REGRESION KERNEL APLICADA A CALIBRACIÓN NIRS

Sebastián Vallejo

8 de diciembre de 2014

INTRODUCCION AL RECONOCIMIENTO DE PATRONES
Facultad de Ingeniería
Version 11
MLnirsProyectoSV v11.tex

Índice

1. Introducción

El término quimiometría [7] fué introducido por Wold y Kowalski a principios de los 70s. En forma similar los términos como biometría y econometría fueron introducidos en los campos de biología y la economía.

Según Wold [30] la Quimiometría (*Chemometrics*, en inglés) estudia cómo obtener información química relevante a partir de datos de mediciones, cómo se representa y despliega esa información, y cómo se convierte esa información en datos. Puede ser considerada como una subdisciplina que provee de teoría básica y metodología a la química analítica moderna.

Dos razones principales [7] han facilitado la evolución de la Quimiometría:

- Es posible adquirir grandes cantidades de datos a través de instrumentos químicos avanzados
- El incremento en el poder de cómputo amplió las capacidades en el procesamiento de señales y la interpretación de los datos químicos

En este contexto, similar al que se produce en otras disciplinas, un conjunto de técnicas basadas en aprendizaje supervisado se constituyen en una tecnología clave para extraer información y darle sentido al océano de bits que nos rodea [22].

En un problema de aprendizaje supervisado, el sistema de aprendizaje debe predecir las etiquetas de los patrones [2], donde las etiquetas pueden ser una clase o un número real.

Con estas herramientas (nuevas tecnologías y teoría) el laboratorio químico analítico se ha transformado. En particular, la aplicación de las técnicas NIRS compiten en reducción de costos y velocidad con muchas técnicas analíticas de la tradicional "química húmeda".

2. Descripción del Problema

El problema planteado consiste en obtener una función que permita predecir valores de un componente químico a partir de datos obtenidos a través de espectroscopía de infrarojo cercano (NIRS), como alternativa al análisis químico tradicional, lento y costoso.

2.1. Datos

Se cuenta con una muestra de 832 elementos, que fue dividida en una muestra de entrenamiento y una de test al azar por los expertos de campo que nos proponen el problema.

Cada elemento de la muestra cuenta con 2201 características que corresponden a las absorbacias, que es el logaritmo del inverso de las reflactancias [17] correspondientes a cada nivel de banda analizado.

Es decir x_i, el elemento i-ésimo de la muestra, es un vector $(x_{i,1}, ..., x_{i,2201})$ tal que $x_{i,j} = log\frac{1}{R_{i,j}}$, siendo $R_{i,j}$ la reflactancia en la banda j-ésima.

Esta transformación es conocida como transformación de absorbancia [31].

Cada elemento de la muestra está etiquetado de la forma:

{vector de características, valor químico)}.

El valor del componente químico específico es medido por métodos de análisis químico tradicional. En particular, utilizaremos el valor asociado al porcentaje de contenido de Nitrógeno Total.

2.2. Objetivo

Se desea un sistema capaz de predecir el Valor Químico en forma automática, en función de un nuevo vector de características obtenido por espectroscopía de infra-rojo cercano.

La capacidad de predicción del modelo se mide en la industria, según los expertos de campo, a través del coeficiente de determinación (R^2) entre los valores predichos y los medidos.

El valor objetivo del R^2, propuesto por los expertos de campo, es de 0,92.

2.3. Enfoque

Buscaremos entonces ajustar una función a la muestra de entrenamiento compuesta por un conjunto de datos de alta dimensionalidad obtenidos por NIRS y su correspondiente variable de respuesta obtenida mediante análisis químico.

La literatura relevada indica que se presentan relaciones no lineales entre los datos del espectro y las variables cuantitativas de interés.

Esa situación y dado que los datos presentan alta dimensionalidad, hace que resulten inadecuadas las técnicas tradicionales de regresión lineal.

Proponemos entonces obtener la función de predicción utilizando métodos de aprendizaje supervisado, en particular métodos kernel de regresión no lineal.

Para el tratamiento de los datos, se evaluará la aplicación de técnicas de reducción de la dimensionalidad de los datos del espectro, utilizando métodos de análisis funcional considerando que, por su naturaleza, los datos espectrales pueden ser considerados funcionales [1].

2.4. Alcance

Si bien el objetivo final es proporcionar un sistema de cálculo automático, para esta etapa plantearemos un modelo y su correspondiente función predictora que cumpla con los requisitos señalados.

3. Trabajos Previos

Se relevaron publicaciones relevantes en forma jerárquica, partiendo de aplicaciones generales de quimiometría y métodos de regresión aplicado a la química, para luego hacer foco en las técnicas de espectroscopía en general. Finalmente

revisamos distintos materiales de referencia aplicados específicamente a la espectroscopía de hojas de vegetales en general y en particular de hojas de tabaco.

El material estudiado nos permite afinar los objetivos de este trabajo, así como plantearnos líneas futuras de investigación.

En el Apéndice A de este documento, se presenta el relevamiento bibliográfico detallado.

A continuación mencionamos las referencias estudiadas que consideramos de mayor relevancia para este trabajo.

3.1. SVM y NIRS aplicado a quimiometría

- Support vectors machines and its applications in chemistry [18]
- A support vector machine-based analysis method with wavelet denoised near-infrared spectroscopy [19]
- Using data mining to model and interpret soil diffuse reflectance spectra [24]

3.2. NIRS aplicado a hojas de plantas

- Estimation of Nitrogen, Phosphorus, and Potassium Contents in the Leaves of Different Plants Using Laboratory-based Visible and Near-infrared Reflectance Spectroscopy: Comparison of Partial Least-square Regression and Support Vector Machine Regression Methods [31]
- A new approach to discriminate varieties of tobacco using vis/near infrared spectra [26]
- Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine [32]

4. Marco Teórico para modelado NIRS

4.1. NIRS

Dado que la materia orgánica controla el metabolismo de las plantas, es importante estimar y monitorear sus componentes bioquímicos.

Tradicionalmente el contenido de estos componentes puede ser estimado utilizando análisis análisis de laboratorio utilizando varias muestras. Si bien estas estimaciones son precisas, estos métodos son costosos, llevan tiempo, son destructivos y complejos [5].

La reflactancia espectral de las hojas de las plantas responde a las acciones combinadas de varios factores químicos y físicos de la planta, como por ejemplo el contenido interno de componentes bioquímicos, distribución y organización de las células y contenido de agua. Es entonces que la reflactancia visible y cercana

al infra-rojo tiene el potencial de permitir estimar el nitrógeno, el fósforo, el potasio u otros componentes bioquímicos [31].

La reflactancia cercana al infra-rojo ha probado ser una herramienta analítica poderosa. Ha sido utilizada ampliamente en las industrias agrícola, petroquímica, textil y farmacéutica. Específicamente, la aplicación de la espectroscopía de infra-rojo cercano para el análisis de muestras farmaceuticas se ha desarrollado en forma significativa en lo que va del siglo [32].

4.2. Regresión Lineal y métodos Sesgados

Dada una muestra de n elementos $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ buscamos encontrar una función f tal que $f(x) = \hat{y}$ sea un buen predictor de y para un input futuro x.

Para el caso de estudio un elemento (x_i, y_i) de la muestra consiste en:

- x_i es un vector de dimensión n que contiene los valores espectrales.
- y_i representa la variable cuantitativa de interés, el contenido de Nitrógeno de la muestra.

Para el tratamiento de este tipo de problemas, la literatura plantea el uso de métodos lineales y sus variaciones sesgadas, así como sus competidores modernos tales como regresión no paramétrica, redes neuronales y support vectors machines.

4.2.1. Métodos de Regresión lineal

"... la mayoría de los otros métodos de regresión son en realidad simplemente elaboraciones o modificaciones de la regresión lineal. Es casi imposible *entender*, como opuesto de *usar*, redes neuronales o support vector machines sin una buena comprensión de la metodología de la regresión lineal." Extraído del prefacio de [29]

La función de regresión y mínimos cuadrados ordinarios(OLS) El modelo de regresión lineal se basa en el supuesto de que la variable de respuesta y está relacionada linealmente con los datos de entrada x.

