Recepción: enero 2013
Aceptación/publicación: abril/julio 2013

REVISTA CUBANA DE Química

QUIMIOMETRIX II, una plataforma automatizada para el procesamiento multivariante de datos químicos y bioquímicos. Experiencias de aplicación

QUIMIOMETRIX II, an Automatic Platform for the Multivariate Processing of Chemical and Biochemical Data. Applications

Dra. Isneri Talavera-Bustamante, MSc. Lázaro Bustio-Martínez, Lic. Yenice Coma-Peña, Dra. Noslen Hernández-González
italavera@cenatav.co.cu

Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), Habana, Cuba

Resumen

Quimiometrix II es un sistema de herramientas concebido para el preprocesamiento, exploración, clasificación y calibración de datos químicos y bioquímicos. Los métodos empleados están sustentados sobre las técnicas más novedosas para el reconocimiento estadístico de patrones. Se implementa en una arquitectura para el desarrollo de software científico basada en *Plugin* que permite la instalación, actualización y control de errores de manera automatizada. En esta versión se incorporan algoritmos novedosos propios para la calibración multivariante, así como una librería matemática que los implementa (Quimilap). Se presenta un caso de estudio donde se muestra el empleo del software.

Palabras clave: análisis exploratorio de datos, clasificación, calibración multivariada.

Abstract

Quimiometrix II is a Toolbox System for preprocessing, exploration, classification and calibration of chemical and biochemical data. The used methods are held on the most innovative techniques in the Pattern Recognition field. It is implemented in an architecture for the development of scientific software in *Plugin* that permits the installation, actualization, and the deal with error management in an automatic way. This version incorporates innovative own algorithms for multivariate calibration purposes, as well as an own mathematical library named (Quimilap). A study case is presented in order to demonstrate de software behavior.

Keywords: exploratory data analysis, classification, multivariate calibration.

Introducción

Cuba ha presentado un crecimiento sostenido del desarrollo de la instrumentación en los laboratorios químicos del país. Desde finales de la década de los años 90 del pasado siglo se han adquirido equipos de última generación que facilitan los análisis a realizar y, a la vez, incrementan la calidad de los mismos. Este hecho provocó que el volumen de información a procesar creciera extraordinariamente, y las capacidades de procesamiento existentes resultaran insuficientes. La Quimiometría, como nueva especialidad dentro de la química, emerge en el escenario mundial para la solución de esta problemática, combinando de forma sistémica el empleo de técnicas químicas, matemáticas e informáticas que permiten la obtención y análisis de la información escondida, tras las inmensas matrices numéricas generadas en los procesos de investigación.

Se hace inminente la necesidad de desarrollar nuevos métodos automatizados para extraer toda la

información útil de los enormes volúmenes de datos que se iban obteniendo.

En Cuba, ante esta necesidad y la imposibilidad de adquirir productos que se iban generando en el mercado como los software (Pirouette. Infometrix Inc. USA /1/, The Unscrambler, CAMO Inc. Suecia /2/, PLS Toolbox, Eigenvector Research Inc. USA /3/), producto del bloqueo y sus altos costos, se acomete por parte del CENATAV, en el año 2006, el desarrollo del sistema de herramientas quimiométricas Quimiometrix /4-7/, producto autóctono, con posibilidades de una amplia generalización para el procesamiento y análisis de datos químicos, que contuviera las mejores características de los software homólogos más reconocidos internacionalmente y que a la vez incluyera los últimos adelantos en el área.

Quimiometrix se encuentra desplegado en más de 18 instituciones que abarcan desde ambientes académicos, como la Universidad de La Habana y la CUJAE, hasta el Laboratorio de Arqueometría perteneciente a la Oficina del Historiador de la Ciudad. Luego de 4 años de finalizado el proyecto Quimiometrix en su primera versión, se puede concluir que superó todas las expectativas en cuanto a aceptación por parte de los especialistas. Esto propició que se realizaran sugerencias, se demandaran nuevos requerimientos y prestaciones, y se detectaran aspectos de diseño e implementación que debían ser mejorados para futuras versiones.

El presente trabajo tiene como objetivo presentar la nueva versión de Quimiometrix, destacando especialmente los aportes teóricos y prácticos que se introducen y mostrar a través de un caso de estudio soluciones alcanzadas en el quehacer investigativo nacional, así como las potencialidades y facilidades que ofrece dicho software para estos fines. Los ejemplos fueron seleccionados sobre la base de que fuesen representativos de varios campos de aplicación de las técnicas quimiométricas, como la calibración y la clasificación multivariante.

