ISSN: 1692-7257 - Volumen 1 – Número 21 - 2013

Revista Colombiana de
Tecnologías de Avanzada

Recibido: 27 de julio de 2012
Aceptado: 16 de octubre de 2012

IMPLEMENTATION OF PATTERN RECOGNITION TECHNIQUES (SUPPORT VECTOR MACHINES - LEAST SQUARE) IN SELECTION OF PARAMETERS USED IN METABOLOMIC SYSTEMS

IMPLEMENTACIÓN DE TÉCNICAS DE RECONOCIMIENTO DE PATRONES (LEAST SQUARE SUPPORT VECTOR MACHINES) EN PROCESOS DE SELECCIÓN DE PARÁMETROS CARACTERÍSTICOS APLICADOS A SISTEMAS METABOLÓMICOS

MSc(c). William Villamizar Rozo, MSc. Luis E. Mendoza, MSc(c). Pablo Santafé G.

Universidad de Pamplona, Facultad de Ingenierías y Arquitecturas.
Km. 1, vía a Bucaramanga, Pamplona, Norte de Santander, Colombia.
E-mail: {wiviro, luis.mendoza, pablosantafe}@unipamplona.edu.co.

Abstract: This paper presents a metodología que involved, multivariate analysis techniques and pre-processing stage in order to determine characteristic metabolites in a given spectrum. This novel approach allowed us to determine that certain metabolites are modified by the addition of different concentrations of about feature SVM LS-NMR data. Validating processes also achieved as peak alignment, normalization, baseline correction and analysis multienergy in metabolomic data in olive oils and pure and mixed with hazelnut alterations 2%, 5%, 10%, 20% and 30% .

Keywords: Metabolomic, HNMR, LS-SVM, COW.

Resumen: En este artículo se presenta una metodologíaque involucra, técnicas de análisis multivariable y una etapa de pre-procesamiento con el fin de determinar metabolitos característicos en un determinado espectro. Este método novedoso permitió determinar que ciertos metabolitos son modificados por las diferentes concentraciones y además de conocer la funcionalidad de LS-SVM en datos NMR. También se logró validar procesos como: alineamiento de picos, normalización, corrección de línea base y análisis multienergía, en datos metabolómicos en aceites de oliva y avellana puros y mezclados con alteraciones de 2%, 5%, 10%, 20% y 30%.

Palabras clave: Metabolómica, HNMR, LS-SVM, COW.

1. INTRODUCCIÓN

La metabolómica es el análisis global de todas o un gran número de sus metabolitos celulares [1]. Siendo la metabolómica originalmente propuesta como un método de genoma funcional [2], además generan una gran cantidad de datos de diferentes orígenes (microorganismos, plantas, animales e inclusive humanos). El valor del nivel del metabolito es de suma importancia ya que permite diferenciar los cambios efectuados en una muestra.

Para su proceso, manipulación y análisis es un claro reto que requiere de matemática especializada, estadística o herramientas bio-informáticas. Estos datos se generan de diferentes técnicas analíticas comunes como: GC-MS, LC-MS, CE-MS, FTIR y finalmente la resonancia magnética nuclear (NMR) [3], siendo esta última, una alta técnica de análisis no destructiva reproducible que provee información acerca de todos los metabolitos.

Con el propósito de reducir el número de variables de los datos metabolómicos generados por el espectrómetro NMR, se optimiza el funcionamiento implicando un cierto riesgo de pérdida de información. Por este motivo las variables deben seleccionarse cuidadosamente; una selección inadecuada de variables puede llevar a un funcionamiento inaceptable del sistema. El uso de características y las formas en los datos para la clasificación es conocida como huella metabolómica [2] y los métodos más utilizados son análisis de componentes principales con análisis de discriminantes lineales (PCA-LDA) [4], mínimos cuadrados parciales y análisis de discriminantes lineales (PLS-LDA) [4], sin embargo estos métodos pueden proveer buenos resultados de clasificación pero suelen ser difíciles de interpretar. Existen otros métodos de clasificación aplicados a datos metabolómicos como: redes neuronales artificiales [5], programación genética [6], algoritmos genéticos [7], los cuales usan un algoritmo genético para alinear picos en datos metabolómicos NMR, siendo estas técnicas de aprendizaje computacionales basadas en la teoría de la evolución de Darwin [8], y son populares para solucionar problemas de optimización. Finalmente fue incorporada la teoría de aprendizaje estadístico al introducir las maquinas de soporte vectorial (SVM) como una nueva clase de algoritmo de clasificación [9].

El objetivo de este estudio es determinar las diferencias relevantes en la composición metabólica de muestras de aceites de oliva y avellana, puros y mezclados con un total de 189 medidas NMR con adulteraciones del 2%, 5%, 10% y 30%.Donde la selección de variables o como es el caso en particular, los metabolitos presentes en las muestras de aceites serán seleccionados aplicando una metodología de preprocesamiento. Una vez obtenida esta información se clasifican los diferentes espectros metabolómicos por medio de maquinas de soporte vectorial con mínimos cuadrados (LS-SVM), permitiendo comprobar las variables influyentes en la alteración del aceite de oliva o avellana.