Se busca una función de la forma $f(x) =< w, x >= \sum_{i=1}^{n} w_j x_j$ que mejor interpole el set de entrenamiento.

Si la muestra tiene n elementos ($m = n$), obtener w se reduce a resolver el sistema de ecuaciones $Xw = Y$. El caso en que $m > n$ es el problema estudiado por Gauss, conocido como aproximación de mínimos cuadrados, que implica resolver el sistema $X'Xw = X'Y$.

Inestabilidad de la solución OLS Cuando la matriz X no es de rango completo, $X'X$ será singular, y por lo tanto los estimadores OLS de w no serán únicos. La singularidad puede producirse cuando la matriz X tiene columnas colineales o cuando hay mas variables que observaciones. [16].

4.2.2. Métodos de regresión lineal sesgados

Una forma de solucionar el problema anterior, es abandonar el requerimiento de que el estimador de w sea insesgado. Existen diversos estimadores sesgados que son superiores al OLS del punto de vista del MSE cuando se presentan estos problemas [16].

Los métodos de regresión sesgados fueron utilizados originalmente en quimiometría (aplicados a la investigación de alimentos, estudios de contaminación ambiental, etc.) donde resulta habitual que el número de variables (dimesión del vector x_i) sea superior a la cantidad de observaciones ($d > n$).

Se han desarrollado entonces métodos sesgados como la Regresión de Componentes Principales, Regresión de Mínimos Cuadrados Parciales y la Regresión Ridge.

Partial Least Squared Regression

La regresión con mínimos cuadrados parciales (PLSR) ha sido utilizada tradicionalmente en química como método de calibración multivariada y en particular es un algoritmo ampliamente usado para modelar datos NIRS. Ofrece resultados aceptables en la mayoría de los casos cuando existe una relación lineal entre el espectro y la propiedad a determinar.

Regresión Ridge

Otro camino frecuentemente elegido es el de restringir de alguna forma la selección de funciones. Esta restricción o sesgo es conocida como regularización. [27]

El problema se puede expresar como el sistema $(X'X + \lambda I_n)w = X'y$.

Si bien es una técnica originalmente utilizada para modelos lineales, incorporaremos la regularización en los métodos Kernel más abajo mencionados.

4.3. Métodos no Lineales: Métodos Kernel

La literatura más reciente indica que la técnica lineal PLSR no es capaz de obtener resultados precisos el modelado NIRS, y que tampoco supera el reto de generar una función predictiva para una muestra limitada. [15]. En este trabajo, verificaremos que dado que la muestra tratada es amplia, se pueden obtener resultados primarios satisfactorios con PLSR.

Las técnicas basadas en Kernels [3] son uno de los desarrollos más importantes dentro de los algoritmos de aprendizaje automático.

Las representaciones Kernel ofrecen una solución alternativa proyectando los datos en un nuevo espacio de características F de alta dimensión para para mejorar la capacidad de las máquinas lineales.

Definimos el mapeo no lineal como:

$$\phi(x) : x \epsilon \mathbb{R}^n \rightarrow \phi(x) \epsilon F \subseteq \mathbb{R}^N$$

El uso de procedimientos lineales en la representación dual hace posible utilizar el paso anterior en forma implícita. Veremos una demostración específica de este resultado para regresión SVM.

Esto es posible dado que los mapeos aparecen siempre en la forma de productos internos entre mapeos de pares de elementos de la muestra. Reemplazando el producto interno por la función kernel apropiada, podemos entonces realizar en forma implícita el mapeo no lineal sin incrementar la cantidad de parámetros a ajustar [8].

$$K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$$

4.3.1. Mínimos Cuadrados Regularizados Kernel (KRLS)

Utilizamos la función de pérdida $V()$ para cuantificar el costo o pérdida de utilizar la función $f(x)$ para predecir los valores de y.

Incorporando un componente de regularización (como en la Regresión Ridge) el problema consiste en encontrar la función que minimiza:

$$\frac{1}{n}\sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda \left\| f \right\|_k^2$$

La utilización de Kernels nos permite, aplicando el teorema de representación [25] convertir el problema en ajustar la función:

$$f(x) = \sum_{i=1}^{m} c_i K(x, x_i)$$

Para la función de pérdida cuadrática [9] el objetivo es encontrar la función que minimiza:

$$\frac{1}{n}\sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 + \lambda \left\| f \right\|_K^2$$

Podemos incluir n en λ, obteniendo:

$$\sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 + \lambda \left\| f \right\|_K^2$$

Utilizando la solución obtenida por el teorema de representación, en notación matricial podemos expresar la función a minimizar como:

$$\left\| Y - Kc \right\|^2 + \lambda \left\| f \right\|_K^2$$

Desarrollamos el segundo sumando:

$$\begin{aligned}
\left\| f \right\|_K^2 &= < f, f >_K =< \sum_{i=1}^{n} c_i k\left(x_i, .\right), \sum_{i=1}^{n} c_i k\left(x_i, .\right) >_K \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j < k\left(x_i, .\right), k\left(x_i, .\right) >_K = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k\left(x_i x_j\right) \\
&= c^t K c
\end{aligned}$$

Oteniendose la siguiente expresion a minimizar:

$$\|Y - Kc\|^2 + \lambda c^t K c$$

Determinacion de c

La expresion anterior es convexa en c, podemos entonces encontrar su mínimo igualando su gradiente a 0:

$$\begin{aligned} -K(Y - Kc) + \lambda K c &= 0 \\ (K + \lambda I)\, c &= Y \\ c &= (K + \lambda I)^{-1} Y \end{aligned}$$

Podemos encontrar entonces c resolviendo un sistema lineal de ecuaciones.

Para cada $\lambda > 0$ la solución existe y es única pues la matriz $(K + \lambda I)$ es simétrica y definida positiva, dado que K lo era,

Necesitamos resolver el sistema lineal:

$$(K + \lambda I)\, c = Y$$

Obtención de c para distintos λ

Realizando la descomposicion en vectores y valores propios de la matriz K:

$$K = Q \Lambda Q^t \text{ con } QQ^t = I$$

Luego:

$$\begin{aligned} (K + \lambda I) &= \left(Q \Lambda Q^t + \lambda I\right) = \left(\Lambda Q^t + Q \lambda I Q^t\right) \\ &= Q\,(\Lambda + \lambda I)\, Q^t \end{aligned}$$

Por lo cual $(K + \lambda I)^{-1} = Q\,(\Lambda + \lambda I)^{-1}\, Q^t$

Una vez calculados Q y Λ, podemos encontrar el c correspondiente a ese λ :

$$c\,(\lambda) = Q\,(\Lambda + \lambda I)^{-1}\, Q^t Y$$

Podemos entonces obtener los c para distintos λ reduciendo al mínimo los cálculos.

Validación cruzada: Leave-One-Out

Para cada punto x_i se obtiene una funcion usando los $n-1$ puntos restantes, y luego se mide el error en x_i.

Problema: Debemos construir n funciones predictores diferentes, en conjuntos de datos de tamaño $n - 1$.

Es posible acelerar este proceso, dado que se puede obtener una expresión analítica del error de validación cruzada[23].

$$LOOE = \frac{c}{diag_v(G^{-1})}$$

Definiendo $G(\lambda) = K + \lambda I$.

Selección del parámetro del Kernel

Se deberá encontrar el óptimo del parámetro λ conjuntamente con la determinación de los parámetros de la definición de la función Kernel. Por ejemplo, para el Kernel Gaussiano es necesario determinar σ:

$$K(u, v) = e^{\|u-v\|^2/-2\sigma^2}$$

Habiendo obtenido una forma eficiente de calcular la solución para distintos λ podremos recorrer una grilla de valores de σ y para cada caso obtener el λ óptimo.

Observación: En la formulación del Kernel Gaussiano para SVM el parámetro utilizado es otro: $\gamma = \frac{1}{2\sigma^2}$.

4.3.2. Support Vector Machines

Las características especiales y el excelente desempeño empírico de las SVMs en el campo de la química se demuestra en destacadas publicaciones científicas de los últimos años [18].