Fundamentación teórica

De una manera muy simple, la Quimiometría puede ilustrarse como una forma de representar una sustancia química mediante una tabla o matriz de datos donde se miden determinadas características. El hecho de contar con una tabla de valores numéricos indica la posibilidad de aplicar operaciones estadísticas y algebraicas sobre dicha tabla para obtener conocimiento subyacente y precisamente sobre esto trata el Reconocimiento de Patrones. Por lo tanto, cabe decir que la Quimiometría es un área válida para la aplicación de los diferentes algoritmos de Reconocimiento de Patrones con el fin de obtener conocimiento oculto en los datos. En suma, el objetivo principal del Reconocimiento de Patrones enfocado hacia la Quimiometría está en la identificación de relaciones y/o vínculos entre objetos (muestras químicas o grupos de estas). Las muestras deben ser previamente caracterizadas a través de los diversos métodos de análisis instrumental que permitan su procesamiento /8/.

Debido a la representación de los datos y a su naturaleza, la Quimiometría utiliza principalmente el enfoque estadístico del Reconocimiento de Patrones, por lo que son perfectamente válidos los diferentes métodos de trabajo, tales como: el Análisis Exploratorio de Datos, los Métodos de Clasificación y Métodos de Regresión Multivariada. Quimiometrix puede ser considerado como un amplio paquete que incluye las herramientas quimiométricas más utilizadas en la actualidad por los especialistas de la materia, abarcando todos los métodos de trabajo de la Quimiometría.

● Métodos y condiciones experimentales

Descripción del software

Quimiometrix consiste en un paquete de herramientas quimiométricas para el análisis multivariante de datos, soportadas en técnicas estadísticas y de Reconocimiento de Patrones. El software está compuesto por tres módulos básicos:

Módulo de Gestión, preprocesamiento, transformación y exploración de datos

✓**Gestión de datos:** Como su nombre lo indica, el área de trabajo para la gestión de datos permite la entrada, ordenamiento, edición y salida de los mismos a través de una amigable interfaz gráfica compuesta por un menú, un explorador de objetos y un área de trabajo.

✓ **Pretratamiento de los datos:** Los datos experimentales deben ser mejorados y preparados convenientemente para el análisis. El objetivo es eliminar matemáticamente fuentes de variaciones indeseables, que pueden influenciar en los resultados finales. Se divide en dos tipos de técnicas, las de preprocesamiento y las de transformación.

a. **Preprocesamientos:** /9-12/ Son un conjunto de operaciones estadísticas que se emplean principalmente para lograr que los datos estén dentro de un mismo rango de valores y siguiendo un comportamiento similar. Entre ellas tenemos: centrado respecto a la media, autoescalamiento, escalamiento por amplitud y escalamiento por la varianza.

b. **Transformaciones:** /9-16/ Conjunto de técnicas matemáticas que se emplean fundamentalmente para disminuir o eliminar las variaciones aleatorias (ruido experimental) y variaciones sistemáticas de la señal medida. Entre ellas tenemos: alisamientos, correcciones de la línea base, corrección multiplicativa de la dispersión (MSC), conversión logarítmica, conversiones espectroscópicas y normalización**.**

✓ **Exploración de los datos**: Las técnicas de exploración de datos, se utilizan para poner de manifiesto y resaltar la información contenida en una matriz de datos multidimensional. Son herramientas estadísticas de proyección y graficación, muy útiles para la identificación de tendencias y patrones, especialmente en matrices de alta dimensión. Entre las más utilizadas y que fueron incluidas en el software tenemos:

a. **Análisis por componentes principales:** /17-20/ (conocida por sus siglas en inglés como "PCA", Principal Component Analysis): es un método de proyección. Proyecta los datos multivariados en un espacio de dimensión menor, reduciendo la dimensionalidad del espacio del conjunto de los datos y por esto se considera un método de "compresión".

b. **Análisis de agrupamiento jerárquico**: /21/ (conocido por sus siglas en inglés como "HCA", Hierarchical Cluster Analysis): El objetivo es formar grupos conteniendo objetos semejantes. Los resultados son presentados en forma de un árbol jerárquico conocido con el nombre de DENDROGRAMA.

✓ **Módulo de clasificación:** Permite la construcción de modelos de clasificación capaces de determinar de forma automática a qué grupo de objetos (clases) pertenece un nuevo objeto, a partir de las características obtenidas por los diferentes análisis químicos realizados. Estos modelos se sustentan en técnicas de reconocimiento de patrones (clasificadores) que son usadas para establecer las semejanzas y diferencias entre diferentes tipos de muestras, comparándolas entre sí. La utilización de uno u otro tipo de clasificador estará en dependencia de las características de los datos y la complejidad del problema de clasificación /22/. Se incluyeron dentro del software los siguientes clasificadores:

a. K-Vecinos más cercanos "k-NN" /23/.

b. Modelado Blando Independiente de Analogías de Clases "SIMCA" /24/.

c. Análisis Discriminante, Regresión por Mínimos Cuadrados Parciales "PLS-DA".

d. Máquinas de Soporte Vectorial para Clasificación "SVC" /25-26/.

