UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

SELECCIÓN GUIADA DE CARACTERÍSTICAS Y BÚSQUEDA DE MODELOS HOMOGÉNEOS EN DATOS DE ALTA DIMENSIONALIDAD: UN ENFOQUE APLICADO A PROBLEMAS DE TELEDETECCIÓN

Autor: Esteban García Cuesta

Directores: Inés Mª. Galván & Antonio J. de Castro González

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Leganés, Noviembre 2009

Tribunal nombrado por el Mgfco. y Excmo. Sr. Rector de la Universidad Carlos III de Madrid, el día ..... de ..... de 2010.

Presidente: D. .....

Vocal: D. .....

Vocal: D. .....

Vocal: D. .....

Secretario: D. .....

Realizado el acto de defensa y lectura de la Tesis el día ..... de ..... de 2010 en .....

Calificación: .....

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

A mis padres y hermano.

En principio la investigación necesita
más cabezas que medios.
– Severo Ochoa (1905-1993)

Agradecimientos

Quisiera agradecer sinceramente a toda la gente que ha estado ayudándome y apoyándome durante estos años de doctorado y sin los cuales este trabajo no pudiera haber sido una realidad.

Agradecer a mi tutor Antonio J. de Castro la oportunidad que me brindó de comenzar mi carrera investigadora y por apoyarme en todo momento durante el proceso. También tengo que agradecer a mi otra tutora, Inés M. Galván, su disposición y ayuda, y por su preocupación por como me encontraba tanto a nivel profesional como personal a lo largo de todos estos años. Gracias a los dos, ha sido una experiencia estupenda y muy gratificante.

Además, ha habido mucha gente alrededor mío durante todo el tiempo que he pasado por la Universidad Carlos III. Como no acordarme de Jacobo Varela y las charlas en el laboratorio sobre álgebra, J. Ramón Martín, Isabel Gómez (pelusaca), Samuel Rodríguez, Jose A. Iglesias, Manuel Gómez, Elisabet Palomo, y de la gente del LIR empezando por el más grande Fernando López, pasasando por Juan Meléndez, José M. Aránda, Susana Briz, Margarita Gallardo, y muchos otros que han estado por ahí.

No puede faltar en los agradecimientos mi no oficial tercer tutor, Fernando de la Torre. El fué quien me ofreció la oportunidad de estar en un entorno de investigación envidiable como es el instituto de robótica de la Universidad Carnegie Mellon. Además, también tengo que agradecerle sus consejos y por introducirme en el campo

de las "maravillosas" componentes principales.

También hay muchos otros compañeros que me han ayudado y acompañado durante mis estancias por Pittsburgh y Pitman. Agradecerles también a ellos su apoyo y compañía, José González, Javier Hernández, Tomás Simón, Aitor Coca, Tejash Patel, Joseph DePasquale, y muchos más.

Aunque he pasado mucho tiempo fuera de mi ciudad natal, ésta siempre ha sido fortaleza y refugio para mí. Allí están muchas de las personas que me han acompañado siempre, estuviera lejos o cerca. Por eso quiero agradecer a todos mis amigos burgaleses, sin excepción alguna, su apoyo y comprensión durante estos años.

Y por supuesto a los últimos que quiero agradecer pero no por ello en menor grado, a mi familia. A mis padres, hermano y a mi reciente cuñada, que me acompañan siempre allá donde vaya.

¡Muchas Gracias!

Resumen

No hay arte abstracto. Debes siempre empezar con algo.
Después puedes quitar cualquier trazo de realidad.
– Pablo Picasso (1881-1973)

Esta tesis estudia los problemas relacionados con la alta dimensionalidad de los datos en un contexto científico de teledetección, con el fin de estimar perfiles de temperatura en el interior de nubes gaseosas a alta temperatura (como es el caso de una llama). El objetivo principal es identificar los problemas de las técnicas existentes en este contexto práctico y proporcionar soluciones.

Para ello se realiza una introducción a los retos presentes en los datos de alta dimensionalidad, y al área de minería de datos que es actualmente la más activa en el estudio y tratamiento de este tipo de datos. La reducción de dimensionalidad aparece como un proceso necesario para solventar algunos de los retos planteados y mejorar el rendimiento de los algoritmos de aprendizaje.

El resto del trabajo está dividido principalmente en dos partes. Cada una de estas partes desarrolla un camino alternativo para reducir la dimensionalidad de los datos y solucionar así los problemas relacionados con la alta dimensionalidad en el contexto de teledetección.

En el primero de ellos, el trabajo se centra en la selección de características no supervisada para buscar la información relevante a la aplicación. El principal problema en la selección de características es la imposibilidad de realizar una búsqueda exhaustiva debido al gran número de posibles soluciones. Por esto, se propone el uso de conocimiento previo específico de la aplicación física a tratar, para guiar el proceso de selección. Los resultados obtenidos muestran que esta solución mejora los resultados

en un entorno de selección no supervisado, o frente a la ausencia de selección.

La segunda parte de esta tesis se centra en la reducción de dimensionalidad desde un punto de vista de extracción de características. En ella se trata de abordar uno de los problemas principales relacionados con la alta dimensionalidad, la multicolinealidad, buscando extraer de un modo supervisado los conjuntos de datos que mantienen un comportamiento similar u homogéneo. Esto va a permitir diferenciar diferentes grupos de datos y, lograr con esta división, aplicar modelos de estimación específicos para los diferentes grupos. La aproximación se basa en estructuras de grafos para incluir la información local de los datos, lo cual es muy útil en nuestra aplicación. Esta solución muestra mejoras significativas en los resultados obtenidos, a la vez que permite obtener estimaciones precisas para los nuevos casos. Además, también posee una interpretación física y ayudará a un mejor entendimiento de la aplicación estudiada.

Abstract

There is no abstract art. You must always start with something.
Afterward you can remove all traces of reality.
– Pablo Picasso (1881-1973)

This thesis studies some of the problems related with high dimensional data in a scientific context, pursuing the estimation of temperature profiles inside a hot gas cloud at high temperature (as it occurs inside a flame). The main objective is to identify the main disadvantages of the actual techniques in this practical context and to provide solutions to them.

For that purpose we introduce currently known challenges related to high dimensional data, and to data mining field which is the most active regarding the study and processing of this type of data. The dimensionality reduction appears as an important step to solve some of the established challenges and to improve the performance of machine learning algorithms.

The work is mainly divided into two parts. Each one of them develops an alternative to reduce the dimensionality of the data solving some of the problems related to high dimensional data in a remote sensing environment.

The first one, focuses on unsupervised feature selection to search for relevant information to the application. The main problem in feature selection is the impossibility to do an exhaustive search due to the huge number of possible solutions. Thus, we propose to use specific physical previous knowledge to guide the selection process. The obtained results show that this solutions improves the results obtained in an unsupervised framework or against non-selection.

The second part of the thesis is focused in dimensionality reduction from a feature

extraction point of view. In it, we try to solve one of the problems related with high dimensionality data, the multicollinearity. For that purpose we extract, in a supervised mode, subsets of data which have similar behavior or are homogeneous. This allows to find out different groups of data and, with this division, to apply specific estimation models for the different discovered groups. This dimensionality reduction approach is based on graph structures which is useful to include local similarity information about the data, which is extremely useful in our application. This solution shows significant improvements and allows better accuracy for new samples. Furthermore, it also has a physical interpretation and it enables a better understanding of the studied application.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

Debido al desarrollo y rápido avance de las tecnologías, cada vez es más común que el hombre se apoye en ellas para el almacenamiento de datos, su procesamiento y posterior uso. En las últimas décadas, la industria de las tecnologías de la información se ha mostrado como una de las que crece con mayor rapidez, y parte de este crecimiento es debido al desarrollo, la gestión, y el análisis de grandes cantidades de datos con fines científicos, médicos, ingenieriles y comerciales. La generación de esta gran cantidad de datos no sólo se debe al mundo científico, sino también a las sociedades, donde muchas de las acciones cotidianas de sus individuos quedan almacenadas, generando así enormes bases de datos.

Algunos ejemplos recientes de fuentes de generación de datos son:

- Bioinformática: su proyecto más renombrado es el "proyecto genoma humano" iniciado formalmente en 1990 y concluido con éxito en 2003. Su objetivo principal era la identificación del genoma humano, y debido a la gran cantidad de datos involucrados fue necesario desarrollar técnicas de almacenamiento masivo, nuevas herramientas de análisis y visualización de datos y, más genéricamente, una adaptación de la computación a problemas propios derivados del uso de una cantidad enorme de datos (p.e. [1, 2]).