A nivel general en la metabolómica, las variables seleccionadas o metabolitos característicos tiene una gran importancia y aplicación en diferentes campos: en comparación de mutantes [10,2], estudio para efectos globales de manipulación genética [11,2], toxicología [12,2], descubrimiento de nuevos medicamentos [13,2], nutrición [14,2], diabetes [15], cáncer [16], y descubrimiento de productos naturales [17].

2. MATERIALES Y MÉTODOS

2.1 Materiales

Los datos espectroscópicos HNMR que a continuación se describen fueron proporcionados por la *Universidad de Rovira i Virgili*; dichos datos constan de señales provenientes de aceite de oliva virgen y aceites de avellanas puras y mezcladas con avellana del tipo av y avp. A continuación se da una explicación del conjunto de medidas.

Se han hecho un total de 189 medidas NMR de aceites de oliva y avellana, puros y mezclados
Las medidas están hechas con 6 aceites:

- 4 de oliva virgen extra:ca, ch, cp, oay.
- 2 de avellana:avp y av.

Las adulteraciones son del 2% (02), 5% (05), 10% (10) y 30% (30) con los aceites de avellana avp y av. Estas medidas representan el valor de la intensidad del espectro (a frecuencia relativa. Van desde un valor aproximado de -2.2 a 11.5, pasando por cero).

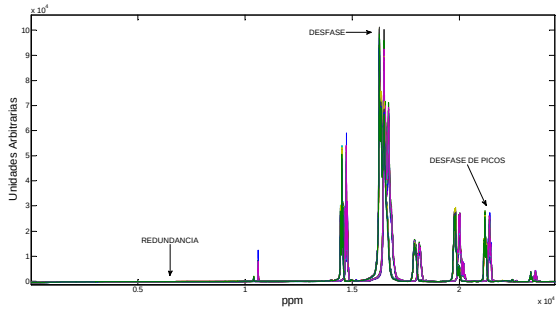


Fig. 1. Señales espectroscópicas NMR de muestras de aceite de oliva con diferentes concentraciones de aceite de avellana.

En la figura 1. Se observa una muestra real de las señales espectroscópicas NMR de aceite oliva con diferentes concentraciones de aceite de avellana. En este ejemplo se evidencia la gran cantidad de información redundante y desfasada que requieren de un tratamiento con técnicas de análisis multivariable y además de un riguroso pre-procesamiento.

2.2 Metodología

Como se ha comentado anteriormente, el tratamiento de los datos, es un claro desafío y requiere matemática estadística especializada, para realizar procesos de alineamientos, corrección de line a base, y normalización entre otras, además de herramientas bioinformáticas, las cuales facilitan el uso de técnicas de Reconocimiento de Patrones, Redes Neuronales, Algoritmos Genéticos, Programación Genética entre otras.

A continuación, se presenta un método novedoso para determinar que ciertos metabolitos son modificados por las diferentes concentraciones y conocer la funcionalidad de LS-SVM en datos NMR.El proceso metodológico para la selección de variables (metabolitos más relevantes) y la validación de estos metabolitos usando un sistema de clasificación es presentado en la figura 2.

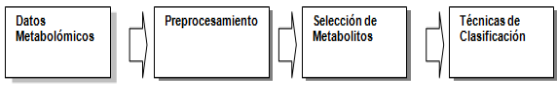


Fig. 2: Proceso de extracción de los metabolitos más relevantes

Como se puede observar el proceso está dividido en 4 etapas: La primera etapa consiste en los datos originales, la segunda etapa el acondicionamiento previo de los datos, la tercera etapa la extracción del metabolito y finalmente se emplea la técnica de clasificación multivariables. A continuación se explica cada una de las etapas.

2.2.1. Datos Metabolómicos

Los espectros provenientes del NMR presentan información en abundancia como ubicación y función de los elementos en las moléculas, presentes en la posición o niveles de los picos. Dando no sólo una determinación cuantitativa sino además cualitativa. Otro problema presente en los espectros NMR es el solapamiento o desfase como se observa en la figura 3.

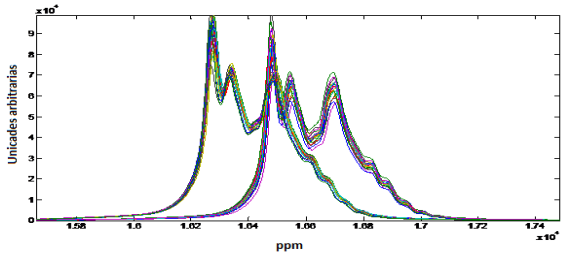


Fig. 3: Muestra de datos metabolómicos.