✓ **Módulo de calibración.** Permite la construcción de modelos de calibración multivariada capaces de predecir de forma automática propiedades químico-físicas de interés de muestras en estudio, a partir de los resultados obtenidos por los diferentes análisis químicos realizados. Estos modelos se sustentan en métodos de regresión con diferentes posibilidades de acuerdo a sus características. Se construye el modelo de calibración con el método de regresión que más se ajuste a las características de los datos y el problema a resolver y se valida el mismo con muestras externas al conjunto de entrenamiento. Se incluyen en Quimiometrix los siguientes métodos de regresión lineal y no lineal:

a. Regresión por Componentes Principales "PCR" /27/.

b. Regresión por Mínimos Cuadrados Parciales Multivariante "PLS1" /28-30/.

c. Regresión por Mínimos Cuadrados Parciales Multivariante "PLS2" /31/.

d. Máquinas de Soporte Vectorial para Regresión "SVR" /26/.

e. **Método Predictivo-Frecuentista** /32/.

f. **FDA-SVM** /33/.

g. **GIR** /34/.

Los tres últimos métodos de regresión son algoritmos propios incluidos en esta nueva versión, y por su importancia como aporte al conocimiento serán descritos:

Método Predictivo-Frecuentista

El método Predictivo-Frecuentista propone un estimador de regresión para el problema de calibración estadística, sobre un enfoque predictivo no Bayesiano. Puede ser descrito partiendo de un vector aleatorio *Y* (puede ser un espectro infrarrojo de una mezcla de sustancias), está relacionado con una variable de interés *X* (concentración de una de las sustancias que componen la mezcla), de acuerdo con un modelo estadístico especificado por la densidad condicional de probabilidad $f(X,\Theta)$, la cual depende de un vector desconocido de parámetros Θ. Fundamentado en un conjunto de datos de entrenamiento (Xi,Yi), se quiere, dado el valor de una Y_0 (observación futura), estimar el valor correspondiente de X_0. Este método, en contraste con los métodos existentes en la literatura, no está basado en máxima verosimilitud, ni es un método Bayesiano, sino que se sostiene en un enfoque predictivo no Bayesiano.

FDA–SVR

El Análisis de Datos Funcionales (FDA) es una extensión del análisis multivariado tradicional que está orientado específicamente a tratar con observaciones de naturaleza funcional. Cada objeto está caracterizado por una o más funciones continuas de valores reales, en lugar de por un vector de dimensión finita. La técnica FDA-SVR, permite solucionar problemas químicos de calibración tanto lineales como no lineales.

GIR

Sea (Ω, F, P) un espacio de probabilidad, *X* un espacio medible abstracto y (X,Y) un elemento aleatorio de Ω que toma valores en $X \times R$. El modelo en que se basa este método tiene la forma $X(.) = \mu(Y, .) + \varepsilon(.)$ donde ε es un proceso aleatorio con función de media μ. Apoyado en realizaciones independientes de (X, Y), dígase (Xi,Yi), $i = 1,\ldots,n$, el objetivo es estimar el valor de un dato futuro y_0 a partir de un valor observado. Este método realiza esa estimación según un estimador de $\gamma(X) = E\ Y/X)$].

Análisis y diseño de Quimiometrix

Quimiometrix en su versión 1.0 fue concebido como un sistema monolítico lo cual dificultó su mantenimiento, extensiones e implementación. Teniendo en cuenta estas experiencias , Quimiometrix v. 2.0 fue diseñado para enfrentar dichas limitaciones y permitir la expansión del proyecto mediante la inclusión de manera sencilla de los nuevos algoritmos que surjan. Para ello se concibió una arquitectura basada en plugin (aplicación informática que interactúa con otra aplicación para aportarle una función o utilidad específica extendiendo así el funcionamiento de la segunda), donde los algoritmos son desplegados sin detener el uso del sistema, ofrece además un mecanismo de actualizaciones automáticas e instalación remota que permite su actualización continua. Implementa un mecanismo de control y detección de errores que reporta los malos funcionamientos de manera automática.

Un valor agregado importante en Quimiometrix es su poder de cómputo al contar con una librería matemática de desarrollo propio nombrada como **Quimilap**. Es una librería genuinamente quimiométrica que contiene los principales algoritmos empleados. Es independiente de librerías matemáticas comerciales y su eficiencia y eficacia es comparable (y en muchos casos mejor), que otras librerías análogas. Quimilap fue implementada en C, y utiliza como motor de cálculo matricial y algebraico a CLapack.