Los modelos SVM fueron originalmente definidos para la clasificación de clases de objetos separables linealmente. Para utilizar los SVM cuando no se presenta la linealidad, las coordenadas de los objetos son mapeadas en un nuevo espacio. En este nuevo espacio de dimension alta, las dos clases pueden ser separadas linealmente.

SVM fue extendido por Vapnik para la regresión [20] y se convierte en uno de los métodos aplicados en las publicaciones científicas más recientes sobre modelado NIRS [18] y específicamente para el caso de estudio [32].

SVM para regresión

Definimos la función de Pérdida ε-insensible:

$$L(y - f(x)) = \begin{cases} |y - f(x)| - \varepsilon & \text{si } |y - f(x)| \geq 0 \\ 0 & \text{en otros casos} \end{cases}$$

Es decir, solo los puntos fuera de la banda ε causan pérdida.

El algoritmo de SVM para regresión busca solucionar el problema de optimización minimizando:

$$\frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n} L(y_i - f(x_i), \varepsilon)$$

donde C es un parámetro de regularización.

Este problema puede expresarse introduciendo la variable $\xi_i^{(*)}$ como:

$$\begin{aligned} \text{Minimizar:} \quad & L(w, b, \xi_i^{(*)}) = \tfrac{1}{2}||w||^2 + \tfrac{C}{n}\textstyle\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{Sujeto a:} \quad & (w^t x_i + b) - y_i \leq \varepsilon + \xi_i, i = 1..n \\ & y_i - (w^t x_i + b) \leq \varepsilon + \xi_i^*, i = 1..n \\ & \xi_i^{(*)} \geq 0, i = 1..n \end{aligned}$$

La función de Lagrange para este problema de optimización resulta en:

$$\begin{aligned} L(w,b,\alpha,\xi_i^{(*)}) &= \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}(\xi_i+\xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i+\eta_i^*\xi_i^*) \\ &- \sum_{i=1}^{n}\alpha_i(\varepsilon+\xi_i+y_i-w^tx_i-b) - \sum_{i=1}^{n}\alpha_i^*(\varepsilon+\xi_i^*-y_i+w^tx_i+b) \end{aligned}$$

Derivando e igualando a cero obtenemos:

$$\begin{aligned} w - \sum_{i}^{n}(\alpha_i^*-\alpha_i)x_i = 0 \\ \sum_{i}^{n}(\alpha_i^*-\alpha_i) = 0 \\ \frac{C}{n} - \alpha_i^* - \eta_i^* = 0 \end{aligned}$$

Sustituyendo esto en la ecuación principal, obtenemos la forma dual del problema de optimización:

$$\begin{aligned} \text{Minimizar:} \quad & \tfrac{1}{2}\textstyle\sum i,j = 1^n(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(x_i^tx_j) + \varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i) - \sum_{i=1}^{n} y_i(\alpha_i^*-\alpha_i) \\ \text{Sujeto a:} \quad & \textstyle\sum_{i=1}^{n}(\alpha_i-\alpha_i^*) = 0 \\ & 0 \le \alpha_i, \alpha_i^* \le \frac{C}{n}, i = 1..n \end{aligned}$$

Resolviendo el problema dual obtenemos:

$$f(x) = \sum_{i=1}^{n}(\alpha_{i,f}^*-\alpha_{i,f})(x_i^tx) + b_f$$

Las características x_i aparecen siempre en el problema dual como productos internos, por lo tanto, las funciones $\phi(x)$ siempre aparecen como productos internos, por lo tanto nunca las tendremos que calcular. Los productos internos los calculamos utilizando $K(x_i,x_j) = \phi(x_i)^t\phi(x_j)$.

4.4. Métodos de extracción de características para datos fucionales

4.4.1. Wavelets

La transformación wavelet (WT) es un método matemático basado en la transformada de Fourier, que puede ser usado en la compresión de datos, filtros para manejar el suavizado y el ruido, validación baseline y análisis de multicomponentes superpuestos de señales [26].

El análisis Wavelet genera una estimación del contenido de una frecuencia local de una señal representando los datos utilizando una familia de funciones wavelet que varían en escala y posición.

La señal puede ser reconstruida con precisión con una cantidad relativamente pequeñ a de componentes [32].

WT puede ser entonces utilizado para comprimir los datos espectrales obtenidos en NIRS.

4.4.2. Transformación Wavelet Discreta (DWT)

La transformación wavelet discreta está basada en un filtro low pass H y un filtro high pass G y una decimaciónn binaria [19].

La DWT de un vector de datos puede ser calculada rápidamente usando un banco de filtros. La estructura básica del banco de filtros comprende un par de filtros H/G seguidos por una operación de muestreo hacia abajo (down-sampling) que consiste en descartar el resto de los puntos de la salida de los filtros.

Los filtros se seleccionan de forma tal que la transformación sea invertible, preservando entonces la información de la señal.

La salida de los filtros del canal low-pass pueden ser descompuestos a su vez por sucesivos pares de filtros hasta cierto número N_{it} de iteraciones.

4.4.3. DWT y FFT

Las plantillas utilizadas en FFT son ondas de senos y cosenos con diferentes frecuencias. De esta forma las técnicas FFT nos pueden decir fácilmente la información de frecuencias global de una señal. Pero [7], en algunos casos lo que se desea es encontrar es algunos picos espectrales correspondientes a ciertos químicos en el análisis espectral. También se necesita determinar las frecuencias locales, lo que no puede ser realizado sencillamente con la información extraída utilizando la tranformada de Fourier. Es entonces cuando se hacen necesarias las técnicas wavelet.

Estas técnicas son aplicadas éxito en [26] y [32] frente a datos similares a los de nuestro problema.

5. Estadística Descriptiva

5.1. Valor químico

El valor químico disponible es el porcentaje de contenido de Nitrógeno total.

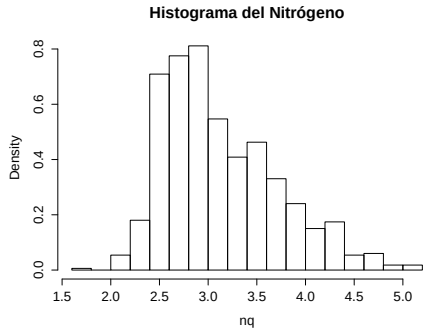

Figura 1: Histograma de los valores de N , $Min = 1{,}7$ $Media = 3{,}1$ $Max = 5{,}1$

5.2. Espectros y su relación con el Nitrógeno

Primero intentamos sin éxito visualizar alguna relación entre los espectros y el nivel de nitrógeno medido. Para ello coloreamos los espectros según su valor de Nitrógeno correspondiente.

Son 832 líneas, en las que se superponen colores, aparentemente no observamos ninguna relación.

Luego procedimos a ordenar los espectros según su etiqueta de valor químico y los graficamos como una superficie. A simple vista no fuimos capaces de encontrar relación alguna.

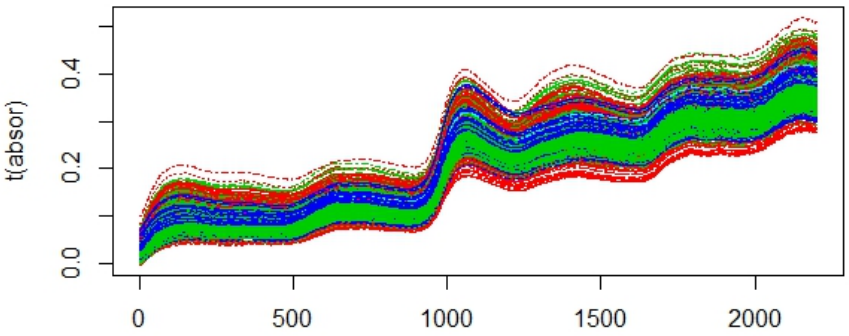

Figura 2: Gráfico de los espectros coloreados según valor químico

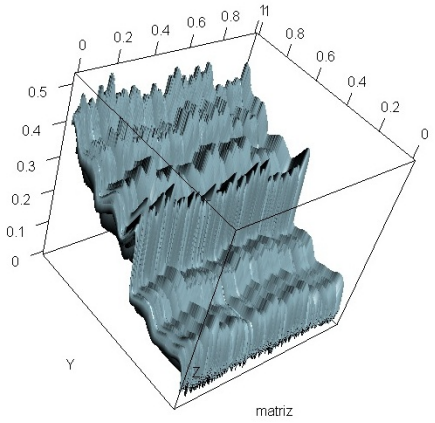

Figura 3: Superficie de los espectros

5.3. Correlación de los componentes del espectro

Calculamos el coeficiente de correlación de cada variable predictora contra la variable de respuesta.