2.2.2 Preprocesamiento de datos.

Las técnicas de preprocesamiento aplicadas en este trabajo son: Normalización, escaldo min-máx., zona de interés, corrección de línea base y alineamiento de picos utilizando programación dinámica.

a) Normalización y Escalado Min- Max.

El objetivo de escalar una serie de datos, es realizar una transformación de estos o reducir la influencia de altas variables inconsistentes, dentro de un conjunto de valores apropiados según el caso.

La técnica escogida en esta metodología es la normalización o escalado min-máx., uno de los tipos de normalización más utilizados y es la técnica más simple: los valores mínimos y máximos de los “scores” se desplazan a los valores 0 y 1, respectivamente y todos los scores se transforman en el rango [0,1], de manera que la distribución original se mantiene (excepto para el factor de escala).

Se realizó un algoritmo de normalización, aplicando la ecuación (1):

$$y' = \left(\frac{y - \min}{\max - \min} \right) (\max' - \min') + \min' \qquad (1)$$

Además, se tuvo en cuenta que al normalizar las columnas, se pierde la relación original entre los componentes de cada vector o ?la. Por ello, se llevó a cabo la normalización de los vectores (norma 1), es decir, se cálculo el módulo de cada vector, y cada una de sus componentes se dividió por este valor. La figura 1, muestra una señal normalizada.

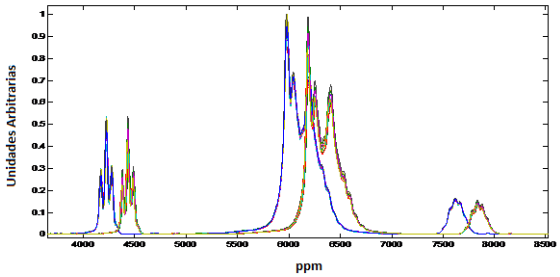


Fig. 4. Normalización de la data metabolómica original con un mínimo de 0 y un máximo de 1

b) Zona de interés

Seguidamente del proceso de normalización y escalado se aplica el algoritmo que se encarga de determinar la zona de interés. Como se representa en la figura 5. Esta zona tiene la mayor cantidad de metabolitos característicos de la muestra metabolómica.

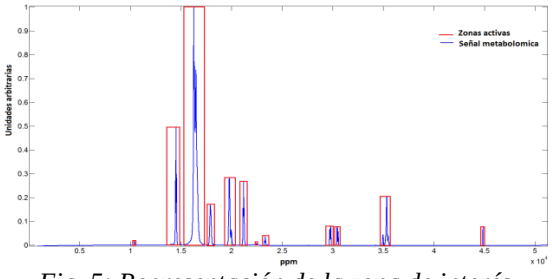


Fig. 5: Representación de la zona de interés.

El criterio de selección de las zonas significativas de los datos metabolómicos se obtuvo con la siguiente expresión:

$$x = \begin{cases} \bar{x}, & x > U \\ 0, & resto \end{cases} \qquad (2)$$

Donde $\boldsymbol{x}$ es la matriz de datos NMR, $\boldsymbol{U}$ es el umbral que determina el criterio de selección de la zona activa y finalmente $\tilde{x}$ es la región más relevante o zona activa.

c) Corrección de línea base

Este método involucra la derivada para encontrar concavidades y se basa en análisis geométrico; los pasos del proceso se describen a continuación:

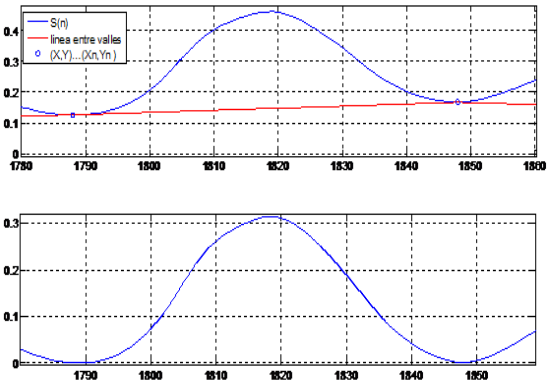


Fig. 6: Efecto de la corrección de línea base, señal superior original y señal inferior es la señal corregida.

En la figura 6, se muestra un resultado de la corrección de la línea base, para llegar a estos, se realizaron cinco procesos:

d) Alineamiento de picos

En Resonancia Magnética Nuclear (NMR) o Cromatografía los picos pueden ser desplazados debido a variaciones por el instrumento o interferencias en el análisis. Estos cambios deben ser corregidos antes de ser analizados o procesados.

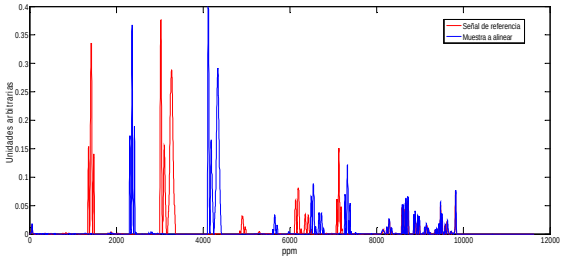


Fig.7: Ejemplo de espectros que necesitan alineamiento.