- Redes sociales y comportamientos de usuarios: el uso de ordenadores se ha extendido del mundo científico a la sociedad general. Hoy en día es habitual el uso

de un ordenador en las tareas cotidianas, y redes sociales como Facebook© o Myspace© cuentan actualmente con más de 200 millones de usuarios. El análisis de estas redes desde un punto de vista de minería de datos (p.e. teoría de grafos) proporciona información sobre el comportamiento de los individuos e igualmente pueden determinar el funcionamiento organizativo de una empresa, roles individuales, o comportamientos frente a un problema. Este mismo tipo de problemas pueden ser encontrados en banca, telefonía o en cadenas de supermercados donde las transacciones efectuadas (compras, ventas, operaciones en bolsa, etc.) quedan registradas y pueden ser analizadas. Estos entornos generan cada día una gran cantidad de datos que son almacenados, analizados, y finalmente usados o vendidos.

- Teledetección: existen numerosos satélites que recogen diariamente información sobre la Tierra lo que genera millones de datos. Además, debido al desarrollo de la optoelectrónica, se han comenzado a utilizar sensores hiper-espectrales con miles de bandas, lo que aumenta en muchos casos la cantidad de datos disponible. El análisis e interpretación de todas las bandas son procesos críticos para mejorar el conocimiento sobre el objeto medido. Se presume que esta gran cantidad de datos podrá revelar información sobre componentes químicos y aportar ayuda para la identificación de cultivos, transmisión de enfermedades en cultivos, mejorar el entendimiento de las sequías y pestes, entre otros. También en un futuro cercano se podrá esperar la aplicación de estas técnicas en otros entornos como la medicina, o la alimentación.

Todos estos ejemplos muestran como ha habido un cambio significativo durante las últimas décadas en lo referente a la cantidad de datos generados en nuestro entorno. Además, no parece que esta tendencia vaya a disminuir, sino al contrario parece que seguirá aumentando. Se desarrollan cada día nuevos sensores, se crean nuevas fuentes de datos, y se escala a mayores densidades. Una nueva tendencia que se está creando en la sociedad es el uso de gran cantidad de datos en la vida diaria. Hoy en día, la gente utiliza una cámara web para grabar fotos, subirlas a un sitio web, modificarlas, etc. cuando hace dos décadas sólo era posible hacerlo en laboratorios. Sin duda alguna,

esta aceptación social provocará una mayor inversión y desarrollo tecnológico.

Investigadores y científicos del área de la computación han realizado enormes esfuerzos para desarrollar algoritmos que ayuden a los expertos de otras áreas científicas a entender y ganar conocimiento a partir de los datos de una manera eficiente y efectiva. Las áreas de aprendizaje automático y minería de datos son las dos más relevantes, y han contribuido de manera significativa en el avance y el desarrollo de herramientas con fines científicos.

1.1. Aprendizaje automático

Aprendemos, o por inducción o por demostración.
La demostración parte de lo universal; la inducción de lo particular.
–Aristóteles, 384-322 a.C.

Simon (1983) define aprendizaje como "los cambios adaptativos de un sistema que permiten que éste realice la misma tarea o tareas cada vez de una manera más eficiente dentro de la misma población " [3]. Esta definición es muy genérica y puede acotarse a una definición de lo que es el aprendizaje inductivo como "el razonamiento realizado a partir de un conjunto de ejemplos/casos proporcionados, para producir reglas generales" [4]. La figura 1.1 muestra la jerarquía del aprendizaje y los bloques sombreados muestran el entorno de trabajo de esta tesis.

El aprendizaje no supervisado y, más concretamente, las técnicas de agrupamiento han sido una valiosa herramienta en diversos dominios, usadas frecuentemente por científicos para el descubrimiento de patrones representativos en los datos. Por ejemplo en [5] se utiliza para encontrar estructuras similares y agrupar 3000 componentes químicos en un espacio de 90 índices topológicos. Igualmente se han desarrollado numerosos métodos de aprendizaje supervisado (p.e. redes de neuronas, árboles de decisión, máquinas de vectores de soporte, etc.) con los que se han obtenido resultados exitosos en muchas y diversas aplicaciones.

Mas recientemente, se ha comenzado a mostrar interés por el aprendizaje semi-supervisado. Este tipo de aprendizaje trata de resolver el dilema que se presenta

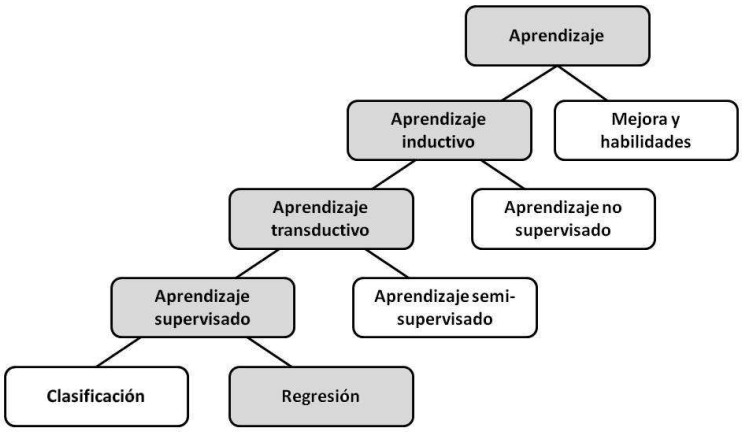

Figura 1.1: Jerarquías de aprendizaje. Los nodos sombreados se corresponden con el aprendizaje supervisado que es objeto de esta tesis.

cuando co-existen datos con etiqueta y sin ella, o cuando el etiquetado de los datos es muy costoso. Para ello, mediante las técnicas de aprendizaje semi-supervisado se propone la utilización de todos los datos no etiquetados, así como los etiquetados para lograr una mejor precisión en el proceso de aprendizaje. El método *Co-training* [6], o algunos métodos basados en grafos como en [7], pertenecen a este tipo de aprendizaje. Algunas de sus aplicaciones más recientes incluyen la clasificación de imágenes hiper-espectrales [8] donde muchas de estas imágenes no están etiquetadas previamente.

1.2. Minería de datos

Una onza de conocimiento vale más
que una tonelada de datos.
–Brian R. Gaines, 1989

Fayyad, Piatetsky-Shaphiro & Smyth (1996) definen minería de datos como la parte algorítmica del proceso no trivial de identificar patrones en datos de una manera válida, novedosa, potencialmente útil, y finalmente inteligible [9]. Esta definición propone

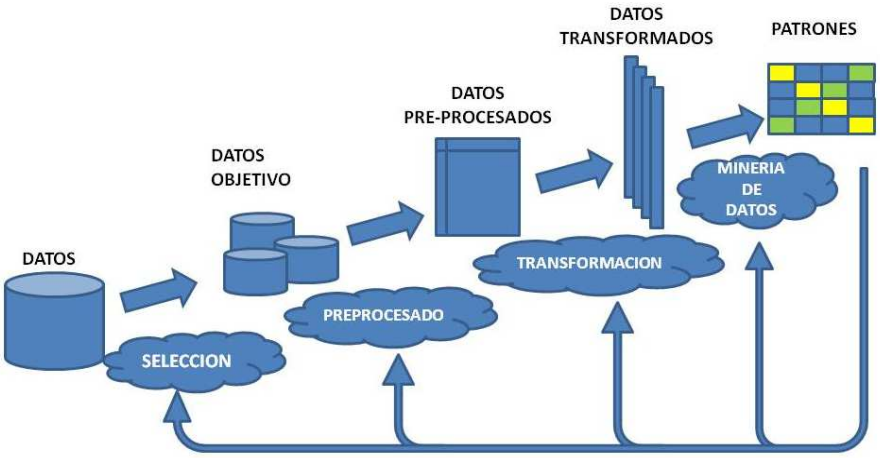

Figura 1.2: Pasos en un proceso de descubrimiento de información en bases de datos.

por tanto la minería de datos como un paso más dentro de un concepto más global denominado descubrimiento de información en bases de datos (*Knowledge Discovery in Databases KDD*). El descubrimiento de la información se define como "el proceso de usar una base de datos junto con cualquier selección, preprocesado y transformación de ésta, para aplicar algún método de minería de datos con el fin de identificar patrones" [9]. La figura 1.2 muestra los diferentes pasos de la metodología y donde se encuentra la fase de minería de datos.

Actualmente es difícil diferenciar en muchos casos un término del otro debido a que algunos algoritmos de minería de datos (p.e. agrupación espectral [10]) incluyen de manera implícita las fases de transformación o selección extendiendo, por tanto, la fase de minería de datos hacia las fases de selección, preproceso y transformación.

Sin duda, el área de minería de datos es una de las que más han contribuido a la adquisición de conocimiento a través del análisis exhaustivo de los datos. Parece también evidente, que el desarrollo asociado a los microprocesadores ha ayudado enormemente a su crecimiento y expansión, y a que hoy en día haya un mayor número de personas trabajando en tareas de análisis de datos que en otras disciplinas más

tradicionales como estadística o matemáticas.