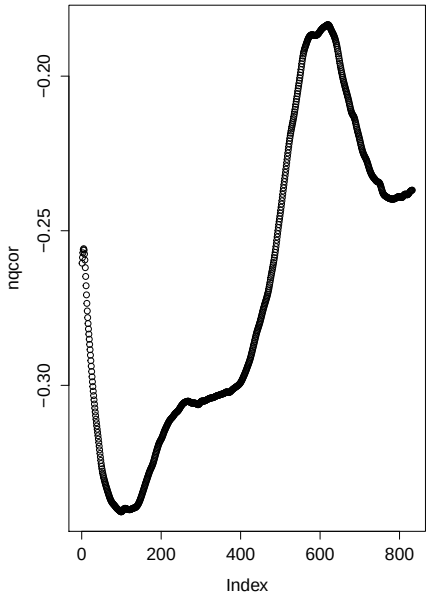

Figura 4: Correlación de las variables predictoras con el valor químico

5.4. Outliers

Outliers funcionales Utilizaremos el concepto de *profundidad estadística* para datos funcionales [10]. Estas medidas de profundidad son útiles para definir medias de posición y dispersión, para clasificación y para detección de outliers. A través del paquete fda.usc [12] de del software R es posible detectar valores atípicos a partir de la profundidad estadística.

En particular, utilizamos una nueva herramienta visual denominada Outliergram [4]

El output generado es el siguiente:

Los outliers de forma detectados son los siguientes: 101 140 159 158 139 71 115 735 116 147 824 138 143 160 287 823 635 130 229 72 289 829 477 141 230 634 104 632 117 642 100 388 633 137 641 640 134 636 133 99 119 285 663 132 630 290 150 157 687 661 219 220 112 221 3 664

Los outliers de magnitud detectados son los siguientes: 554 623 624

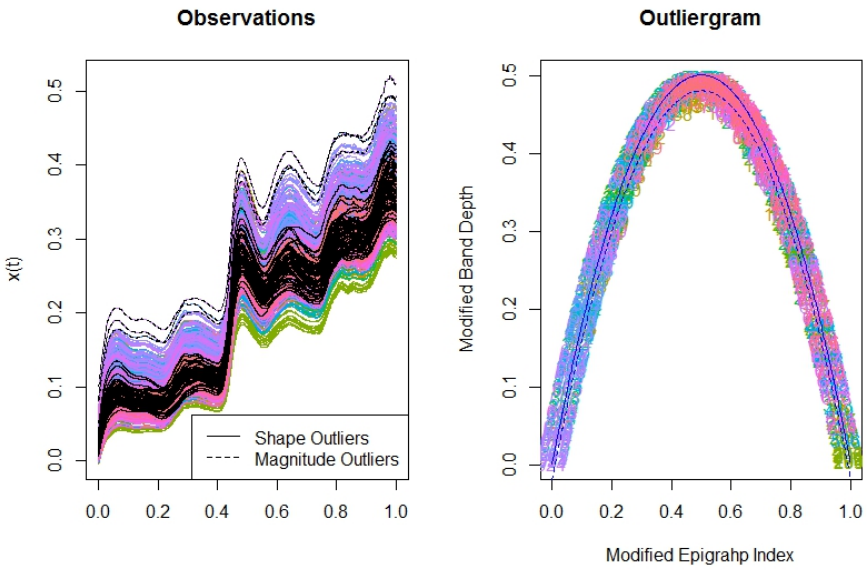

Figura 5: Salida del Outliergram

Outliers de datos expresados como coeficientes wavelet

Se utilizan algoritmos aplicables a la detección de outliers en datos multivariados.

En particular utilizamos el algoritmo LOF [6] y [28].

Los primeros outliers detectados con este criterio son los siguientes elementos de la muestra: 665 666 164 556 72.

6. Metodología

6.1. Pre-procesamiento

Las variables predictoras están muy correlacionadas (se puede decir que son datos funcionales). En este contexto se hará necesaria la utilización de prácticas de reducción de la dimensionalidad aplicada a datos funcionales.

Como problema adicional, en los datos originales, tenemos más dimensión de datos que tamaño de la muestra.

Tomamos dos caminos para tratar estos problema:

- Utilizar métodos sesgados aplicables cuando $n < d$, que son estándares en la literatura NIRS, que adicionalmente reducen el ruido.
- Reducción de la dimensionalidad funcional

Por las características de los aparatos de captura de datos, es razonable considerar que los datos tienen ruido. Esto puede ser observado si se mira en detalle las curvas de los espectros originales.

PLS al seleccionar componentes, reduce ruido y además reduce cantidad de variables.

Como alternativas para reducir dimensionalidad utilizaremos la transformación discreta wavelet (DWT).

Para mejorar los resultados de los métodos Kernel, realizaremos reducción del ruido aplicando técnicas de suavizado wavelet.

6.2. Evaluación y Selección de Modelos

A continuación detallamos las metodología para evaluar y seleccionar los modelos. Buscamos un correcto ajuste de la función predictora, pero sin caer en el sobreajuste.

Selección de modelo sobre datos de entrenamiento Sobre la muestra de entrenamiento evaluaremos los distintos modelos, a través de la validación cruzada de 10 folds. Usaremos el indicador MSE CV10, es decir el error cuadrático medio que resulta del promedio de las 10 evaluaciones realizadas. El modelo que tenga el menor indicador, es seleccionado para pasar a la siguiente etapa de evaluación, estableciéndose entonces los parámetros del modelo.

Modelo	Parámetros a determinar
PLSR	cantidad de componentes
SVM R	parámetro de regularización C, parámetro del kernel γ
KRLS	parámetro de regularización λ, parámetro del kernel σ

Evaluación sobre datos de entrenamiento Una vez definido el modelo en la etapa anterior, se evalúa el desempeño del mismo calculando el MSE sobre el total de la muestra de entrenamiento y el R^2 entre los valores predichos y los medidos.

Para ello ajustan los valores predichos contra los valores medidos con una regresión lineal simple, para obtener los indicadores necesarios.

Procedimiento de evaluación:

- Se utiliza el modelo seleccionado para predecir los valores correspondientes para todos los elementos de la muestra de entrenamiento.
- Se ajusta un modelo lineal entre los valores predichos y los medidos
- Se calculan MSE y R^2.

Evaluación sobre datos de Test Se procede en forma similar que en el modelo evaluado sobre datos de entrenamiento. Se realizan las predicciones con los datos de Test.

7. Experimentos realizados

A continuación se procede a evaluar en la práctica el poder predictivo de los distintos modelos estudiados.

7.1. Extracción supervisada de características con PLSR

Es un método supervisado pues, en forma similar a LDA (utilizado en problemas de clasificación [11]), para la generar los componentes se utilizan las variables de respuesta. No solo se proyectan los predictores (como en PCA) sino que también las variables de respuesta.

Extracción de características La muestra es proyectada en ese nuevo espacio, obteniendo una nueva muestra con la misma cantidad de características. Podemos ordenar las nuevas características según su puntaje.

El procedimiento consiste en evaluar modelos lineales con mínimos cuadrados ordinarios para distintos conjuntos de componentes. Los conjuntos de componentes son seleccionados en forma incremental uno a uno según su puntaje descendente.

Encontramos que el menor error se encuentra para el conjunto de las 12 primeras características. Este error es evaluado según el criterio 10 fold cross validation.

En la figura 7 vemos que se minimiza el error para el modelo de 12 componentes.