En la figura 7. Se muestra partes de dos espectros NMR que necesitan alineamiento. Donde la señal de color azul esta desalineada con respecto a la señal de referencia indicada en color rojo.

Uno de los procedimientos más robustos en alineamiento de picos hecho inicialmente para datos cromatográficos es el llamado Alineamiento de Correlación Optimizado (COW).

Esta técnica de alineamiento que opera por segmentos y utiliza programación dinámica, hace alineamiento de una muestra con otra de referencia por medio de dilatación o expansión del segmento de la muestra usando interpolación lineal.

Los pasos que involucran al alineamiento COW se describen en [20]. En este trabajo el tamaño del *slack* tiene un rango de 1 a15.

2.2.3. Selección del metabolito o selección de variables.

Inicialmente se seleccionan dos espectros o muestras, uno de aceite puro y otro mezclado, utilizando el acondicionamiento de la señal tomando como base la metodología anteriormente mencionada. Donde las datas a analizar se procesan, tomando todos los niveles máximos o picos. Este análisis no resultó muy conveniente ya que se presentaron similitudes y no se pudo extraer un patrón característico que permitiera diferenciar los aceites puros y mezclados. Para lograr determinar o extraer la zona que presenta mayor diferencia entre estos tipos de aceite, se determinó un umbral por medio de niveles micros en sus amplitudes y con ayuda de técnicas de procesamientos implícitos en la metodologías, específicamente con el método de alineamiento de picos COW tomando parámetros con un "*slack*" y longitud del segmento dentro de un intervalo de 26 a 33 respectivamente; para el análisis multienergía el tamaño de las ventanas se seleccionó un rango entre 5 y 20 puntos, obteniéndose con estas técnicas de exploración y determinación de los metabolitos una diferencia notoria entre las muestras de señales de aceites.

2.2.4 Técnicas de Clasificación

Una vez, realizado el proceso de alineamiento, se procede a validar esta extracción de información, con un método de clasificación.

- LS-SVM

En el estudio de clasificación se empleó LS-SVMLab v1.8 de Matlab, con el propósito de

detectar cambios que se puedan presentar en las señales de aceite de oliva mezclados con diferentes tipos de concentración de aceite de avellana, tomándose 5 datas de entrenamiento y 5 de validación para un total de 25 repeticiones aleatorias, discriminadas en tres clases, con una longitud de cada vector de 11626 datos, cada muestra de ejecutándose en el siguiente orden:
Los datos de entrenamiento se distribuyen en tres grupos de aceites: oliva, avellana avp y avellana av. Para la prueba se seleccionan al azar señales de aceites de oliva con concentraciones de 2%, 5%, 10% y 30%de aceite de avellana av y avp.

Finalmente se obtiene la clasificación de estas señales. Antes de realizar estos pasos, se debe tener en cuenta algunos parámetros de referencia: El coeficiente de dispersión (gamma) en un intervalo de 700 a 1000, la varianza (sigma2) con un margen de 8000 a 10000 y la función de espacios característicos (Kernel) de RFB. Las pruebas obtenidas con estos parámetros resultaron relevantes en el proceso de clasificación. El proceso de validación se aplica en el siguiente orden:

a) Señal original sin tratamiento (data cruda).
b) Normalización o escalado de los datos.
c) Zonas de interés.
d) Corrección de línea base.
e) Alineamiento de picos.
f) Análisis multienergía.

3. RESULTADOS

Los resultados relevantes obtenidos de la metodología presentada se describen a continuación. Estos son presentados en el orden mostrado en la metodología.

3.1 Análisis data cruda (señal original)

Los datos HNMR de aceite (ca), fueron repartidos en 9 clases, para un total de 44 x 65535 muestras sin procesar. Un ejemplo de una muestra de este tipo de señal es observado en la figura 9.

A continuación se describe el funcionamiento de la técnica de LS-SVM multiclase, en el proceso de clasificación. Tomando como datos de entrenamiento 5 muestras (ca, ca05p, ca10p, ca20p, ca30p), del total de las muestras; 5 de prueba al azar con diferentes porcentaje de mezclado del 5% al 30%con los parámetros de gamma en un rango de 1 a 1000 y un sigma en un intervalo de 1000 a 9000, un Kernel RBF, generando los siguientes resultados de porcentaje de clasificación verdaderos; ver tabla 1.

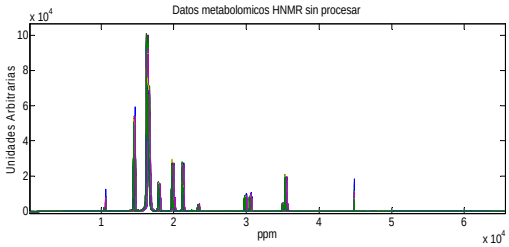


Fig. 9: Datos metabolómicos HNMR de aceite de oliva puros y mezclados con aceite de avellana.