Aunque se han visto muchos casos exitosos en diversos campos científicos, a menudo estos algoritmos carecen de las características necesarias para trabajar en problemas reales como los mencionados al comienzo de este capítulo. Eso es debido a que todos ellos son diseñados para problemas de escala pequeña, o porque hacen asunciones de partida que a menudo no se cumplen en los casos reales. Cuando se obtienen datos de una manera masiva y de manera sistemática, normalmente se desconoce que variables medidas son relevantes para el fenómeno de interés. De este modo, en vez de trabajar con un conjunto de variables pequeño y relevante, se trabaja con datos de alta dimensionalidad donde sólo unas pocas variables contienen la información deseada.

Por ejemplo, en datos de tipo hiper-espectral, los datos se corresponden con curvas asociadas a los espectros de energía. Cada espectro de energía está compuesto por miles de longitudes de onda y cada una lleva un tipo de información. Un criterio de búsqueda de información relevante podría ser la búsqueda de los picos de esas curvas. De ese modo, cada medida es una curva que puede ser analizada para posteriormente seleccionar las zonas correspondientes a los picos, y reduciendo así su dimensionalidad.

Una de las características de los datos de alta dimensionalidad es que a menudo la asunción de que $D < N$, y $N \to \infty$, siendo D el número de variables y N el de ejemplos, no se cumple. Por el contrario, suelen producirse situaciones donde $D \to \infty$ y N se mantiene. Este hecho provoca que los métodos clásicos fallen y sean poco robustos, promoviendo el desarrollo de nuevas herramientas para bases de datos de alta dimensionalidad.

1.3. Motivación y objetivos

Esta tesis se centra en el estudio de técnicas de descubrimiento de información dentro de un contexto de aplicaciones científicas y, más concretamente, en teledetección.

La motivación de este trabajo de investigación viene de nuestro proyecto en identificación y medición de partículas contaminantes en atmósfera. En este entorno y en

otros como en fundiciones, la estimación de los perfiles de temperatura de una nube de gases es muy valioso debido a que la temperatura proporciona información sobre la eficiencia energética o sobre condiciones atmosféricas.

Actualmente, los sensores de teledetección son capaces de realizar mediciones de emisión energética con una resolución sin precedentes. Esto es una oportunidad única para lograr mejorar los resultados obtenidos hasta ahora, pero también presenta nuevos problemas debido a la alta dimensionalidad de los datos con los que se tiene que trabajar.

Presumiblemente sólo hay unos pocos predictores que son clave, pero debido a la alta dimensionalidad y a la multicolinealidad observada no es trivial encontrar estos predictores o alguna transformación que permita mejorar las estimaciones.

1.3.1. Reconstrucción de temperaturas en teledetección

La predicción es muy difícil, especialmente sobre el futuro.
–Niels Bohr, 1885-1962

Se define teledetección como la medición o adquisición de información sobre un objeto o fenómeno a través de un dispositivo que no está físicamente en contacto con ese objeto o fenómeno. Normalmente, se ha denominado con este término a las observaciones terrestres y climatológicas, aunque hoy en día también se utilizan este tipo de mediciones en muchas aplicaciones en tierra, por ejemplo para obtener información de lugares inaccesibles o peligrosos.

En la actualidad, la regulación de sustancias perjudiciales es cada vez más restrictiva en plantas comerciales que usan procesos de combustión (p.e. turbinas de gas, calderas, incineradores). El control y la reconstrucción de la temperatura es un factor importante para entender el mecanismo de combustión, y así minimizar su impacto medioambiental y mejorar su eficiencia [11, 12, 13, 14, 15].

Aunque tradicionalmente se han usado termopares para la obtención de temperaturas, este método tradicional posee una variedad de inconvenientes debido a su interacción con el objeto medido. Se han realizado varios esfuerzos para utilizar métodos que no interaccionen con el objeto de medición aunque la utilización de sensores

infrarrojos multiespectrales es bastante reciente. El reto consiste en buscar un modelo inverso al modelo de emisión de energía conocido como Ecuación de Transferencia Radiativa (RTE) [16]. Este modelo inverso permitirá la predicción de la temperatura en futuros casos a partir de los datos correspondientes a las diferentes longitudes de onda λ. Resaltar que a partir de ahora podemos referirnos indistintamente a este concepto como longitud de onda λ, o como número de onda $\overline{\nu}$, siendo $\overline{\nu} = \frac{1}{\lambda}$.
La ecuación RTE es definida como:

$$R_{\overline{\nu}} = I_{0,\overline{\nu}}\ \tau_{\overline{\nu}}(z_o) + \int_{z_0}^{\infty} B_{\overline{\nu}}\{T(z)\}\frac{d\tau_{\overline{\nu}}(z)}{dz}dz \tag{1.1}$$

donde el subíndice $\overline{\nu}$ se corresponde con las diferentes longitudes de onda, $B_{\overline{\nu}}\{T(z)\}$ es la función de Planck, la cual indica la radiancia emitida por un cuerpo negro a la temperatura T, $\tau_{\overline{\nu}}(z)$ es el coeficiente de transmitancia entre el sensor y una profundidad dada z para una longitud de onda $\overline{\nu}$, e I_0 es la intensidad de emisión inicial. Resaltar que el valor de la transmitancia τ depende, a través de la ley de Beer, de una manera no lineal del perfil de temperatura T. Esta no linealidad es debida a la dependencia de absorción de los gases con respecto a la temperatura para cada longitud de onda.

Por tanto si se considera que no existe ninguna emisión inicial, es decir el término $I_{0,\overline{\nu}}\ \tau_{\overline{\nu}}(z_o) = 0$, entonces la ecuación 1.1 expresa la cantidad de energía emitida por una nube de gases para cada una de las longitudes de onda $\overline{\nu}$, y a partir de ahora nos referiremos a esto como el **modelo directo**.

El objetivo buscado en este trabajo es la inversa de esta ecuación, es decir, dadas las medidas de energía obtenidas por el sensor a distintas longitudes de onda $\overline{\nu}$, queremos obtener los perfiles de temperatura $T(z)$ asociados a esas medidas. A este modelo se le denomina **modelo inverso**. A partir de ahora nos referiremos a los datos $R_{\overline{\nu}}$ obtenidos por el sensor como a las entradas $\mathbf{x} \in \Re^{p\times 1}$, y a los perfiles de temperatura $T(z)$ como a las salidas $\mathbf{y} \in \Re^{s\times 1}$.

Se han realizado algunos trabajos con fines similares durante los últimos años. Por ejemplo, en el área de monitorización atmosférica el objetivo es reconstruir el

estado de la atmósfera y sus constituyentes (p.e. agua). En este contexto hemos identificado principalmente dos aproximaciones. La primera aproximación está basada en variaciones y usa el modelo directo para calcular la radiancia emitida por un estado específico de la atmósfera. Este método optimiza de manera iterativa un vector de estado en función de la distancia entre la radiación medida y la estimada hasta lograr un resultado válido [17]. La principal problemática de este método es la cantidad de tiempo necesario para lograr una solución aceptable debido a que el cálculo del modelo directo tiene un alto coste computacional. La segunda aproximación es más reciente y esta basada en una reducción de la dimensionalidad de las bandas espectrales haciendo uso de las componentes principales [18]. Para obtener la estimación final utiliza las proyecciones de los datos originales como nuevas características, y posteriormente aplica un modelo de redes neuronales utilizando estas proyecciones como entradas.

Para el problema aquí planteado hemos realizado estudios previos [19] y hemos concluido que el comportamiento de un modelo similar al propuesto por [18], y de otros modelos de reducción tipo kernel, producen resultados con un alto grado de desviación del error entre los diferentes ejemplos.

Esta variabilidad nos induce al estudio de los motivos que la originan, y a la búsqueda de nuevos modelos y/o técnicas para reconstruir los perfiles de temperatura a partir de los datos de energía de las distintas longitudes de onda. A continuación se presentan algunos de los problemas y retos asociados a esta aplicación.

1.3.2. Retos asociados a la alta dimensionalidad de los datos

Non quia difficilia sunt, non audemus;
sed quid non audemus, difficilia sunt.
–Seneca(Epistuale Morales 104.26)

Los sensores de teledetección modernos son capaces de realizar mediciones de radiancia con una resolución sin precedentes. Esta alta resolución en las mediciones proporciona datos de alta dimensionalidad asociados a los espectros de energía. Esta alta dimensionalidad en los datos de la aplicación motivadora de este trabajo plantea dos retos fundamentales en el contexto de aprendizaje supervisado.

- Primeramente, el conjunto de características que describe los datos puede contener muchas características irrelevantes o ruidosas, lo cual a menudo hace empeorar los resultados de los algoritmos. Se han encontrado evidencias tanto teóricas como empíricas que muestran que la existencia de características irrelevantes y/o redundantes afectan a la velocidad y a la precisión de los algoritmos de aprendizaje [20, 21, 22].