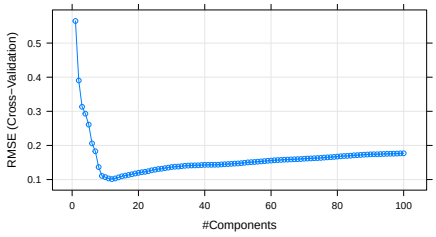

Figura 6: 12 componentes, MSECVentr=0.010201

Seleccionamos entonces el modelo de 12 componentes.

Procedemos a realizar una predicción con el modelo seleccionado de 12 componentes con los datos de entrenamiento. Evaluamos la predicción realizando la regresión lineal simple entre los valores predichos y los medidos. El error de mínimos cuadrados es de 0,0086. Es levemente menor, pues el error anterior de la selección era resultado del promedio de los 10 modelos evaluados. El R^2 ajustado obtenido es de 0,9748.

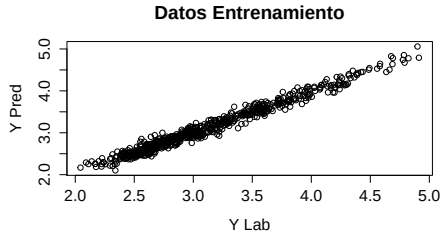

Figura 7: Datos entrenamiento, $MSE = 0{,}0086$ $R^2 = 0{,}9748$

Realizamos la predicción con el modelo seleccionado para los datos de test. Evaluamos en forma similar. Obtenemos un error levemente mayor, con un $R^2 = 0{,}9624$ que nos permite estimar la capacidad predictiva del modelo (no hubo sobreajuste).

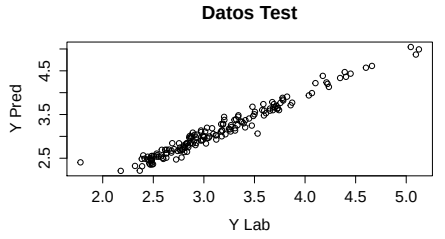

Figura 8: Datos test, $MSE = 0{,}0137$ $R^2 = 0{,}9624$

El modelo PLS funciona en forma satisfactoria superando el 0.92 del objetivo. Esto se debe a que, según la bibliografía, el PLS no funciona tan bien con muestras pequeñas (50 elementos) pero si con una muestra grande como es nuestro caso (más de 600 elementos).

7.2. Regresión SVM de los datos originales del espectro

Procedemos a ajustar un modelo SVM con kernel Gaussiano. Por lo tanto debemos elegir el parámetro de regularización C y el γ del kernel, que minimicen el error cuadrático medio.

Selección de los parámetros del modelo Para ello realizamos una búsqueda en una grilla de valores. Primero con una grilla que abarca un rango grande, para luego ir reduciendo el tamaño de la misma.

Luego de varios intentos, con una grilla evaluada entre valores de C entre 140 y 250 y γ entre 0.05 y 0.08 obtenemos:

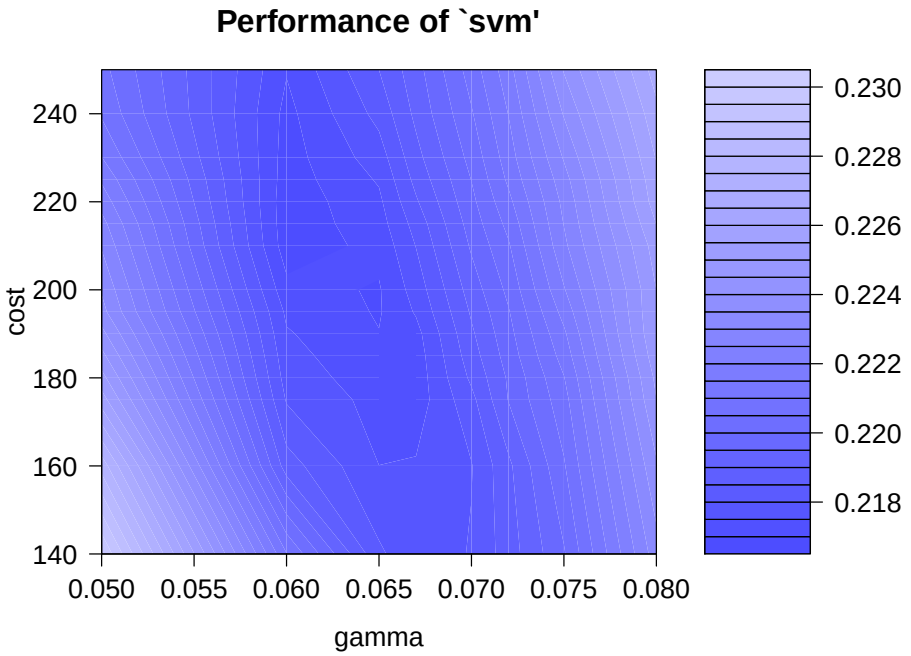

Figura 9: Gridsearch:parámetros $\gamma = 0{,}06$ y $C = 220$, $MSECVentr = 0{,}2166$

Selección del modelo Evaluamos en una grilla más fina en el entorno de óptimo encontrado en la etapa anterior.

Entonces, evaluamos una grilla final de para valores:
$C \in \{210, 215, 217, 218, 219, 220, 221, 223, 225, 226, 227, 228\}$ y
$\gamma \in \{0{,}055, 0{,}057, 0{,}059, 0{,}06, 0{,}062, 0{,}065\}$

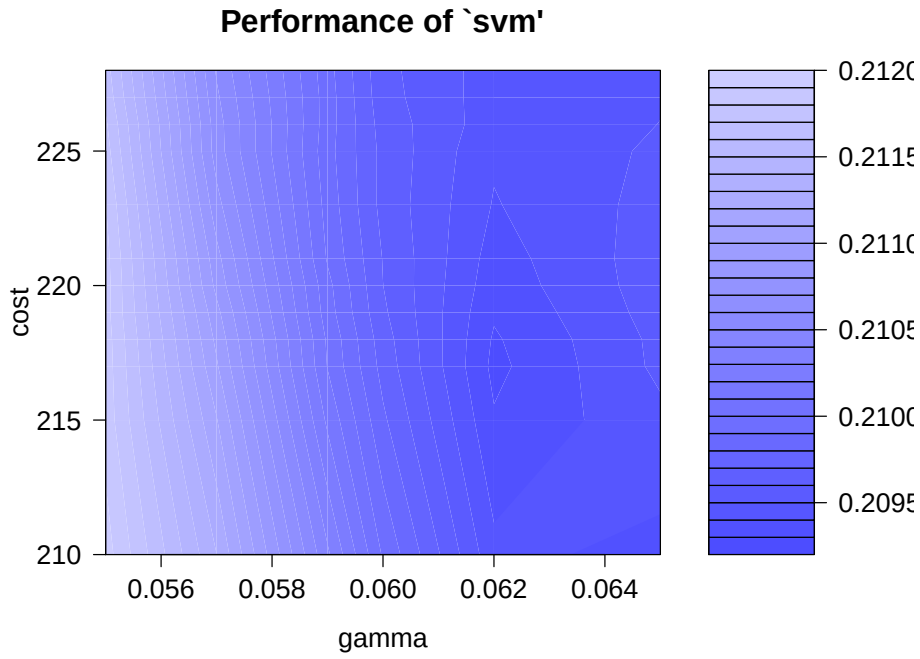

Figura 10: Gridsearch:parámetros $\gamma = 0{,}062$ y $C = 217$, $MSECVentr = 0{,}2092$

Los parámetros seleccionados son $\gamma = 0{,}062$ y $C = 217$. Se mejora levemente el MSE de validación cruzada de 0,2155 a 0,2092.

Realizamos predicción con los datos de entrenamiento, para evaluar el modelo hacemos una regresión de los valores predichos contra los valores medidos.

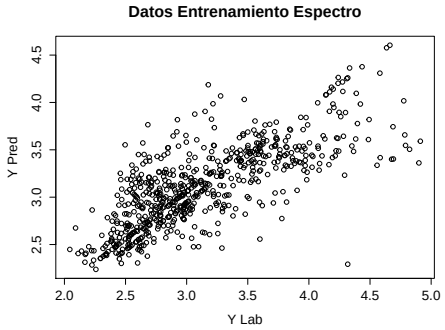

Figura 11: Datos Entrenamiento $MSE = 0{,}1493$ $R^2 = 0{,}5652$

El R^2 no cumple con el objetivo.