• Resultados y discusión

Se presenta un caso de estudio donde se muestran los resultados de la aplicación del software Quimiometrix a modo de validación de su uso en soluciones reales en el país.

Caso de estudio: Clasificación de combustibles derivados del petróleo por espectroscopia FT-MIR

La determinación del tipo de combustible, es una tarea a la que se enfrentan especialistas en diversos sectores del país, entre ellos la industria del petróleo y el frente forense-criminalista. El reto actual radica en cómo lograr una identificación rápida y certera de la mayor cantidad de combustibles posibles con métodos de análisis sencillos y la ayuda de técnicas quimiométricas para el procesamiento multivariante de los datos.

Varios métodos han sido reportados en la literatura que aborda esta temática. Los ensayos físico-químicos oficialmente establecidos como procedimientos de referencia por diferentes organizaciones internacionales, como por ejemplo: la ASTM (American Society for Testing Materials), la ISO (Internacional Standard Organization) y el IP (Institute of Petroleum) son muy utilizados para estos fines. Se debe destacar que las propiedades físico-químicas determinadas a través de los métodos de referencia emplean un tiempo de medición considerable y requieren de apreciable cantidad de muestra; además, cuando los productos presentan pequeñas diferencias o anomalías en su composición química, estas no son fácilmente detectables a través de los mismos.

De forma alternativa, la cromatografía de gases y la espectroscopía ultravioleta, también han sido utilizadas en los últimos años. A pesar de su reconocida sensibilidad y exactitud, la cromatografía gaseosa presenta desventajas relacionadas con el tiempo de preparación y ejecución de los análisis, por otra parte, la espectroscopía ultravioleta si bien es rápida y sencilla, no brinda toda la información espectral necesaria para poder lograr una diferenciación entre destilados medios del petróleo, así como entre la gasolina especial y regular /35-36/.

El uso de la espectroscopía infrarroja tanto media como cercana, combinada con la utilización de técnicas quimiométricas, se ha impuesto en los últimos años por su sencillez, rapidez y poder discriminativo de una amplia gama de tipos de derivados del petróleo /37-41/.

Entre los clasificadores más utilizados se encuentran el método SIMCA, PLS-DA y la novedosa incorporación de las Máquinas de Soporte Vectorial para clasificación /24, 26/.

Tomando en consideración lo antes expuesto, el objetivo de este caso de estudio estuvo dirigido a la obtención de un método de clasificación supervisada para la identificación de seis tipos de combustibles derivados del petróleo a partir del empleo de la espectroscopía en el infrarrojo medio (FT-MIR) y el empleo de técnicas quimiométricas para el procesamiento multivariante de los datos.

Preparación de los Conjuntos de calibración y validación. Técnicas quimiométricas empleadas.

Para el caso de estudio se cuenta con una data de entrenamiento de 64 muestras. Dichas muestras son representativas de seis tipos de combustibles destilados del petróleo provenientes de la Refinería de Petróleo "Ñico López" en Ciudad de La Habana: Gasolina Regular GR (1), Gasolina Especial GE (2), Diesel Regular DR (3), Nafta N (4), Turbo Combustible TC (5) y Kerosina K (6), el número entre paréntesis notifica el número de asignación de la clase patrón.

El conjunto de validación está compuesto por un total de dieciséis muestras representativas de las seis clases de patrones seleccionadas.

La adquisición de los espectros fue hecha con un Espectrofotómetro Infrarrojo Medio con Transformada de Fourier (FT-MIR); en transmitancia, en un Rango de 4 000 a 600 cm^{-1} y una resolución de 4 cm^{-1}. Se empleó el aditamento de Reflectancia Total Atenuada (ATR), marca PIKE con cristal de diamante KRS5.

Todas las muestras fueron sometidas a las transformaciones: Corrección de ATR, Corrección de Línea Base, Alisamiento y Eliminación del CO_2 con el software propietario del equipo Spectra Manager versión 2 (figura 1a). Ante la persistencia de la presencia de un efecto Offset se aplicó la corrección MSC (del inglés Multi Scatering Correction) con el software Quimiometrix (figura 1b). Se aplicó como preprocesamiento el centrado por la media para

