.

[33] HE Ping, *Classification Methods and Appplications to Mass Spectrometry Data*, PhD. Thesis, Hong Kong Baptist University, 2005.

[34] XU Q. ,MOHAMED S.S. ,SALAMA M.M.A.,KAMEL M. y RIZKALLA K., *Mass Spectrometry-Based Proteomic Pattern Analysis for Prostate Cancer Detection Using Neural Networks with Statistical Significance Test-Based Feature Selection* , Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference.

[35] GUYON, I., GUNN, S., NIKRAVESH, M., ZADEH, L.A., *Feature Extraction – Foundations and Applications*, pp. 90, Studies in Fuzziness and Soft Computing, Vol. 207, Springer Publications, ISBN 978-3-540-35488-8.

[36] SINGH Ajit P. , HALLORAN John , BILMES Jeff A. , KIRCHOFF Katrin , NOBLE William S. , *Spectrum Identification using a Dynamic Bayesian Network Model of Tandem Mass Spectra*, Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI2012), ISSN 2159-5399.

[37] BJM Webb-Robertson , *Support vector machines for improved peptide identification from tandem mass spectrometry database search*, Mass Spectrometry of Proteins and peptides: Methods in Molecular Biology Journal, Vol 146. Humana Press, New York, NY, ISSN 1064-3745.

[38] WU Baolin, ABBOTT Tom, FISHMAN David, MCMURRAY Walter, MOR Gil, STONE Kathryn, WARD David, WILLIAMS Kenneth and ZHAO Hongyu,, *Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data*, March 6, 2003, Bioinformatics Journal, ISSN 1367-4803.

[39] QU Yinsheng, ADAM Bao-Ling, YUTAKA Yasui, WARD Michael D., CAZARES Lisa H., SCHELLHAMMER Paul F., FENG Ziding, SEMMES O. John, and WRIGHT JR. George L., *Boosted Decision Tree Analysis of Surface-enhanced Laser Desorption/Ionization Mass Spectral Serum Profiles Discriminates Prostate Cancer from Noncancer Patients*, October 2002 vol. 48 no. 10 1835-1843, Clinical Chemistry Journal, ISSN 0009-9147.

[40] NARSKY Ilya , PORTER Frank C., *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*, Wiley-VCH; 1 edition (October 24, 2013), ISBN: 9783527677290 - 3527677291.

[41] HETHELYI E., TETENYI P.,DABI E, DANOS B., *The role of mass spectrometry in medicinal plant research*, Biological Mass Spectrometry Journal, Online ISSN: 1096-9888.

[42] IDOYAGA Moliona Natilde y LUXARDO Natalia, *Medicinas no convencionales en cáncer*, Medicina (B. Aires) [online]. 2005, vol.65, n.5, pp. 390-394. ISSN 1669-9106.

[43] MANZANO SANTANA Patricia , ORELLANA LEÓN Tulio , MARTÍNEZ MIGDALIA Miranda C., ABREU PAYROL C. Juan , RUÍZ Omar , PERALTA GARCÍA C. Esther L., *Algunos parámetros farmacognósticos de Vernonanthura patens (Kunth) H. Rob. (Asteraceae) endémica de Ecuador*, Rev Cubana Plant Med vol.18 no.1 Ciudad de la Habana ene.-mar. 2013, ISSN 1028-4796.