  En el aprendizaje no supervisado el problema de la existencia de características irrelevantes concierne con su efecto sobre la distancia o medida de similitud. Eventualmente, según aumente el número de características irrelevantes, la similitud entre dos objetos en alta dimensionalidad tiende a desaparecer. Esto hace que la separación y/o identificación de grupos sea más complicada. Ha sido probado en [23, 24], que la distancia entres dos puntos al incrementar la dimensionalidad se mantiene casi constante bajo ciertas condiciones en la distribución de los datos. Esto significará que si consideramos una métrica de distancia que no sea L_1, la distancia entre dos puntos tenderá a ser 0, y el conjunto de puntos formará una sola agrupación.

  En el aprendizaje supervisado, el aumento de características irrelevantes provoca un aumento de colinealidad (entre dos características) o multicolinealidad (más de dos características) en los datos. El término multicolinealidad se refiere a la gran correlación (dependencia lineal) que existe entre varias variables/características que de acuerdo a su naturaleza deberían ser independientes. Siguiendo el mismo razonamiento descrito para el caso de aprendizaje supervisado, la multicolinealidad aumentará debido a que las distancias entre puntos son menores en alta dimensionalidad y, por tanto, aumenta el número de características que sufren de multicolinealidad.

  Los efectos de la multicolinealidad son fácilmente observables cuando se construye un modelo de estimación. Si consideramos un modelo de regresión con dos características x_1 y x_2, y suponemos que están estandarizadas eliminando la media de las características para cada observación, y dividiendo entre la raíz cuadrada de su suma corregida al cuadrado, logramos que la matriz $\mathbf{X}\mathbf{X}^T$

tenga la estructura de una matriz de correlaciones. Esta matriz tiene $1's$ en la diagonal principal y los otros términos contienen la correlación simple entre la característica x_i y la x_j, siendo i y j los índices de la matriz. Entonces, el modelo de regresión para los diferentes casos k puede representarse como

$$y_k = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \epsilon_i \quad k = 1 \dots m \ . \tag{1.2}$$

Suponiendo que la salida y también está centrada, entonces $\beta_0 = 0$ y la solución de mínimos cuadrados se reduce a calcular la pseudoinversa de la matriz $\mathbf{X}^+$ tal que $\mathbf{X}^+ = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$, siendo la solución del sistema $\boldsymbol{\beta} = \mathbf{X}^+\mathbf{Y}^T \in \Re^{p\times s}$ siendo p las dimensiones de la entrada y s las de la salida. Para este modelo,

$$(\mathbf{X}\mathbf{X}^T)^{-1} = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix} \tag{1.3}$$

siendo r_{12} la correlación simple entre x_1 y x_2. Entonces, si hay multicolinealidad, x_1 y x_2 están altamente correladas, y $|r_{12}| \to 1$. En esta situación, los valores de varianza y covarianza de los coeficientes de $\mathbf{X}\mathbf{X}^T$ son muy grandes lo que implica que los coeficientes de regresión $\boldsymbol{\beta}$ son vagamente estimados. Es destacable que el efecto de la multicolinealidad se debe a la dependencia lineal casi exacta entre las características de $\mathbf{X}$, y que cuando $r_{12} \to \pm 1$ esa linealidad es exacta.

En general, siempre existe multicolinealidad en los datos y especialmente en datos de alta dimensionalidad. Por eso debemos hablar de grados de multicolinealidad (severa o suave) en vez de su existencia o ausencia. Uno de los posibles efectos de la multicolinealidad severa es la degradación del proceso de aprendizaje. Este efecto no es causado por la propia multicolinealidad sino por su falta de homogeneidad. Así, si la nueva predicción se hace en la región del espacio de entrada donde ocurre la multicolinealidad, se obtendrán resultados satisfactorios, porque aunque el término individual β_j sea mal estimado, la función $\sum_{j=1}^{p} \beta_j x_{ij}$ puede ser bien estimada. Esta falta de homogeneidad hace que pequeñas variaciones en el conjunto de los datos usados provoque grandes variaciones en los

modelos aprendidos. Este factor de gran variabilidad en el aprendizaje es realmente el que provoca el empeoramiento del proceso, afectando tanto a modelos lineales como no lineales.

- El **segundo gran problema** es causado por la dispersión en los datos que ocurre especialmente cuando se trabaja en alta dimensionalidad. Aunque, como se ha comentado, toda esta generación y tratamiento de grandes cantidades de datos es moderadamente reciente, ya en 1961 Bellman acuñó la hoy bien conocida frase de "la maldición de la dimensionalidad" [25] en relación a la dificultad de optimización por enumeración exhaustiva en espacios de productos. Con esta frase, Bellman nos recuerda que si consideramos una malla cartesiana con unidad de equiespaciado $\frac{1}{10}$ de la unidad de un cubo de 10 dimensiones, tendremos 10^{10} puntos; pero si el cubo tiene 20 dimensiones, entonces tendremos 10^{20} puntos. Su interpretación fue: que si el objetivo es optimizar una función en el espacio continuo del dominio del producto de un par de docenas de dimensiones haciendo una búsqueda exhaustiva en su espacio discreto, fácilmente nos encontraremos realizando trillones de evaluaciones de la función.

  En la aproximación de funciones, si deseamos aproximar una función con D variables de entrada y únicamente conocemos que es una función de Lipschitz[1], entonces necesitamos del orden de $(\frac{1}{\epsilon})^D$ evaluaciones en la malla para obtener una aproximación con un error uniforme ϵ. Si se desea realizar una estimación, la alta dimensionalidad de los datos afecta en la convergencia del límite superior del error. De este modo, es necesaria una gran cantidad de ejemplos para reducir el límite, y además su convergencia es muy lenta a medida que aumenta la dimensión.

  En nuestro caso el número de ejemplos disponibles no es mucho mayor que el número de dimensiones de esos datos. Por lo tanto, nos encontramos con un problema cuando deseamos aproximar a una función objetivo debido al mal

[1]Una función de Lipschitz es un requisito de continuidad de una función donde dado $f : X \to Y$, y siendo d_x y d_y las métricas de la función en X e Y, se cumple que : $d_y(f(x_1), f(x_2)) \leq K \cdot d_x(x_1, x_2)$.

acondicionamiento de los datos.

Para aliviar estos problemas, muchos algoritmos de aprendizaje son a menudo precedidos por una reducción de dimensionalidad. Siguiendo la clasificación en [26], las diferentes posibles aproximaciones para reducir la dimensionalidad se encuentran clasificadas principalmente en dos grupos : la selección de características y la extracción de características.

Los métodos basados en la selección de características tratan de realizar una selección óptima de un subconjunto de características de acuerdo a la función objetivo. La mayoría de los estudios relacionados con la selección de características pertenecen a problemas de clasificación [21, 27, 28, 29], y de regresión [30].

Por el contrario, las técnicas de transformación de características no están limitadas a usar únicamente un subconjunto de las variables originales, sino que se crea un conjunto pequeño de nuevas características realizando una transformación general a partir de los datos de alta dimensionalidad. Normalmente esta transformación involucra todas las características y puede ser lineal o no lineal. Igualmente, las técnicas de extracción de características pueden ser supervisadas o no supervisadas, dependiendo de que se utilicen las variables de salida o no. Nos referimos a técnicas no supervisadas cuando la extracción de características se realiza utilizando únicamente la matriz de datos de entrada y sin usar la información de salida. El análisis de componentes principales (PCA) [31], también conocido como la transformación Karhunen-Loeve o simplemente transformación KL, es posiblemente el método de extracción de características más conocido. Es común referirse como *principal component regression* (PCR) cuando estas nuevas características son utilizadas con fines de regresión.

Entre estas dos aproximaciones, la selección de características tiene la ventaja de que el subconjunto de características final puede tener un significado inteligible y puede ayudar a un mejor conocimiento sobre la problemática tratada. Por el contrario, los métodos de transformación utilizan todas las características para buscar una transformación que ayude a representar mejor el problema. A menudo, la selección de características es considerada más general que la extracción por considerarse como un

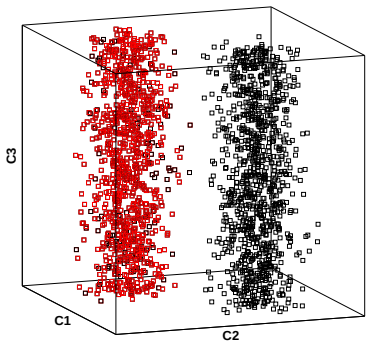

Figura 1.3: Conjunto de datos conteniendo una característica (C3) irrelevante para su clasificación.

caso especial de transformación. Por ejemplo, el método LASSO [32] realiza de manera embebida una selección usando ℓ_1[2] como métrica de penalización y estableciendo así que características son o no son útiles en el modelo.