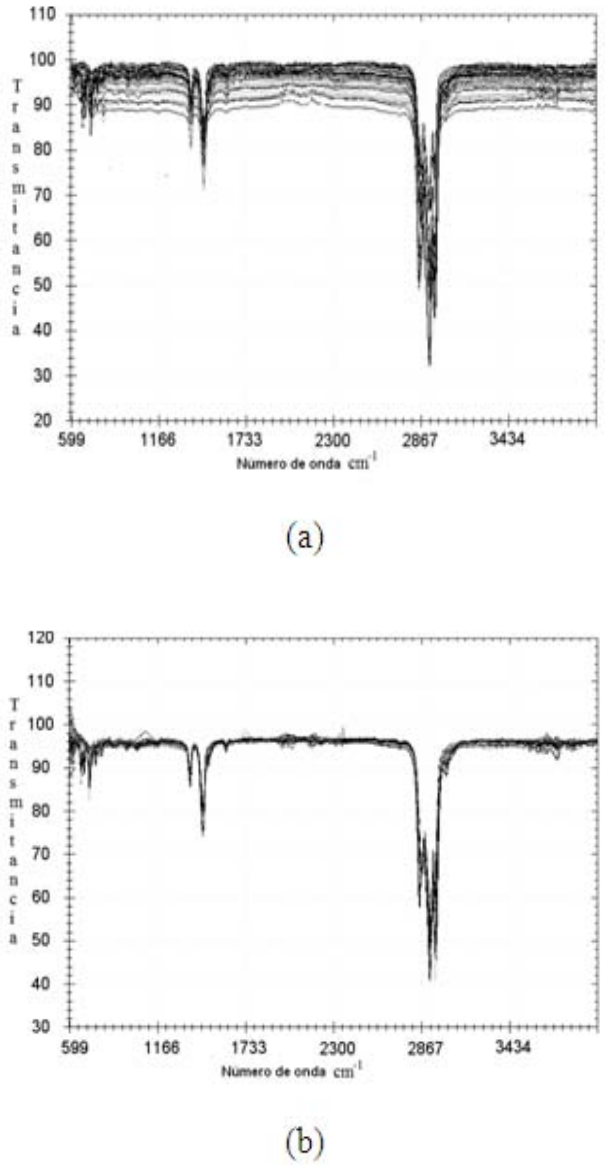

Fig. 1 Espectros de los datos de entrenamiento (a). Antes de realizarle ninguna transformación, obsérvese el corrimiento del Offset, (b) Después de realizarle los preprocesamientos y transformaciones, obsérvese la corrección del offset.

lograr un adecuado análisis exploratorio por PCA.

El análisis exploratorio PCA reveló que toda la información correspondiente a las variables analizadas, puede ser agrupada en solo siete nuevas variables (factores) con un 95,55 % de varianza acumulativa. Con vistas a mejorar aún más la modelación a partir de la eliminación de información redundante se procedió a estudiar la posibilidad de una reducción de variables utilizando el análisis del poder de modelación y de discriminación de cada una de las variables originales, posibilidad que ofrece Quimiometrix en el análisis exploratorio PCA, como se mostró en el caso de estudio I. Fueron eliminadas por su bajo poder de modelación y discriminación las variables entre los 4 000 y 3 200 cm^{-1} y desde los 2 700 hasta los 1 900 cm^{-1}.

Véase el gráfico de los "scores" en función de los dos primeros componentes que son los que acumulan el mayor por ciento de la varianza (figura 2). En el gráfico se muestra con claridad que el factor 1 está íntimamente relacionado con el tipo de combustible, mostrándose una discriminación en función de su peso, hacia la derecha los más ligeros y hacia la izquierda los más pesados. Sin embargo, la marcada similaridad entre las kerosinas y turbocombustibles no permite su discriminación por parte de los dos primeros componentes, es precisamente el factor 3 el que ostenta la información discriminante.

El análisis exploratorio por PCA realizado indica la factibilidad de la utilización de la espectroscopía infrarroja para extraer la información requerida con vistas a lograr una posible clasificación de este tipo de compuesto y con ello la predicción automática de muestras desconocidas de combustibles del petróleo en función de su tipicidad.

Terminado el análisis exploratorio de la data, están las condiciones creadas para proceder a crear y entrenar diferentes modelos de clasificación, con el objetivo de encontrar aquel que mejor se comporte ante la tarea en cuestión. Se evaluaron tres clasificadores SIMCA, PLS-DA y SVM y se comparó su comportamiento en función de los errores de clasificación cometidos por cada uno de ellos durante el proceso de validación de los modelos (figura 3).