3.2 Normalización y escalado min-máx.

Para el caso de estudio de los espectros de aceite de oliva con mezclas de aceite de avellana presentaron niveles máximos y mínimos entre un margen de -437.5780 ppm a 78464ppm respectivamente.
El escalado min-máx no mejoró los resultados como se evidencia en la tabla 1.

3.3 Corrección de línea base

En el preprocesamiento de datos metabolómicos es importante y de gran relevancia llevar los datos a una línea común debido a que estos presentan una variación a lo largo del espectro.

El método aplicado a la totalidad de las señales, presentó mejoras con resultados de aproximadamente en un 100%, en diferentes sitios del espectro; esto se evidenció cuando se le examinó la energía tanto en la señal con zona activa y la señal con corrección de línea base, generando una energía de 163.5287, conservándose la energía en ambas señales. Es así como se concluye que el porcentaje de error en la línea base es del 0%.Al aplicar la señal generada por el método de corrección de línea base al algoritmo de LS-SVM aplicado en Matlab los resultados obtenidos no presentaron variaciones con respecto a la zona de interés, como se puede evidenciar en la tabla 1. Con los mismos parámetros, como: sigma, gamma, kernel, datos de entrenamiento y prueba.

3.3 Alineamiento de picos

Para corregir dicho desalineamiento, es necesario seleccionar una señal de referencia respecto a la cual serán alineadas todas las señales de HNMR. La señal de referencia seleccionada para la base de datos de aceites fue una de las señales de aceite de oliva al 100% o puro de color azul como se representa en la figura 12.

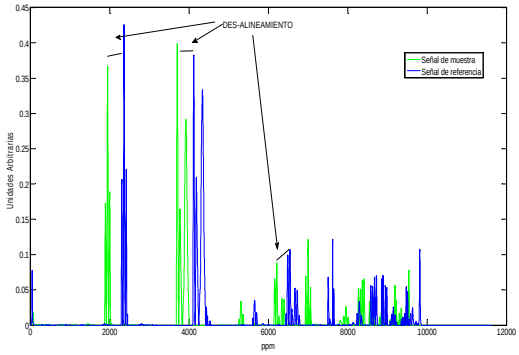
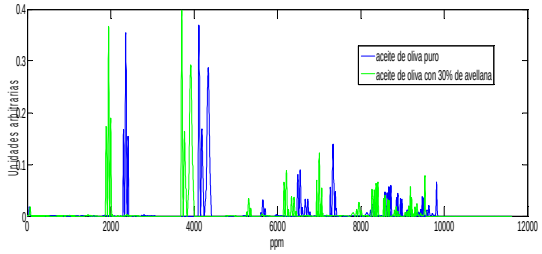


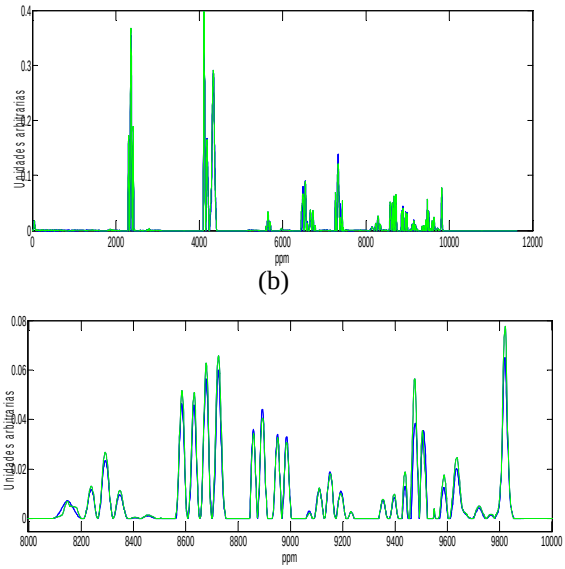
Fig. 12: Espectros HNMR de aceite de oliva puros (azul) y aceite de oliva mezclados con aceite de avellana al 30% (verde) desalineados.

La aplicación del método de alineamiento dinámico en el tiempo (DTW); tomando como referencia una señal de oliva pura en color azul y la señal a alinear una de oliva mezclada al 30% con aceite de avellana, indicada en color verde. Los resultados no fueron muy óptimos en ciertas zonas donde se presentaban la concentración máxima de picos, más exactamente en la región comprendida entre 3500 y 4000 ppm. Además este método consume un tiempo de aproximadamente 50 minutos de ejecución en Matlab.