Normalmente todas estas técnicas de reducción de dimensionalidad se asientan sobre un concepto de "relevancia" para realizar la selección o la transformación de características. Por ejemplo, PCA preserva la varianza de los datos realizando una transformación lineal a partir de las variables originales. Por tanto, uno de los grandes retos para diseñar una técnica de reducción de dimensionalidad es establecer el criterio para definir este concepto de "relevancia". Hay que tener en cuenta que, a menudo, un mismo criterio puede tener buenos resultados para un conjunto de datos, pero puede comportarse muy mal para otro distinto. Por ejemplo, si consideramos la técnica PCA, pueden construirse situaciones para las que la dirección de máxima varianza de los datos no es la mejor solución para separar los distintos grupos. La figura 1.3 muestra visualmente ese efecto, donde la componente principal está muy próxima a la dirección de la característica C3 que a su vez proporciona una mala separación entre los grupos. Debido a estas dificultades, la reducción de dimensionalidad es una cuestión que presenta diversos problemas y sigue siendo actualmente motivo de investigación [29].

[2] ℓ_1 es la norma del espacio vectorial de dimensión 1 y queda definida como su valor absoluto $|\mathbf{v}|$.

1.3.3. Objetivos

El objetivo de esta tesis es la estimación de los perfiles de temperatura de una llama a partir de la radiación que emite.

Cuando una combustión no se realiza de manera completa emite elementos contaminantes, y la temperatura aparece como un parámetro clave para la monitorización y control de estos procesos. Por tanto, su estimación es muy valiosa debido a que proporciona información sobre la eficiencia energética de una combustión.

Para lograr este objetivo es necesario estudiar los problemas que presenta la alta dimensionalidad en este contexto y, desarrollar soluciones que permitan extraer la información relevante para así obtener estimaciones precisas.

Existen dos aproximaciones fundamentales para reducir la dimensionalidad, la selección y la extracción de características. La selección de características tiene una propiedad muy útil para la aplicación propuesta. Esta propiedad es la obtención de un subconjunto de características originales las cuales se corresponden con diferentes longitudes de onda propiciando tanto una interpretación física como el diseño de sensores específicos. Por esto, un objetivo será el desarrollo de una técnica de reducción de dimensionalidad, basada en la selección de características, la cual proporcione un subconjunto de características originales representativas para la estimación.

A priori, es lógico pensar que la utilización de todas las características para realizar la estimación proporcionará mejores resultados. Por eso también se presenta como un segundo objetivo el desarrollo de un método de reducción de dimensionalidad que utilice todas las características y obtenga resultados con la mayor precisión posible.

Para poder lograr estos objetivos es necesario la identificación previa de los problemas que surgen al aplicar técnicas ya conocidas pertenecientes a los campos de aprendizaje automático y minería de datos.

1.4. Solución: Selección de características basada en características no redundantes e información previa

En la primera solución planteada en esta tesis, el problema de la alta dimensionalidad ha sido tratado desde una aproximación de selección de características no supervisada. Esta selección tiene dos ventajas principales. La primera es que al obtener un subconjunto de las variables originales se facilita la interpretabilidad de los resultados y aumenta el conocimiento sobre el problema, o incluso puede corroborarse información sobre el problema previamente presumible o ya conocida, permitiendo así la validación en ambos sentidos. La segunda es que este conocimiento permite el diseño de sensores específicos para ese subconjunto de características obtenidas, siendo muy útil en problemas con contextos específicos y menos genéricos que los estudiados en esta tesis.

Dado el conjunto de datos de alta dimensionalidad correspondiente a los espectros de energía, el método propuesto se basa en realizar una selección de características utilizando la estructura de las direcciones principales obtenidas por el análisis de componentes principales. Simultáneamente se analiza la información física conocida a través de la funciones de peso asociadas a la ecuación de transferencia radiativa. Esto es explicado con más detalle en la sección 3.3. Por ahora es suficiente conocer que esta información, conocida a priori, establece una aproximación a la probabilidad de que una característica esté relacionada con una determinada información de salida. Finalmente, se establece un compromiso entre la reducción de dimensionalidad proporcionada por el análisis de las estructuras de las componentes PCA y la información física de que se dispone. Para ello se aplica una selección de características sobre los vectores de dirección principales teniendo en cuenta la información física asociada a cada característica.

De este modo, se soluciona uno de los problemas que presenta la técnica PCA el cual se debe a que la nueva base preserva la varianza en los datos pero no necesariamente proporciona una representación con significado, es decir, que no se corresponde

con cualidades físicas. Además, el uso de la información a priori va a evitar seleccionar características redundantes, dispersando las características seleccionadas a lo largo de todo el espectro.

Existen dos problemas principales para el desarrollo de esta propuesta:

1. cómo generar la combinación de criterios entre la técnica PCA y la información física

2. cómo establecer el número de características a seleccionar y qué grado de variabilidad debe ser mantenido.

El primer problema planteado se corresponde realmente con el establecimiento del criterio de "relevancia" que antes se ha mencionado. Nosotros hemos propuesto usar un selector de máximos locales que simultáneamente busca aquellos coeficientes de mayor valor y, por tanto, con más información, pero también mantiene la dispersión de la selección buscando diferentes características (asociadas con distintas longitudes de onda y por tanto cualidades físicas) de manera que se mantenga la diversidad de la información seleccionada. Recientemente, en otros estudios [33] se ha utilizado también las direcciones obtenidas por PCA para reducir la redundancia de las características en el proceso selección, pero al no disponer de información a priori realizan una agrupación sobre los coeficientes de las direcciones principales, para finalmente elegir una característica representante para cada grupo obtenido.

Con respecto al segundo problema, habitualmente se utiliza el valor de la varianza acumulada para determinar el número de dimensiones mínimo necesario para conservar la varianza del espacio original. También es muy común buscar el "bajo codo" de la representación de la varianza acumulada o *scree graph* [31]. Nosotros hemos utilizado el criterio de la varianza acumulada para seleccionar el número de dimensiones y después hemos asignado un número de características a seleccionar en cada una de ellas. El método propuesto ha superado a otras técnicas de reducción de dimensionalidad utilizadas en este contexto.

1.5. Solución: Extracción de características y búsqueda de modelos homogéneos.

La segunda parte de esta tesis investiga la utilización de técnicas de extracción de características en datos de alta dimensionalidad aplicados a la teledetección. Como ya se ha comentado en el apartado 1.3.2, uno de los problemas al que nos enfrentamos al tratar los datos de alta dimensionalidad pertenecientes al problema de reconstrucción de temperaturas es la multicolinealidad. Recordar que los efectos de la multicolinealidad se muestran o acentúan especialmente cuando existe falta de homogeneidad en los datos. En la práctica, esto se ve reflejado en que pequeñas variaciones en el conjunto de los datos usados provocan grandes variaciones en los modelos aprendidos empeorando el proceso de aprendizaje.

Si nos referimos a falta de homogeneidad en un problema de regresión estamos diciendo que las relaciones entre el conjunto de datos de entrada y de salida están mezcladas, estableciendo relaciones equívocas entre los dos conjuntos de datos.

En consecuencia, vamos a estar interesados en descubrir qué subconjuntos de datos asociados a un contexto físico específico pueden ser estimados correctamente independientemente del resto de subconjuntos. De esta manera, se va a caracterizar el problema en diferentes modelos que poseen características similares, tanto desde un punto de vista del aprendizaje de estos modelos, como desde un punto de vista físico.

En la aplicación de teledetección en la que se ha desarrollado este trabajo, este tipo de estudio va a proporcionar dos valores importantes. El primero es que los científicos pueden obtener información para identificar qué tipo de problemas reales pueden ser abordados como si se tratasen de un único sistema, y aplicar un modelo homogéneo y específico para su resolución. El segundo valor es que permitirá abordar el problema global utilizando todas las características y logrando una mayor precisión en las estimaciones.

La segunda solución propuesta en esta tesis es aprender un modelo que capture la correlación entre las estructuras de dos conjuntos de datos, manteniendo al mismo tiempo la estructura local de cada uno de ellos. La búsqueda de una estructura

intrínseca a los dos conjuntos de datos y que mantenga su homogeneidad permite una división del modelo global en varios modelos locales a la vez que se alivia el problema de la multicolinealidad.

Para ello, asumimos que existe un subespacio común a ambos espacios de entrada y salida, y éste contiene la estructura con la información que buscamos. Para poder realizar esta aproximación es necesario:

1. definir el modo en el que se mide la distancia entre los diferentes ejemplos y la técnica de representación.

2. detectar las estructuras de los subespacios para encontrar zonas homogéneas de datos y agruparlas

3. elaborar diferentes modelos para los diferentes sub-conjuntos de ejemplos identificados como homogéneos.

El primer punto se corresponde con la métrica utilizada para establecer la distancia entre dos ejemplos y el modo de representar dicha distancia. Para esto vamos a seguir una representación basada en grafos la cual nos permite incorporar tanto información local como global de una manera sencilla [34, 35, 36] pudiendo así explotar la homogeneidad local de los datos.

En cuanto al modo de detectar las estructuras en el nuevo subespacio generado utilizamos un algoritmo basado en densidades debido a que los conjuntos de datos que posean características de homogeneidad quedarán en una misma zona del subespacio generado frente a otros grupos. Un resumen de algunas de las características de varios de estos algoritmos basados en densidades puede encontrarse en [37].