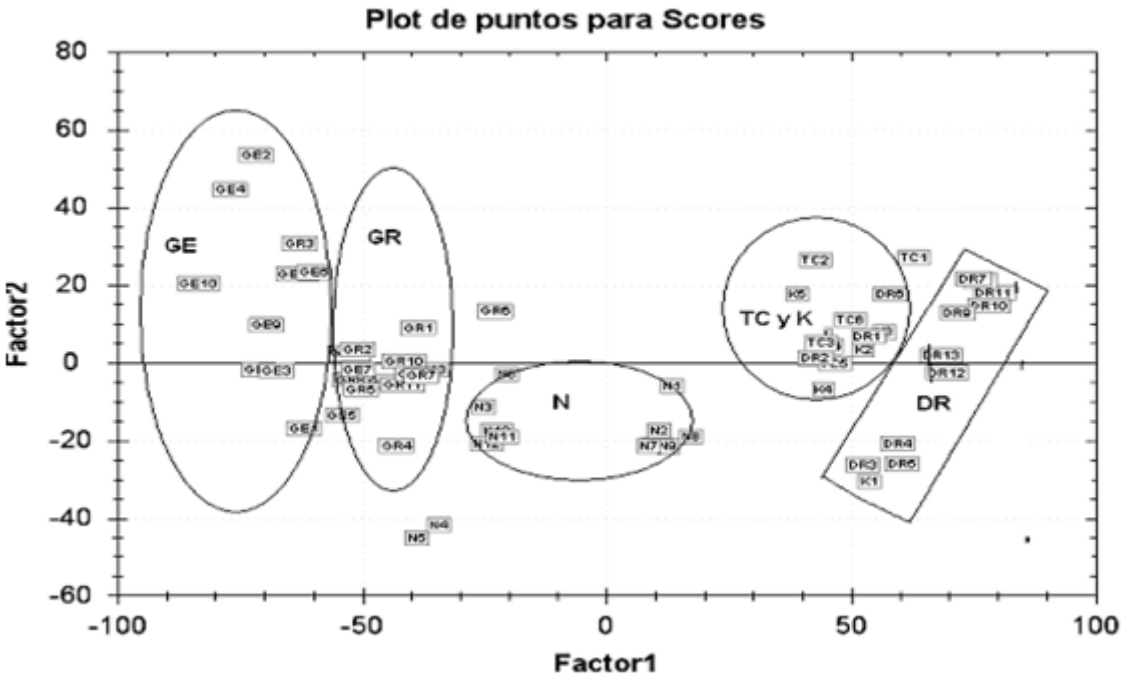

Fig. 2 Análisis por PCA realizado al conjunto de entrenamiento. Gráfico de los "Scores" en función de las componentes principales (factor 1 y factor 2).

SIMCA

			X						
			1	2	3	4	5	6	7
			PredC1	PredC2	PredC3	PredC4	PredC6	PredC5	No clasificadas
1	C1	☐	3	0	0	0	0	0	0
2	C2	☐	0	3	0	0	0	0	0
3	C3	☐	0	0	3	0	0	0	0
4	C4	☐	0	0	0	3	0	0	0
5	C6	☐	0	0	0	0	2	0	0
6	C5	☐	0	0	0	0	0	2	0

PLS-DA

			X						
			1	2	3	4	5	6	7
			PredC1	PredC2	PredC3	PredC4	PredC6	PredC5	No clasificadas
1	C1	☐	3	0	0	0	0	0	0
2	C2	☐	0	3	0	0	0	0	0
3	C3	☐	0	0	3	0	0	0	0
4	C4	☐	0	0	0	3	0	0	0
5	C6	☐	0	0	0	0	1	1	0
6	C5	☐	0	0	0	0	1	1	0

SVM

			X						
			1	2	3	4	5	6	7
			PredC1	PredC2	PredC3	PredC4	PredC6	PredC5	No clasificadas
1	C1	☐	3	0	0	0	0	0	0
2	C2	☐	0	3	0	0	0	0	0
3	C3	☐	0	0	3	0	0	0	0
4	C4	☐	0	0	0	3	0	0	0
5	C6	☐	0	0	0	0	2	0	0
6	C5	☐	0	0	0	0	0	2	0

Fig. 3 Resultado del proceso de validación de los modelos de clasificación.

Los resultados muestran que los clasificadores SVM y SIMCA fueron los de mejores resultados, ya que no presentan errores de clasificación, el PLS-DA no logra discriminar adecuadamente las muestras de kerosinas y turbocombustibles. Aunque ambos clasificadores dan buenos resultados, las ventajas que se atribuyen a las SVM en la literatura están relacionadas con: su bajo costo computacional en la validación, al estar sustentado el entrenamiento solo sobre parte de las muestras (vectores de soporte), su robustez ante el sobrentrenamiento, su capacidad de resolver casos no linealmente separables y su independencia de la dimensionalidad del sistema. Todo esto hace que la selección del modelo SVM resulte la más indicada.