Por otro lado, los parámetros óptimos asignados al método de alineamiento de picos COW como: El tamaño del "*slack*" y la longitud del segmento están dentro de un rango de 32 y 24 respectivamente. Los espectros a alinearse se tomaron teniendo como referencia una señal de oliva pura en color azul y la señal a alinear una de oliva mezclada al 30% con aceite de avellana indicada en color verde. Tal como se muestra en la figura 14. Donde el alineamiento es de aproximadamente de un 97.72% en el total de los datos. En la figura 14c. Se muestra una ampliación entre el intervalo de 8000 a 10000 ppm donde se aprecia la potencialidad del método de alineamiento de picos por correlación optimizado COW. Además se logra exitosamente el alineamiento, sin que interese que los picos desalineados queden relativamente cerca o lejos.



(a)



(b)

(c)

Fig.14: Alineamiento de picos utilizando el método COW. a) espectro de referencia (azul) y espectro a alinear (verde). b) alineamiento de las dos señales y c) ampliación de la zona alineada.

Una vez obtenida la base de datos de espectros alineados inicialmente se les aplicó el análisis multienergía con el propósito de buscar los patrones relevantes (metabolitos) que permitieran diferenciar los espectros metabolómicos (selección de parámetros) de los aceites en estudio. Los resultados arrojados por este análisis fueron los siguientes. En la figura 15, se observan una imagen que representa 44 señales (número de filas) de aceite de oliva con mezclas de diferentes concentraciones de avellana, alineadas por el método de COW, y tratadas por análisis multienergía. En este caso se muestran ventanas de 10, 50 y 150 datos, con el propósito de escoger la ventana optima, que nos permita visualizar y determinar las diferencias entre los grupos de las señales metabolómicas.

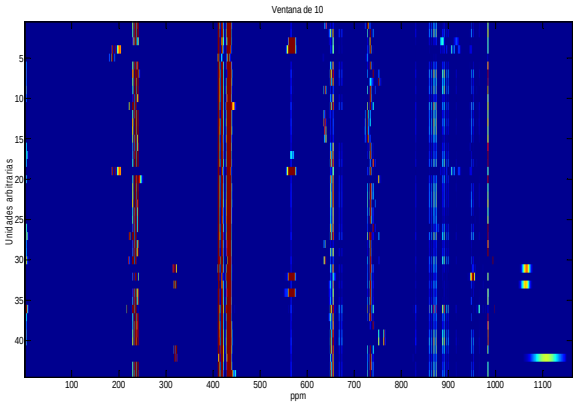


Fig.15: Análisis multienergía, con ventanas de 10.

Una vez seleccionadas las señales relevantes, de esta base de datos siendo elegidos16, de los 44 espectros; resultando ser las señales de aceite de oliva mezcladas con Avellana tipo avp, en sus diferentes concentraciones. Con esta base de datos se aplicó el clasificador LS-SVM, con los mismos parámetros antes expuestos, teniendo como resultado el siguiente porcentaje de aciertos de clasificación; ver tabla 1. Por otro lado, en la figura 16, se observa el resultado del análisis de componentes principales, en este caso se escogió la primera componente representando un porcentaje del 98.7% de la energía total de la señal, Donde la dispersión de las clases equivalentes a las diferentes concentraciones, representadas con colores (amarillo, violeta, rojo y azul) es muy alta y por ende no permite escoger y clasificar por simple inspección las diferentes concentraciones.

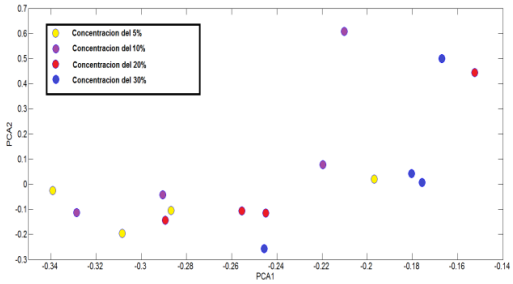


Fig.16. Diagrama de dispersión de Análisis de componentes principales de los espectros de aceite de oliva mezclados con aceite de avellana en concentraciones del 5% ,10% ,20% y 30%.

Los problemas presentes en los picos de amplitud grande y sin poder determinar el parámetro característico que hace que una muestra de aceite de oliva mezclado con aceite de avellana presente diferencias en comparación con una de aceite de oliva puro, se hizo una exploración en los picos del orden de los **niveles más pequeños** y siguiéndose con la misma metodología se obtuvieron los siguientes resultados:

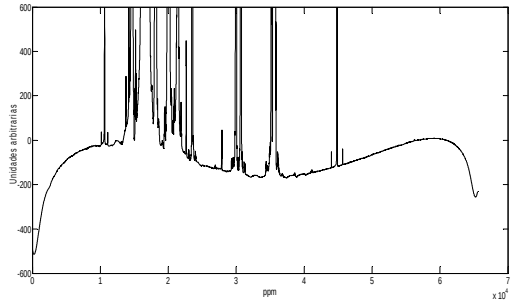


Fig.17: Zona umbralizada de un espectro HNMR de aceite de oliva con concentraciones del 30% de aceite de avellana.