El tercer paso de la propuesta consiste en construir diferentes modelos de estimación para los diferentes sub-conjuntos obtenidos. En nuestro caso, se utilizan modelos lineales debido a que la propuesta se ha basado en la localidad lineal de los datos y por tanto los modelos de estimación podrán también ser lineales.

Además, esta aproximación propuesta abre una nueva forma de abordar problemas de regresión en alta dimensionalidad debido a que la reducción de la dimensionalidad del modo planteado está muy relacionada con la agrupación de datos y, por tanto, con la discriminación entre modelos.

1.6. Organización

El resto de esta tesis está organizada como sigue. El capítulo 2 realiza una explicación de la física de una combustión haciendo especial hincapié en los elementos importantes para aplicaciones de teledetección.

La Parte I contiene los capítulos del 3 al 5, y estudia la selección de características no supervisada en datos de alta dimensionalidad usando información a priori. El capítulo 3 hace una introducción a los conceptos generales necesarios para poder desarrollar la técnica propuesta de selección, la cual combina la información física y estadística, y es explicada en el capítulo 4. El capítulo 5 muestra los resultados obtenidos al aplicar dicha técnica.

La Parte II contiene los capítulos del 6 al 8 donde se estudia el problema de la búsqueda de modelos homogéneos en los conjuntos de datos de entrada y salida. El capítulo 6 hace una introducción a la terminología de grafos y a la reducción de dimensionalidad usando grafos de similitud. Estos conceptos son necesarios para el desarrollo de nuestra propuesta que se muestra en el capítulo 7. El capítulo 8 recoge los resultados obtenidos correspondientes a esta segunda parte y sus conclusiones. Finalmente, el capítulo 9 presenta las conclusiones generales obtenidas con el desarrollo de esta tesis y se resumen las contribuciones que hemos realizado.

Capítulo 2

Física de la combustión

Este cosmos, el único, no lo hizo un dios ni un hombre,
sino que siempre fue, es y será, Fuego siempre vivo
que se propaga siguiendo un patrón, y se extingue según un patrón
–Heráclito (Filósofo griego)

2.1. Introducción. Procesos de combustión y su control

Los procesos de combustión han estado íntimamente ligados a la actividad cotidiana del hombre desde la prehistoria con la aparición del fuego. Hoy en día siguen siendo casi en un 95 % la fuente principal de energía, además de estar involucrados en muchos procesos industriales. El principal inconveniente que presentan es la contaminación atmosférica que generan. Aproximadamente un 90 % de la polución tiene su origen en estos procesos de combustión. Otro problema asociado, es la previsible futura escasez de combustibles fósiles que son la principal fuente de energía.

Frente a estos problemas, el enfoque medioambiental más correcto es el de la prevención. Para evitar en lo posible las emisiones contaminantes y contribuir al ahorro energético, es necesario un conocimiento profundo de estos procesos, así como una mejora y control de la eficiencia de los sistemas de combustión.

Una llama está generada por una reacción química de oxidación entre un reactivo y un oxidante, que se autopropaga y en la que se desprende calor (reacción exotérmica). El reactivo o combustible puede ser sólido, líquido o gas, y el oxidante suele ser oxígeno puro o aire. En estas reacciones hay además un transporte de calor y una difusión de las especies reactivas. Los productos de esta reacción son principalmente H_2O , CO_2 ,y nitrógeno molecular [38], pero también se suelen observar otros tales como CO e hidrocarburos inquemados, cuando la combustión no se realiza de forma completa, y óxidos de nitrógeno (NO_2 y NO) cuando se alcanzan temperaturas elevadas.

La monitorización precisa del proceso de combustión en un horno industrial juega un papel fundamental tanto para optimizar el propio proceso de combustión, minimizando las pérdidas energéticas, como para controlar la producción y emisión a la atmósfera de estos gases contaminantes y partículas. En algunos casos es posible realizar manualmente el proceso de monitorización a partir de la experiencia del operario combinada con información parcial proporcionada por algún sistema de control. Sin embargo, en la mayoría de los casos es necesario un sistema inteligente y automático de control para asegurar la correcta monitorización del proceso.

Uno de los parámetros más importantes de control es la temperatura de la llama. Pequeños cambios en las condiciones de operación (decenas de grados) son responsables de importantes incrementos en la generación de contaminantes atmosféricos [14]. Los parámetros medidos habitualmente en llamas son la temperatura y la concentración de los distintos productos de combustión, así como la distribución espacial y temporal de estas magnitudes. Las primeras medidas de llamas fueron realizadas mediante termopares, pirómetros de succión y medidas extractivas para gases. Estas técnicas no siempre pueden ser utilizadas debido a las altas temperaturas asociadas al proceso de combustión. Además, resultan muy restrictivas debido a que son medidas puntuales promediadas en el tiempo, y no pueden dar cuenta de la alta variabilidad temporal y espacial de estos procesos. Como ejemplo de ésto, es conocido que al introducir un termopar en el interior de la llama el valor de la temperatura se ve afectado, ya que se produce un efecto de enfriamiento en el entorno de medida (*flame quenching*) asociado a la conductividad térmica del termopar.

Las técnicas ópticas han superado estas deficiencias gracias a su mayor resolución

espacio-temporal. Además, poseen la ventaja de ser técnicas no intrusivas por lo que no perturban el sistema a medir. Las técnicas ópticas surgen como consecuencia del desarrollo de la tecnología de los láser (de alta potencia, pulsados, diodos láser sintonizables, etc.), y son en la actualidad una de las principales herramientas para el estudio de procesos de combustión y de los contaminantes emitidos en estos procesos. Estas técnicas utilizan sensores ultravioleta, visible o infrarrojo, y están basadas en la dispersión Raman anti-Stokes, Rayleigh, fluorescencia inducida por láser y diferentes técnicas espectroscópicas [39]. Utilizando estas técnicas se consigue recuperar valores de temperaturas con errores entre el 1-5 % [11, 14, 40, 41, 42]. Dos de sus desventajas son que requieren que las concentraciones de los gases a medir sean altas y que se realicen en laboratorios con salas limpias [43].

2.2. Teledetección. Interacción radiación-materia

No tenemos el derecho de asumir que las leyes físicas existan,
o si han existido hasta ahora,
que seguirán existiendo en el futuro de manera similar.
–Max Planck (1858-1947)

Frente a estos inconvenientes y debido a su carácter no intrusivo, las técnicas de teledetección aparecen como más adecuadas para resolver este problema. En concreto, la teledetección en el infrarrojo aparece como una técnica muy interesante, ya que los gases calientes que componen la llama (principalmente dióxido de carbono CO_2 y agua H_2O) presentan bandas de emisión en la región infrarroja de longitudes de onda entre 2 y 20 μm

La capacidad de un objeto para reflejar, absorber, dispersar o emitir radiación depende de la naturaleza y estado del objeto. Por consiguiente, mediante el estudio de la interacción entre un objeto y la radiación, se podrían determinar las características y el estado del objeto en cuestión. La teledetección de cualquier objeto consiste en la medida de alguna propiedad característica del objeto mediante algún procedimiento que no implique el contacto directo con él. Por tanto, la teledetección

involucra procesos de propagación de señal tales como la energía que es absorbida, reflejada, dispersada o emitida por ese objeto generalmente en forma de ondas electromagnéticas. El fenómeno físico en que está basada la teledetección es la interacción radiación-materia.

Cualquier cuerpo a una temperatura superior a 0 K emite y absorbe radiación electromagnética. La agitación de átomos y moléculas implica el movimiento acelerado de cargas electrónicas, que emiten radiación de acuerdo a las leyes del electromagnetismo. La magnitud física que describe de manera más completa la emisión de radiación electromagnética de un cuerpo es la radiancia, que nos da la energía radiada por la fuente por unidad de tiempo, área y dirección (ángulo sólido). Esta magnitud depende solamente de la naturaleza de la fuente y de su temperatura, pero no de sus características geométricas.

En realidad la magnitud más interesante no es la radiancia sino la radiancia espectral, que tiene en cuenta cómo se distribuye la radiancia en las diferentes longitudes de onda que constituyen el espectro electromagnético. Por tanto, la magnitud que se emplea para caracterizar espectralmente la energía es la longitud de onda λ. Sin embargo, en espectroscopia es mucho más común utilizar el número de onda $\overline{\nu} = \frac{1}{\lambda}$, ya que el número de onda es proporcional a la energía que llega al detector. Las unidades utilizadas generalmente para el número de onda son cm^{-1}.

Se define un cuerpo negro como un cuerpo ideal caracterizado porque absorbe toda la radiación que le llega y emite toda esa radiación (no transmite ni refleja radiación para ningún número de onda). La característica principal de un cuerpo negro es que su radiancia depende exclusivamente de la temperatura, y viene dada por la ley de Planck

$$B(\overline{\nu}, T) = \frac{c_1 \overline{\nu}^3}{exp(\frac{c_2 \overline{\nu}}{T}) - 1} \tag{2.1}$$

donde $c_1 = 1{,}191 \cdot 10^{-12}$ y $c_2 = 1{,}4388$.