Conclusiones

Como resultado del análisis y diseño de Quimiometrix v2.0 se obtuvo un software de calidad elaborado según los parámetros más modernos del desarrollo de software. Quimiometrix v2.0 implementa una arquitectura basada en "plugin" la cual permite la inclusión de nuevos algoritmos promoviendo el constante desarrollo de este proyecto.

Con Quimiometrix se explora la posibilidad de desarrollar mecanismos de control de errores, de actualización y de instalación automáticos y remotos basados en servicios webs, además en el diseño de software científico de rápido desarrollo, permitiendo a los especialistas centrarse en la implementación de sus algoritmos y dedicar menores esfuerzos a la concepción del sistema.

Quimiometrix, mediante el uso de Quimilap, alcanza tiempos de ejecución eficientes lo que lo hacen competitivo tanto en eficiencia como eficacia. Es un software competitivo con los de su tipo a nivel mundial.

Se pone de manifiesto a través del caso de estudio presentado, cómo el empleo de las técnicas quimiométricas automatizadas para procesamiento multivariante de datos mejoran y amplían la capacidad de respuesta de los investigadores en diversas formas, entre ellas: capacidad de mayor procesamiento y extracción de información útil de los datos; posibilidad de fácil acople con rápidas y modernas técnicas de análisis instrumental generadoras de datos de interés; mayor rapidez en la obtención de los resultados sin disminuir su veracidad y exactitud; capacidad de predicción para la identificación de sustancias desconocidas; capacidad de predicción cuantitativa de propiedades químico-físicas de sustancias.

Contar con un software autóctono que integre una gran parte de las mismas con funcionalidades de visualización y graficación es incuestionablemente una ventaja para investigadores, especialistas y técnicos cubanos relacionados con esta rama del saber, sin desestimar por supuesto otros de factura internacional a los cuales se pudiese acceder libremente.

Algunas de las funcionalidades del software nacional Quimiometrix se han presentado en este artículo como parte de su aplicación en la solución de problemas concretos, y que han sido el fruto del trabajo conjunto entre desarrolladores e investigadores para lograr la introducción en la práctica social de un producto de beneficio nacional.

Bibliografia

1. INFOMETRIX. *Infometrix Pirouette* [en línea]. 2012 [ref. de 3 julio 2012]. Disponible en Internet: <http://www.infometrix.com>.
2. CAMO. *The Unscrambler* [en línea]. 2012 [ref. de 3 julio 2012]. Disponible en Internet: <http://www.camo.com>.
3. EIGENVECTOR. *Eigenvector Research Incorporated* [en línea]. [ref. de 3 julio 2012]. Disponible en Internet: <http://www.eigenvector.com>.
4. NÚÑEZ, O., *et al*. "Nuevo sistema automatizado para el análisis de datos químicos y bioquímicos". En: *Memorias de la Convencion de Salud e Informática.* La Habana: 2009.
5. _____. "Quimiometrix. Nuevo sistema automatizado para el análisis de datos químicos". *Revista CENIC Ciencias Quimicas.* 2010, 42, 1, p. 13.
6. _____. "El uso de las técnicas quimiométricas en la solución de problemas prácticos". En: *Congreso Nacional de Reconocimiento de Patrones RECPAT*, 2009.
7. TALAVERA, I., *et al*. "Quimiometrix: Sistema de herramientas Quimiométricas para el Preprocesamiento, Clasificación y

Predicción de datas químicas espectrales". En: *Conferencia Internacional FIE*. Santiago de Cuba, 2008.