Ensayamos la predicción datos de test, aunque ya con los datos de entrenamiento no dió resultados satisfactorios.

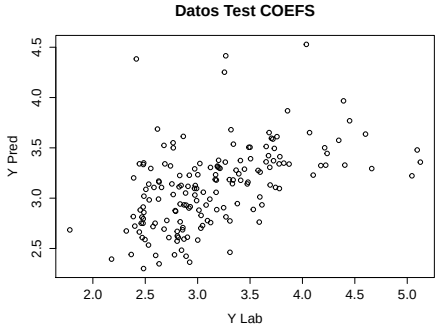

Figura 12: Datos Test $MSE = 0{,}2805$ $R^2 = 0{,}2473$

Los resultados están lejos de ser satisfactorios.

7.3. Regresión SVM, de coeficientes wavelet de espectros originales

Procedemos a reducir la dimensionalidad de los datos del espectro.

Las técnicas con mejores resultados en este campo son los coeficientes wavelet. En particular, utilizamos un wavelet HAAR nivel 5

Utilizamos entonces, como variables predictoras a los 69 coeficientes obtenidos.

Selección del modelo Buscamos en una grilla de valores, la que vamos afinando paso a paso. En la última grilla evaluada obtenemos los siguientes resultados.

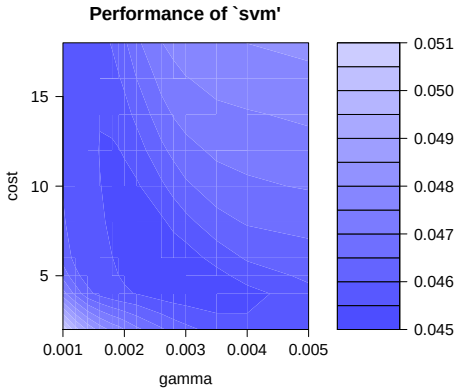

Figura 13: Gridsearch:parámetros $\gamma = 0{,}003$ y $C = 5$, $MSECVentr = 0{,}0448$

El menor valor del error obtenido es para los parámtros $\gamma = 0{,}003$ y $C = 5$. El error calculado es mucho menor que cuando tratamos los datos de los espectros originales.

Evaluación del modelo con los datos de entrenamiento.

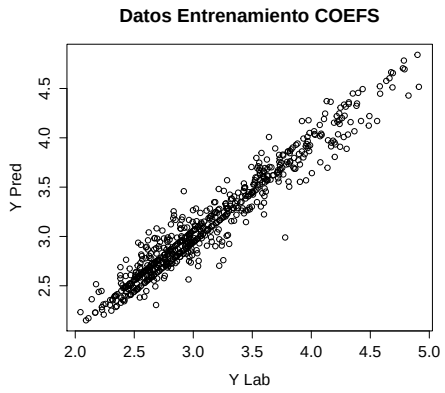

Figura 14: Datos Entrenamiento $MSE = 0{,}0206$ $R^2 = 0{,}9408$

Evaluación del modelo con los datos de test.

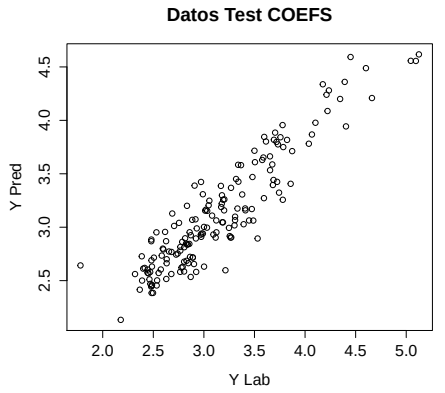

Figura 15: Datos Test $MSE = 0{,}0553$ $R^2 = 0{,}8484$

Se ha logrado obtener un modelo con mejores resultados que con los datos de los espectros originales. No alcanzamos el R^2 objetivo con los datos de test (0,84), pero si con los datos de entrenamiento (0,94).

7.4. KRLS, de coeficientes wavelet de espectros originales

Hainmuller y Hazlett del Departamento de Ciencias Políticas del M.I.T. proponen en [13] la utilización de regresion con minimos cuadrados regularizados kernel en las ciencias sociales. Los autores han desarrollado un paquete de R: KRLS [14].

Este paquete nos permite, para un σ dado obtener el λ óptimo correspondiente.

Por lo tanto recorreremos una grilla de valores de σ.

Para cada modelo resultante de una combinación de σ y su λ óptimo evaluamos, en forma similar a lo ya realizado con SVM, en la muestra de entrenamiento y test.

El R^2 para los datos de test se maximiza para $\sigma = 71$

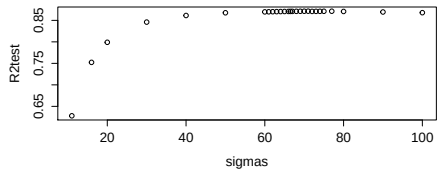

Figura 16: Evaluación del R^2 para los datos de test en la grila de σ

Para KRLS incluimos en la iteración el cálculo de R^2 para los datos de entrenamiento y de test.

Un subset de la tabla de datos obtenidos:

```
      sigmas    lambdas    R2func    R2entr    R2test
[17,]   69.0 0.11727456 0.9901312 0.9912999 0.8711333
[18,]   70.0 0.11716759 0.9898935 0.9910678 0.8712389
[19,]   71.0 0.11706063 0.9896534 0.9908333 0.8713369*
[20,]   72.0 0.16694832 0.9835318 0.9853243 0.8708421
[21,]   73.0 0.16684136 0.9832031 0.9849975 0.8709383
```

Los resultados son mejores que los de SVM. Alcanzamos el objetivo para los datos de entrenamiento (0,99), pero no para los datos de test (0,87)

7.5. Regresión SVM, de coeficientes wavelet de espectros suavizados

Dado que SVM es sensible al ruido, procedemos a utilizar distintas técnicas del extracción del ruido.

En este caso, aplicamos wavelet denoising a los datos del espectro, con un wavelet DB8 (Daubechies de vanishing-monent 8).

Reprocesamos luego los espectros suavizados, disminuyendo la dimensionalidad a través de los coeficientes wavelet en forma similar a lo realizado con los espectros originales.

Proseguimos luego con el procedimiento similar al realizado para los espectros originales.

Busqueda en grilla de valores:

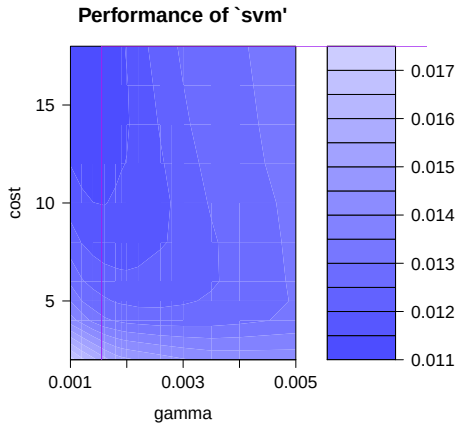

Figura 17: Gridsearch:parámetros $\gamma =$ y $C =$, $MSECVentr = 0{,}011$

Evaluación del modelo con los datos de entrenamiento.

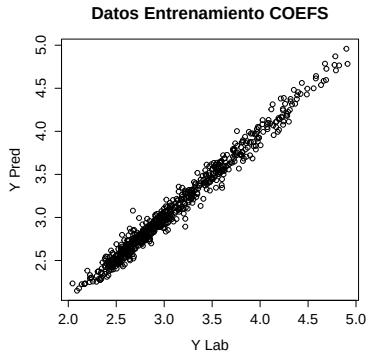

Figura 18: Datos Entrenamiento $MSE = 0{,}006$ $R^2 = 0{,}9805$

Evaluación del modelo con los datos de test.