Sin embargo, sólo unas pocas superficies reales se aproximan al comportamiento emisivo del cuerpo negro (negro de carbón, carborundo). La mayoría de las superficies absorben sólo una parte de la energía que reciben, y no se comportan exactamente como cuerpos negros. Al no absorber toda la radiación que reciben, un cuerpo no negro

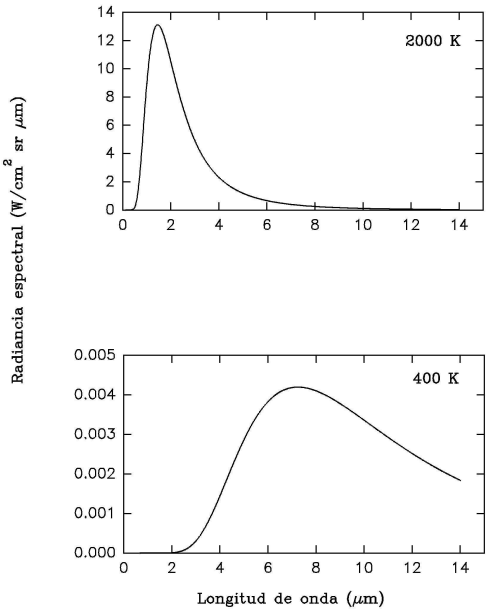

Figura 2.1: Radiancia espectral de cuerpo negro para dos temperaturas diferentes.

siempre emite menos que un cuerpo negro a la misma temperatura. Para cuerpos reales, la ecuación anterior se modifica de la siguiente manera

$$L(\overline{\nu}, T) = \epsilon(\overline{\nu}) \frac{c_1 \overline{\nu}^3}{exp(\frac{c_2 \overline{\nu}}{T}) - 1} \tag{2.2}$$

donde ϵ se conoce como la emisividad de la superficie en cuestión. Para un cuerpo negro, $\epsilon = 1$ para todas las longitudes de onda. Se denominan cuerpos grises aquellos cuerpos para los cuales $\epsilon < 1$, pero su valor no depende de la longitud de onda. Se denominan cuerpos selectivos aquellos cuerpos para los cuales $\epsilon < 1$, y además su valor depende de la longitud de onda. Los gases calientes constituyen el ejemplo más típico de cuerpos selectivos. Su emisividad es nula en casi todas las longitudes de onda, excepto en unos determinados intervalos que corresponden con sus bandas de emisión. La posición espectral de estas bandas de emisión es característica del propio gas. Además, la mayor parte de la energía emitida por gases calientes se corresponde con la región espectral del infrarrojo medio (entre 2 y 20 mm).

Cuando una molécula absorbe o emite un fotón, su estado energético cambia. En general, el cambio en energía se traduce en un cambio traslacional, o como un

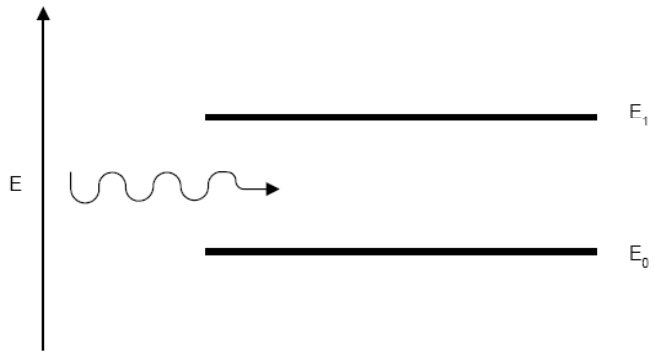

Figura 2.2: Absorción de un fotón en un sistema de dos niveles cuánticos.

cambio en el estado electrónico, vibracional o rotacional de la molécula. Debido a que las energías involucradas en estos cambios son muy diferentes, estos procesos pueden tratarse de manera independiente. Exceptuando los cambios traslacionales (que permiten básicamente un rango continuo de energías), los estados energéticos de las moléculas están cuantizados. Esto significa que la emisión y absorción de luz sólo puede tener lugar a unas frecuencias bien determinadas. Además, cada molécula tiene un conjunto de frecuencias de absorción/emisión propio haciendo que su espectro de absorción constituya una "huella digital"que permite identificar la especie química.

La figura 2.2 muestra el ejemplo más sencillo de un sistema de dos niveles de energía. Un fotón cuya frecuencia ν_0 venga dada por la diferencia de energías $E_1 - E_0 = h \cdot \nu_0$ será absorbido por el sistema. Por lo tanto, el espectro de absorción de este sistema presentará una línea de absorción para la frecuencia ν_0; el resto de las frecuencias no interacciona con el sistema.

Por lo tanto, la radiación electromagnética sólo es absorbida cuando su energía iguala la diferencia entre dos niveles de energía. Además, este proceso de absorción no se produce para cualquier pareja de niveles. Las denominadas reglas de selección determinan entre qué niveles están permitidas estas transiciones.

La absorción de luz en las regiones ultravioleta y visible del espectro, es el resultado de transiciones entre estados de energía electrónicos en átomos o moléculas. Las absorciones que se producen en la región espectral del infrarrojo involucran energías pequeñas, que no estarán relacionadas con la estructura electrónica de la molécula.

Se pueden obtener resultados sorprendentemente buenos si uno considera un modelo muy sencillo de molécula, como un conjunto de masas unidas por muelles (que representan los enlaces), y sin considerar la nube electrónica que genera las transiciones electrónicas. Según este sencillo modelo mecánico de la molécula, ésta puede absorber energía para rotar o para vibrar, es decir, para cambiar su estado energético rotacional o vibracional. Las energías que se involucran en estos procesos corresponden a frecuencias del infrarrojo. En general, una molécula vibra y rota simultáneamente, por lo que se observa un espectro de absorción denominado vibracional-rotacional,

Por último, realizar un breve comentario sobre la forma de las líneas de absorción. Como se ha indicado anteriormente, los procesos de absorción de fotones tienen lugar a frecuencias bien determinadas, por lo que en principio el ancho de las líneas de absorción vendría determinado por el principio de incertidumbre: el ancho es inversamente proporcional al tiempo de vida del estado excitado, siendo valores típicos del ancho $\Delta = 10^{-7}\text{cm}^{-1}$. Sin embargo, existen una serie de mecanismos que producen un ensanchamiento de las líneas de absorción.

Uno de éstos, se debe a la velocidad de las moléculas lo que provoca un desplazamiento Doppler en la frecuencia de la línea. Para un gas isótropo, la dirección de las moléculas es aleatoria, por lo que el ensanchamiento Doppler de la línea de absorción tiene un perfil gaussiano. Este ensanchamiento depende de la temperatura, y predomina sólo en condiciones de baja presión.

Otro de los mecanismos de ensanchamiento de línea es el asociado a las colisiones entre moléculas. En este caso, el perfil de línea es un perfil lorentziano, cuya ecuación viene dada por

$$g_c(\overline{\nu} - \overline{\nu}_0) = \frac{\gamma}{\pi} \frac{1}{[(\overline{\nu} - \overline{\nu}_0)^2 + \gamma_c^2]}$$

donde γ_c es el semiancho de línea, y es proporcional tanto a la temperatura como a la presión.

Como se ha indicado anteriormente, las vibraciones y las rotaciones de una molécula son las responsables de las bandas de absorción de la misma en el rango espectral del infrarrojo. Sin embargo, no todas la vibraciones y/o rotaciones producen una

absorción de la radiación incidente. Sólo los modos vibracionales y rotacionales de moléculares con momento dipolar diferente de cero, o bien aquellos modos que induzcan un momento dipolar diferente de cero en la molécula, son activos en el infrarrojo. Para que exista un espectro rotacional activo en el infrarrojo se requiere que la molécula sea polar (momento dipolar neto no nulo), mientras que para tener un espectro vibracional activo en el infrarrojo basta que el movimiento vibracional de los átomos de la molécula induzca un momento dipolar no nulo.

Por esta razón, las moléculas diatómicas como N_2, O_2 o H_2, así como las moléculas de los gases nobles, no presentan espectros de absorción en el infrarrojo. Sin embargo, y como veremos a continuación, no es el caso del dióxido del carbono.

2.3. Cálculo de la emisividad de un gas. Ley de Lambert-Beer

La ley de Lambert-Beer expresa cuantitativamente la absorción de radiación en un medio. La cantidad de radiación absorbida está relacionada con la estructura de niveles de la molécula, con la concentración del gas absorbente (Pa), y con la longitud del camino óptico recorrido (Z).

La figura 2.3 muestra un esquema de absorción donde a una nube de gases llega una señal de cierta intensidad I_0 y tras pasar a través de la nube con un camino óptico (Z) la señal se ha visto modificada y convertida a una nueva intensidad I.