8. BRERETON, R. G. *Chemometric, Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons, Ltd, 2002.
9. ESBENSEN, K. H. *Multivariate Data Analysis In Practice*. 5th ed. Esbjerg: Alborg University, 2002.
10. FERREIRA, M. *Curso de Quimiometría* [en línea]. [ref. de 1 agosto 2012]. Diponible en Internet: <http://www.cenatav.co.cu>.
11. *Pirouette user Guide. Multivariate Data Analysis. Version 3.11 Infometrix, Inc* [en línea]. [ref. de 1 agosto 2012]. Diponible en Internet: <http://www.infometrix.com>.
12. *Manual de ayuda Quimiometrix* [en línea]. [ref. de 1 agosto 2010]. Diponible en Internet: <http://www.cenatav.co.cu>.
13. SAVITZKY, A.; GOLAY, E. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". *Anal. Chem*. 1964, 36, p. 1627-1639.
14. GORRY, A. "General Least-Squares Smoothing and Differentiation by the Convolution (Savitzky-Golay) Method". *Anal. Chem*. 1990, 62, p. 570-573.
15. ISAKSSON, T.; NAES, T. "The effect of multiplicative scatter correction (MSC) and linear improvement in NIR spectroscopy". *Applied Spectroscopy*. 1988, 42, p. 1273.
16. KORTUM, G. *Reflectance Spectroscopy*. New York, Springer, 1969.
17. HOTTELING, H. "Analysis of a Complex Statistical Variables into Principal Components". *J. Edu. Psychol*. 1933, 24, 417-424, p. 498-520.
18. JACKSON, J. E. *A User's Guide to Principal Components*. New York: J. Wiley & Sons, 1991.
19. WOLD, S.; SJÖSTRÖM, M. "Chemometrics: Theory and Application". En: *ACS Symposium Series*. B. R. KOWALSKI (Ed). 1977, 52, p. 243-282.
20. MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. *Multivariate Analysis*. London: A. Press, 1980.
21. HASTIE, T. T.; FRIEDMAN, J. Hierarchical clustering, in The Elements of Statistical Learning. New York: Springer, 2009.
22. DERDE, M. P.; MASSART, D. L. "Supervised Pattern Recognition: The Ideal method". *Anal. Chim. Acta*. 1986, 191, p. 1-16.
23. KOWALSKI, B. R.; BENDER, C. F. "The K-Nearest Neighbor Classification Rule (Pattern Recognition)". *Anal. Chim. Acta*. 1972, 44, p. 1405-1411.
24. WOLD, S.; SJÖSTRÖM, M. *SIMCA: A Method for Analying Chemical Data in Terms of Similarity and Analogy*. Research Group for Chemometrics, Institute of Chemistry, Umea' University, 1977.
25. VAPNIK, V. *The nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.
26. CHEN, N.; WENCONG, L.; JIE, Y.; GOZHENG, L. *Support Vector Machines in Chemistry*. World Scientific, 2004.
27. MARTENS, H.; NAES, T. *Multivariate Calibration*. New York: Wiley, 1989.
28. WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. "PLS-regression: a basic tool of chemometrics". *Chemometrics Intell. Lab. Syst*. 2001, 58, p. 109-130.
29. WOLD, S.; TRYGG, J.; BERGLUM, A.; ANTII, H. "Some recent development in PLS modelling". *Chemometrics Intell. Lab. Syst*. 2001, 58, p. 131-150.
30. HAALAND, D.; THOMAS, E. "Partial Least-Squares methods for spectral analyses. Relation to other quantitative calibration methods and the extraction of qualitative information". *Anal. Chem*. 1988, 60, 11, p. 1193-1202.
31. BRERETON, R. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Wiley & Sons, Ltd, 2003.
32. HERNÁNDEZ, N.; BISCAY, R. J.; TALAVERA, I. "A non Bayesian predictive approach for statistical calibration". *Journal of Statistical Computation and Simulation*. 2010.
33. HERNÁNDEZ, N., *et al*. "Support vector regression for functional data in multivariate calibration problems". *Analytica Chimica Acta*. 2009, vol. 642 (Issues 1–2), p. 6.
34. HERNÁNDEZ, N.; BISCAY, R. J.; VILLA, N.; TALAVERA, I. *Gaussian Inverse Regresión (GIR): A class of nonparametric functional regression estimators based on the estimation of the inverse model*. 2009.
35. BRUDZEWSKI, K; KESIK, A; KOLODZIEJCZYK, K; ZBOROWSKA, U; ULACZYK, J. "Gasoline quality prediction using gas chromatography and FT-IR spectroscopy: An artificial intelligence approach". *Fuel*. 2006, 85, 4, p. 553-558.
36. FLUMIGNAN, D. L.; TININIS, A. G.; FERREIRA, F. O.; DE OLIVEIRA, J. E. "Screening brazilian C gasoline quality: Application of the SIMCA chemometric method to gas chromatographic data". *Anal. Chim. Acta*. 2007, 595, 1-2, p. 128-135.
37. HIDAJAT, K.; CHONG, S. M. "Quality characterization of crude oils by partial least squares calibration of NIR spectral profile". *J. of Near Infrared Spect*. 2000, 8, 1, p. 53.
38. REBOUCAS, M. V.; BARROS, N. "Near infrared spectroscopic prediction of physical properties of aromatic hydrocarbon mixtures". *J. Near Infrared Spect*. 2001, 9, p. 263.
39. CANECA, A. R., *et al*. "Assessment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricanting oils". *Talanta*. 2006, 70, 2, p. 344-52.
40. PASADAKIS, N.; KARDAMAKIS, A. A. "Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis". *Anal. Chim. Acta*. 2006, 578, 2, p. 250-255.
41. BALABIN, R. M.; SAFIEVA, R. Z.; LOMAKINA, E. I. "Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction". *Chem. and Int. Lab. Syst*. 2007, 88, p. 183-188.