En la figura 17 se muestra un ejemplo, de cómo se detectaron los picos de amplitudes menores a 600, con el fin de extraer la zona con picos de amplitudes más pequeñas. Este umbral se escoge por inspección, ya que funcionan para el 100% de las señales (44 en total).

Seguidamente al proceso de exploración en picos pequeños, se aplicaron, las técnicas de alineamiento de picos, por correlación optimizada, tomando como referencia una longitud del segmento en un rango de 24 a 28 y una flexibilidad o "*slack*" en un intervalo de 28 a 32. Finalizando el pretratamiento con el análisis multienergía, donde el parámetro de la ventana fue de 7. Reflejando las diferencias presentes en los dos espectros antes mencionados, específicamente en la zona de 100 ppm a 300 ppm. Como se muestra en la figura 18.

En la figura 18, se reflejan los picos de interés o metabolitos que hacen diferentes los espectros de aceites, presentándose en la figura superior un máximo de siete picos en comparación con la figura inferior con un máximo de 5 picos. A la base de datos realizada con esta información relevante de metabolitos se le aplicó el método de clasificación LS_SVM multiclase, donde se presentó un conjunto de espectros de 44 señales de aceites de oliva con sus diferentes concentraciones, además de 10 señales entre, aceites de avellana puros tipo av y avp. Generándose los siguientes resultados de aciertos de clasificación. Ver tabla 2.

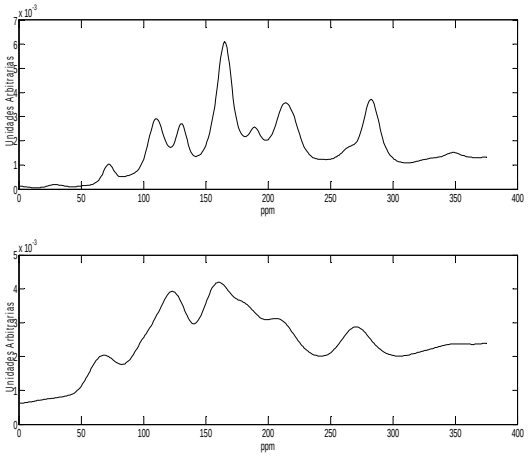


Fig. 18: Diferencia de metabolitos, de un espectro de aceite de oliva mezclado con avellana (superior) con uno de aceite de oliva puro (inferior).

Con los resultados de la tabla 1, se verifica que los aceites de oliva mezclados con aceite de avellana, presentaban alteraciones, ya que la cantidad de picos es superior en número.

Tabla 1: Resultados obtenidos con técnicas LS-SVM para datos con información relevante de metabolitos presentes en aceites de oliva puros y mezclados con avellana.

Mezclado	2%	5%	10%	20%	30%	Clasificación
Avellana av	5	5	5	5	5	100%
Avellana avp	5	5	5	5	5	100%
Oliva puro	4	4	4	4	4	100%

Finalmente estos resultados se compararon con estudios realizados en aceites de oliva con mezclas de aceite de avellana [18, 19,] determinándose que el metabolito presente es el acidolinolénico; en esta región en particular del espectro HNMR de aceite de oliva mezclado con avellana, sin importar la concentración. Específicamente en la zona de alrededor de 15100 ppm a 15350 ppm como se aprecia en la parte superior de la figura 19.

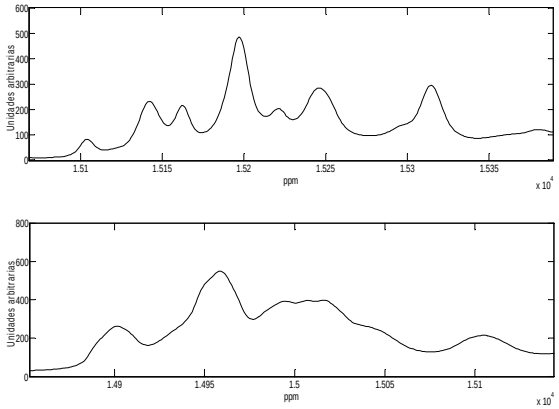


Fig. 19. Diferencias de cantidad de picos. La figura superior corresponde a presencia de ácido linolénico debido a mayor numero de picos, en comparación con la grafica inferior.

Tabla 2: Resultados obtenidos con técnicas LS-SVM para los datos en sus diferentes etapas de preprocesamiento de aceite de oliva puros y mezclados con avellana.