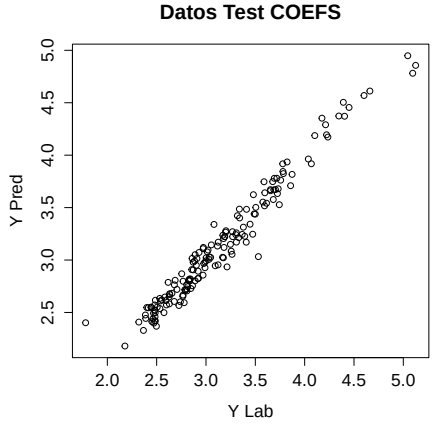

Figura 19: Datos Test $MSE = 0{,}0141$ $R^2 = 0{,}9612$

Al eliminar el ruido, registramos resultados notoriamente mejores. Logramos superar el R^2 objetivo de 0,92 tanto para los datos de entrenamiento (0,98) como para los datos de test (0,96).

7.6. Resumen resultados obtenidos

Cuadro de resumen resultados obtenidos:

Modelo	MSE CV	Entr MSE	Entr R^2	Test MSE	Test R^2
PLS Espectro Original, 12 componentes	0.0102	0.0086	0.9748	0.0137	0.9624
SVM Espectro Original	0.2092	0.1493	0.5652	0.2805	0.2473
SVM Espectro Coeficientes DWT	0.0448	0.0206	0.9408	0.0553	0.8484
KRLS Espectro Coeficientes DWT	N/A	N/A	0.9908	N/A	0.8713
SVM Espectro Suavizado Coeficientes DWT	0.011	0.006	0.9805	0.0141	0.9612

Salvo para el tratamiento directo con SVM de los datos originales del espectro, en todos los casos logramos superar el R^2 objetivos para los datos de entrenamiento.

Para los datos de test, superamos el objetivo con SVM aplicado a los coeficientes de los espectros suavizados (0,9612).

8. Conclusiones

- Se superó ampliamente el objetivo de 0.92 planteado por los expertos de campo
- Se logró mejorar los resultados de las técnicas SVM a través de la eliminación de ruido y reducción de dimensionalidad funcional.
- Aunque la muestra es grande, lo que favorece a PLSR, se obtuvieron resultados similares con SVM. Para muestras futuras más pequeñas hemos logrado una técnica con resultados prometedores.
- Si bien los tiempos de ajuste del modelo son largos (media hora en el peor de los casos), los tiempos de predicción son casi instantáneos.

8.1. Cumplimiento del objetivo

Se logró cumplir en forma holgada con el objetivo planteado. Para medir los resultados utilizamos el R^2 obtenido en el set de datos de test.

Modelo	Test R^2
Valor Objetivo planteado por Expertos de Campo	0.9200
Valor Obtenido con PLS en publicación [32]	0.9411
Valor Obtenido con SVM en publicación [32]	0.9724
PLS Espectro Original, 12 componentes	0.9624
SVM Espectro Suavizado Coeficientes DWT	0.9612

Es de destacar que los R^2 resultantes de nuestros experimentos, son los calculados con la muestra de test. Si utilizaramos el R^2 de la muestra de entrenamiento, superaríamos inclusive los resultados de publicados en [32].

8.2. Mejora de resultados eliminando ruido y reduciendo dimensionalidad

El modelo SVM tratando los datos originales del espectro no logra resultados últiles, obteniéndose un R^2 lejos de 1, que indica el fracaso de la función predictora obtenida.

Utilizamos con éxito la reducción de la dimensionalidad de los datos funcionales a través de la utilización de los coeficientes wavelet. Esos nos permitió ajustar una función más efectiva con SVM.

Procedimos a utilizar técnicas de eliminación de ruido que mejoraron aún más los resultados, permitiéndonos superar el R^2 objetivo y aproximarnos al R^2 publicado en [32].

8.3. Regresión PLS y SVM

En las publicaciones relevadas se utiliza como modelo de referencia el PLS. PLS no tiene los mejores resultados, dado que fracasa en obtener una buena función predictora cuando las muestras son peqeñas ($n = 50$).

En nuestro caso, para una muestra muy grande de mas de 600 en la muestra de entrenamiento, PLS funciona con resultados razonables tratando directamente los datos originales del espectro. Esto se produce como ya dijimos por el tamaño de la muestra, así como por la selección de características, método con el cual de hecho estamos eliminando ruido implicitamente.

SVM alcanza resultados competitivos con PLS recién al eliminar el ruido de la muestra y reducir las dimensiones utilizando los coeficientes wavelet.

Los buenos resultados de SVM sobre datos suavizados y coeficientes DWT, nos permiten disponer un método que potencialmente puede dar resultados satisfactorios cuando se traten muestras pequeñas.

En un escenario de recalibración del sistema, dónde el costo de cada análisis químico tradicional de una muestra es muy alto, será crucial poder disponer de esta técnica.

8.4. Tiempos de cómputo

Las distintas técnicas implicaron la utilización de un alto poder de cómputo, registrándose tiempos largos de procesamiento.

Esta situación no genera problemas en la utilización de estos modelos, dado que el cómputo efectivo de la función predictora aplicado a una nueva muestra son muy pequeños.

8.5. Robustez de SVM

Si bien encontramos outliers en los datos, no fue necesario descartar estos datos para cumplir con el objetivo. Nos planteamos en un futuro inmediato, reprocesar los modelos obtenidos descartando los outliers y verificar con los expertos los registros de las muestras etiquetadas como valores atípicos.

9. Trabajo futuro

Tareas siguientes a la culminación de este trabajo:

Luego de la etapa que intentamos cumplir con este trabajo, continuaremos trabajando sobre el problema en cuestión. Algunas de las tareas que se considera relevante abordar a corto plazo son:

- Simular una muestra pequeña a partir de la muestra de entrenamiento
- Generar modelos SVMr y KRLS independientes para predecir valores de otras 3 substancias disponibles en el conjunto de datos.
- Probar otras técnicas de suavizado/eliminación de ruido
- Aplicar KRLS a datos suavizados
- Introducir outliers artificiales y reprocesar
- Extraer outliers y reprocesar

Lineas de investigación tentativas:

- Generar un modelo que busque la función $f : \mathbb{R}^{2201} \rightarrow \mathbb{R}^4$ o para para el caso de los coeficientes wavelet $f : \mathbb{R}^{69} \rightarrow \mathbb{R}^4$ que permita predecir los valores de las 4 substancias simultáneamente [21]

10. Referencias

[1] Aguilera Ana M, Escabias Manuel, Valderrama Mariano J, and Aguilera-Morillo M Carmen. Functional analysis of chemometric data. *Open Journal of Statistics*, 2013, 2013.

[2] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[4] Ana Arribas-Gil and Juan Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 2014.

[5] Tracy M Blackmer, James S Schepers, and Gary E Varvel. Light reflectance compared with other nitrogen stress measurements in corn leaves. *Agronomy Journal*, 86(6):934–938, 1994.

[6] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May 2000.

[7] F.T. Chau, Y.Z. Liang, J. Gao, and X.G. Shao. *Chemometrics: From Basics to Wavelet Transform*. Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications. Wiley, 2004.

[8] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[9] F. Cucker and D.X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007.

[10] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007.

[11] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[12] Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012.

[13] Jens Hainmueller and Chad Hazlett. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 2013.

[14] Jens Hainmueller and Chad Hazlett. *KRLS: Kernel-based Regularized Least Squares (KRLS)*, 2013. R package version 0.3-2.

[15] Hui hua Yang, Feng Qin, Yong Wang, Qiong lin Liang, Yi ming Wang, and Guo an Luo. Laplacian regularized least squares regression and its dynamic parameter optimization for near infrared spectroscopy modeling. *2013 International Conference on Computing, Networking and Communications (ICNC)*, 1:591–595, 2007.

[16] A.J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer-Verlag New York, 2008.

[17] Zahra Monsef Khoshhesab. *Infrared Spectroscopy - Materials Science, Engineering and Technology*. Intech, 2012.

[18] Hongdong Li, Yizeng Liang, and Qingsong Xu. Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, 95(2):188 – 198, 2009.

[19] Liwen Liang, Bin Wang, Ye Guo, Hong Ni, and Yulin Ren. A support vector machine-based analysis method with wavelet denoised near-infrared spectroscopy. *Vibrational Spectroscopy*, 49(2):274 – 277, 2009.