Preprocesamiento	Mezclado	2 %	5 %	10 %	20 %	30 %	Clasificación
Datos crudos		5	5	5	5	5	40%
Datos Normalizados		5	5	5	5	5	40%
Datos zonas de interés		5	5	5	5	5	40%
Datos con corrección de línea base		5	5	5	5	5	50%
Datos con alineamiento		5	5	5	5	5	66%

4. CONCLUSIONES

El alineamiento de picos COW y el análisis multienergía es de vital importancia, cuando se trata de señales HNMR, ya que juega un papel importante en la detección de metabolitos que permiten diferenciar los aceites de oliva y avellana. Con esta metodología y con ayuda de LS-SVM multiclase se permitió detectar el metabolito que hace la diferencia entre los aceites de avellana y oliva, en el caso en particular es el ácido linolenico. Que está en la región de 15100 ppm a 15350 ppm del espectro de aceite.

REFERENCIAS

[1]. Oliver S.G, Winson MK, Kell D.B, et al. Systematic functional analysis of the yeast genome.TrendsBiotechnol 1998; 16:373–8.

[2]. Vladimir Shulaev, Metabolomics technology and bioinformatics. Briefings in Bioinformatics.2006; Vol. 7. No 2. 128 -139.

[3]. Viant MR, Rosenblum E.S, Tieerdema RS. NMR-based metabolomics: a powerful approach for characterizing theeffects of environmental stressors on organism health. Environ SciTechnol 2003; 37:4982–9.

[4]. Hong-Seok Son, Ki Myong Kim, Frans Van Den Berg,Geum-Sook Hwang, Won-Mok Park, Cherl-Ho Lee, and Young-ShickHong,*J.* ^{1}H Nuclear Magnetic Resonance-Based Metabolomic Characterization of Wines by Grape Varieties and Production Areas.*Agric. Food Chem.* **2008,** *56,* 8007–8016

[5]. D. F. Brougham, G. Ivanova, M. Gottschalk, D.M. Collins, A. J. Eustace, R. O'Connor, and J. Havel,Artificial Neural Networks for Classification in Metabolomic Studies ofWhole Cells Using ^{1}H NuclearMagnetic Resonance.

[6]. Richard J. Gilbert, Helen E. Johnson, Michæl K. Winson, Jem J. Rowland,RoystonGoodacre, Aileen R. Smith, Michæl A. Hall and Douglas B. Kell. Genetic Programming as an Analytical Tool for MetabolomeData;Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion.

[7]. Z. Ramadan, D. Jacobs, M. Grigorov, S. Kochhar. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. Elsevier, Talanta 68 (2006) 1683–1691

[8]. Darwin on the origin of species by means of natural selection. *Canadian Naturalist and Geologist* 5:100-120.

[9]. Vapnik, V. (1998b).The support vector method of function estimation. In J. A. K. Suykens, & J. Vandewalle, (Eds.), *Nonlinear Modeling: Advanced Black-box Techniques*. Boston: Kluwer Academic Publishers.

[10]. Jan Luts, Fabian Ojeda, Raf Van de Plasa,Bart De Moora, Sabine Van Huffela, Johan A.K. Suykensa , A tutorial on support vector machine-based methods for classification problems in chemometrics. Elsevier AnalyticaChimicaActa 665 (2010) 129–145

[11]. Raamsdonk LM, Teusink B, Broadhurst D, et al. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 2001;19:45–50.

[12]. Catchpole GS, Beckmann M, Enot DP, et al. Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. PNAS 2005;102:14458–62.

[13]. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics' understanding the metabolic responses of living systems topathophysiological stimuli via multivariate statistical analysisof biological NMR spectroscopic data. Xenobiotica 1999; 29:1181–9.

[14]. Watkins SM, German J. B. Metabolomics and biochemical profiling in drug discovery and development.CurrOpinMolTher 2002;4:224–8.

[15]. Watkins SM, Reifsnyder PR, Pan HJ, et al. Lipid metabolome-wide effects of the PPARgammaagonistrosiglitazone. J Lipid Res 2002;43:1809–17.

[16]. [16]Jennifer L. Spratlin,NatalieJ. Serkova, and S. Gail Eckhardt, Clinical Applications ofMetabolomics in Oncology: AReview.ClinCancerRes2009;15:431-440.

[17]. Fiehn O, Kopka J, Trethewey RN, et al. Identification of uncommon plant metabolites based on calculation ofelemental compositions using gas chromatography andquadrupole mass spectrometry. Anal Chem 2000; 72: 3573–80.

[18]. Georgia Vigli, AngelosPhilippidis, Apostolos Spyros, and PhotisDais, Classification of Edible Oils by Employing 31P and 1H NMR Spectroscopy in Combination with Multivariate Statistical Analysis. A Proposal for the Detection of Seed Oil Adulteration in Virgin Olive Oils,*J. Agric. Food Chem.* **2003,** *51,* 5715-5722

[19]. Mannina Luisa, Segre Annalaura,High Resolution Nuclear Magnetic Resonance: From Chemical Structure to Food Authenticity,Grasas y Aceites, Vol. 53. Fasc. 1 (2002), 22-33.

[20]. Niels-Peter Vest Nielsen, Jens Michael Carstensen, JørnSmedsgaard ,*Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping. Journal of Chromatography A, 805 (1998) 17–35.