[20] K.B. Lipkowitz, T.R. Cundari, and D.B. Boyd. *Reviews in Computational Chemistry*. Number v. 23 in Reviews in Computational Chemistry. Wiley, 2007.

[21] Heng-Hui Lue. Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference*, 139(8):2656 – 2664, 2009.

[22] T. Poggio and S. Smale. The mathematics of learning: Dealing with data *. In Wesley Chu and Tsau Lin, editors, *Foundations and Advances in Data Mining*, volume 180 of *Studies in Fuzziness and Soft Computing*, pages 1–19. Springer Berlin Heidelberg, 2005.

[23] Ryan M. Rifkin and Ross A. Lippert. Notes on regularized least-squares. Technical report, 2007.

[24] RA Rossel and Thorsten Behrens. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1):46–54, 2010.

[25] Bernhard Scholkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *In Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426, 2001.

[26] Yongni Shao, Yong He, and Yanyan Wang. A new approach to discriminate varieties of tobacco using vis/near infrared spectra. *European Food Research and Technology*, 224(5):591–596, 2007.

[27] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[28] L. Torgo. *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010.

[29] S. Weisberg. *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley, 2013.

[30] Svante Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109 – 115, 1995. InCINC '94 Selected papers from the First International Chemometrics Internet Conference.

[31] Yanfang Zhai, Lijuan Cui, Xin Zhou, Yin Gao, Teng Fei, and Wenxiu Gao. Estimation of nitrogen, phosphorus, and potassium contents in the leaves of different plants using laboratory-based visible and near-infrared reflectance spectroscopy: Comparison of partial least-square regression and support vector machine regression methods. *Int. J. Remote Sens.*, 34(7):2502–2518, April 2013.

[32] Y. Zhang, Q. Cong, Y. Xie, . JingxiuYang, and B. Zhao. Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine. *Spectrochim Acta A Mol Biomol Spectrosc*, 71(4):1408–13, 2008.

11. Apéndice - Relevamiento bibliográfico anotado

A continuación mencionamos las principales referencias prácticas estudiadas, relevantes para este trabajo.

SVM en química y en NIRS genérico

En una primera etapa, relevamos trabajos sobre técnicas estándar y SVM aplicados en la química, para luego buscar en concreto su aplicación en el modelado NIRS.

Support vectors machines and its applications in chemistry

En este trabakp [18] se presenta a SVM como un método apto tanto para clasificación como para regresión aplicados a problemas de la química. Específicamente se trata un problema de regresión en el campo de la quimiometría, donde se plantea predecir el punto de ebullición a partir del espectro de infrarrojo cercano de muestras de diesel. Los resultados del modelo SVMr se comparan con los de un modelo PLS (partial least squares). Se destaca que al modelo PLS, usado como referencia, se le hace imposible tener en cuenta las relaciones no lineales.

Se dispone de una muestra de 246 elementos. Se dividen los datos al azar en una muestra de entrenamiento y otra de test. El rango del espectro utilizado para ajustar el modelo es 760-1100 nm.

Puntos destacables:

- En su introducción teórica, en el punto 5.2 se explica como se transforma el problema de clasificación SVM en uno de regresion.
- En el punto 6.3 se plantea la extracción de outliers de la muestra de entrenamiento antes de generar el modelo.
- Se transforman los datos NIRS con diferenciación de primer órden.
- Para SVMr se utiliza el método ν-SVR con kernel RBF
- Se utiliza el parámetro γ sugerido por defecto por el paquete *libsvm* Se seleccionan los parámetros C y ν a través de un algorimo genético.
- En 7.3 se indica que se le hace dificil explicar los resultados al investigador.

*****A support vector machine-based analysis method with wavelet denoised near-infrared spectroscopy

En este trabajo [19] se propone utiliza la regresión SVM para predecir la concentración de ciertos elementos a partir de datos NIRS de tabletas de medicamentos.

Se dispone de una muestra de 36 elementos. 24 se usan como datos de entrenamiento y 12 de test. El espectro se registra en el rango 760-1100 nm.

Destacamos:

- En la introducción se indica que se usa la transformación wavelet para preprocesar los datos NIRS, de forma de eliminar parte del ruido. Para ello se descartan los coeficientes pequeños obtenidos antes de realizar la transformación wavelet inversa para obtener la señal sin ruido.
- En 3.1 se indica que se utiliza la wavelet de Daubechies db8
- Se utiliza PLS como modelo de referencia
- En el registro de los datos se realiza un promedio de cuatro escaneos.
- En 4 se menciona el preprocesamiento SNV.
- En 4 se menciona que se utiliza ν-SVM. Kernel RBF con validación cruzada 5 para determinar los parámetros del modelo.(ν, σ, C).

Using data mining to model and interpret soil diffuse reflectance spectra

En este trabajo [24] se presenta un abundante comparativo de distintos algoritmos para calibrar espectros de reflactancia vis-NIR para para predecir el contenido SOC, CC y PH de muestras de suelo.

Se utilizan métodos PLSR, MARS, RF, BT, SVM. Se plantea el preprocesamiento de los datos utilizando DWT. Se utilizan métodos de selección de atributos o procedimientos de ranqueo en el dominio del espectro y también en el dominio wavelet.

Se dispone de una muestra de 1104 elementos. El espectro se registra en el rango 350-2500 nm en 876 anchos de banda. Al usar técnicas de reducción de características se utilizan entre 72 y 137 coeficientes wavelet y entre 11 y 31 componentes principales.

Tomamos nota especialmente de:

- Se destaca el poder de la espectroscopía NIR que permite la determinación de la concentración de varios elementos a partir de un solo escaneo no destructivo.
- En 3.2.1.2 al modelar con los coeficientes wavelet se ordenan los coeficientes por su varianza como criterio de selección.
- El modelo SVM es el que obtiene las mejores predicciones, que inclusive son mejoradas con el tratamiento DWT.

NIRS aplicado a hojas de plantas

Observamos que el modelado NIRS cambia cualitativamente dependiendo del objeto de análisis. Es decir, son muy distintos los tipo de relaciones entre los valores del espectro y la variable cuantitativa a predecir, según sea el objeto de estudio el suelo, carne u hojas de vegetales.

Para aproximarnos a nuestro objeto de estudio, el tabaco, revisamos trabajos relativos a técnicas de modelado NIRS sobre hojas de vegetales en general y para el tabaco en particular.

Estimation of Nitrogen, Phosphorus, and Potassium Contents in the Leaves of Different Plants Using Laboratory-based Visible and Near-infrared Reflectance Spectroscopy: Comparison of Partial Least-square Regression and Support Vector Machine Regression Methods

En este trabajo [31] se buscan modelos para estimar los algunos de los componentes de la materia orgánica de las plantas: Nitrógeno, fósforo y potasio.

Se compara modelos PLSR y SVM para predecir los componentes mencionados.

Se utiliza una muestra de 95 hojas de muy diversas especies.

Se aplican varias técnicas de preprocesamiento.

Tomamos notas especialmente de:

- Técnicas de preprocesamiento
- Técnicas de análisis de correlación
- Los R^2 nunca son superiores a 0,7, considerando que son variedades de hojas de distintas especies.

A new approach to discriminate varieties of tobacco using vis/near infrared spectra

En este trabajo [26] se plantea un modelo de clasificación para predecir la marca de cigarrillos en base a datos NIRS.

En la investigación se usa el rango de 340 a 1000 nm. La muestra es de 100 elementos con 661 datos de espectro. Luego del proceso WT se usan 21 datos. La muestra se divide al azar en 80 elementos para entrenamiento y 20 para testing.

Se destacan:

- Se aplican dos tipos de preprocesamiento. Suavizado Savizky-Golay con un gap de tres puntos. Mutiplicative scatter correction (MSC).
- Se utiliza la transformación wavelet Daubechies.
- Se utilizan modelos de BP-redes neurales,

Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine

En este trabajo [32], se trata un problema muy similar al nuestro, pero para una muestra reducida.

Se detacan:

- Utilización de DWT para compresión del espectro.
- Diversas técnicas aplicadas al pre-procesamiento de los datos
- Suavizado previo al procesamiento de los datos