Para la absorción de luz monocromática por parte de una molécula de un gas de la atmósfera, la ley de Lambert-Beer vendrá dada por la expresión

$$I(\overline{\nu}, L) = I_0 \; exp[-\alpha(\overline{\nu}) \cdot Pa \cdot Z] \tag{2.3}$$

donde $\alpha(\nu)$ es el coeficiente de absorción del gas y Pa es la concentración de dicho gas.

Por tanto, la transmitancia τ de un medio para un valor dado de número de onda

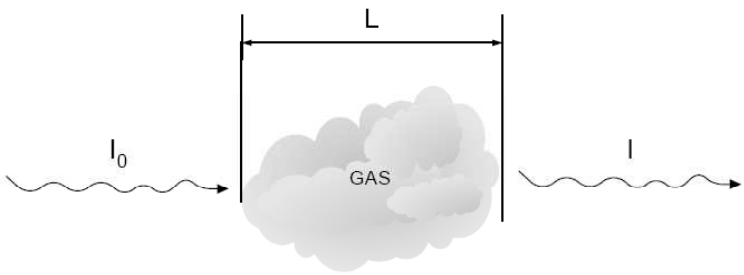

Figura 2.3: Esquema de absorción electromagnética al pasar por un medio.

se define como

$$\mathrm{I}(\overline{\nu}) = \frac{\mathrm{I}(\overline{\nu}, \mathrm{Z})}{\mathrm{I}_0} = exp[-\alpha(\overline{\nu}) \cdot Pa \cdot Z] \tag{2.4}$$

Vamos a ver cómo se relaciona para un gas el coeficiente transmitancia con el coeficiente emisividad. Cuando un cuerpo interacciona con radiación electromagnética, ese cuerpo puede absorber, transmitir o reflejar dicha radiación. La ley de conservación de la energía se puede enunciar como

$$\alpha + \rho + \tau = 1 \tag{2.5}$$

donde α, ρ y τ son los coeficientes de absorción, reflexión y transmisión.

Esta ecuación nos dice que el flujo de energía incidente es igual al flujo absorbido por la superficie, más el flujo reflejado por la superficie, más el flujo transmitido por la superficie. Si estamos en condiciones de equilibrio termodinámico, se verifica además la llamada ley de Kirchhoff, que queda reflejada en la siguiente ecuación

$$\epsilon = \alpha \tag{2.6}$$

y que relaciona la emisividad ϵ y el coeficiente de absorción α. Para un cuerpo negro ideal, $\rho = \tau = 0$ y $\alpha = 1$. Sin embargo, para un gas tendremos que $\rho = 0$ con lo que $\alpha + \tau = 1$. Teniendo en cuanta la ley de Kirchhoff, la emisividad de un gas viene dada por

$$\epsilon(\overline{\nu}) = 1 - \tau(\overline{\nu}) \ . \tag{2.7}$$

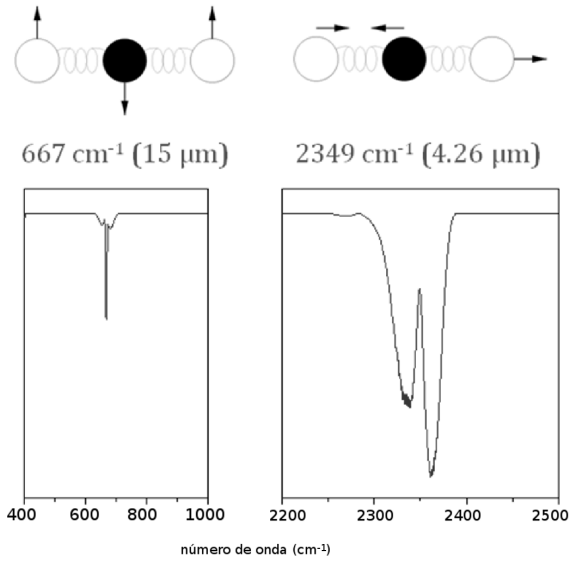

Figura 2.4: Bandas de transmitancia del CO_2.

Dióxido de carbono

El dióxido de carbono, CO_2, es el principal componente gaseoso producto de una combustión de combustibles fósiles. Se produce principalmente durante la fase de llama de la combustión, donde las altas temperaturas alcanzadas permiten la completa oxidación del carbono. Para un combustible orgánico, que contiene carbono e hidrógeno en su composición, la reacción general de combustión viene dada por

$$C_xH_yO_z + nO_2 \rightarrow xCO_2 + \frac{y}{2}H_2O + energía$$

La molécula CO_2 es una molécula lineal, no polar. Su espectro de absorción presenta sólo bandas vibracionales, que se presentan en forma de un espectro muy compacto. Cabe destacar la banda centrada en 2347 cm^{-1}(4,26 μm), que es la banda de absorción más intensa. Otra banda muy importante es la banda centrada en 667 cm^{-1}(15 μm), ya que es la responsable de la contribución del CO_2 al efecto invernadero (ver figura 2.4)

2.4. Ecuación de Transferencia Radiativa

Cuando la fuente emisora es un medio homogéneo, con una temperatura dada y una concentración de un gas dada, el cálculo de la radiancia emitida es sencillo, y viene dado por

$$\mathrm{R_{gas}} = (1 - \tau_{\mathrm{gas}})\mathrm{B} \tag{2.8}$$

donde τ_{gas} es la transmitancia del gas y depende del número de onda, de la temperatura, concentración del gas y de la longitud de la nube del gas (camino óptico). B es la ley de Planck, y depende del número de onda y de la temperatura.

Sin embargo, una llama es un medio fuertemente inhomogéneo, con gradientes de temperatura y concentración en su interior. En este caso, la ecuación que nos da la radiancia emitida por esa nube gaseosa es la bien conocida ecuación de transferencia radiativa

$$\mathrm{R_{gas}} = \int_0^Z B(T)\frac{d\tau}{dz}dz \tag{2.9}$$

donde Z es la longitud total de la nube gaseosa, z localiza un punto en el interior de esa nube, y $\tau(z)$ nos da la transmitancia de la nube para un espesor z de la misma. El perfil de temperaturas $T(z)$ está implícitamente incluido tanto en la ley de Planck B como en la función $\frac{d\tau}{dz}$, mientras que el perfil de concentraciones $c_{gas}(z)$ está incluido en la función $\frac{d\tau}{dz}$. Por lo tanto, el espectro de energía emitido por la llama contiene información tanto del perfil de temperaturas como del perfil de concentración del gas.

En la figura 2.5 se puede observar la distribución espectral de la energía emitida por una llama. En él, pueden distinguirse las emisiones de varios gases producto de la combustión: H_2O (en las bandas centradas en 3700 y 1590 $\mathrm{cm^{-1}}$), CO_2 (en 3700, 2325 y 670 $\mathrm{cm^{-1}}$), CO (entre 2130 y 2200 $\mathrm{cm^{-1}}$), e hidrocarburos inquemados (en torno a 2900 $\mathrm{cm^{-1}}$). También se pueden observar las correspondientes absorciones de los gases atmosféricos H_2O y CO_2 en las mismas bandas. Por ejemplo, es fácil apreciar la absorción del CO_2 frío de la atmósfera por el efecto que causa en la banda de emisión del CO_2. Por este efecto, la banda queda dividida en dos picos, denominados en la literatura especializada 'pico rojo' (más intenso y centrado alrededor de 2250 $\mathrm{cm^{-1}}$) y el 'pico azul' (centrado en torno a 2390 $\mathrm{cm^{-1}}$). Como se observa en dicha figura, la

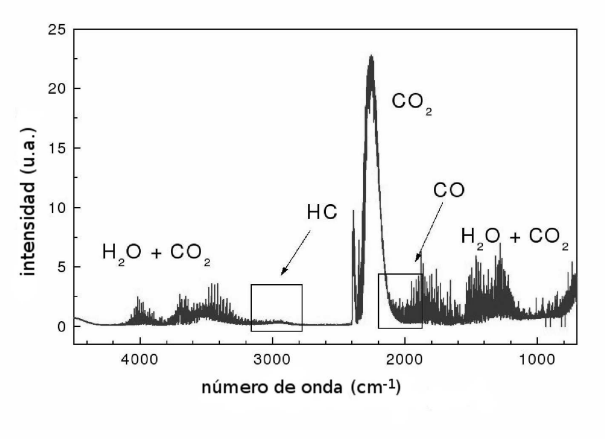

Figura 2.5: Bandas de emisión de los diferentes gases de una llama.

emisión asociada al CO_2 es con mucho la característica espectral más intensa en el espectro de emisión, y es precisamente la emisión de este gas la que se utilizará para la monitorización a distancia del proceso de combustión y reconstruir los perfiles de temperatura a partir de su información espectral.

2.5. Diseño de la experimentación

Para poder realizar una evaluación de las técnicas propuestas en esta tesis es necesario definir un conjunto de situaciones sintéticas que reflejan el comportamiento de combustiones típicas. Este conjunto de casos han sido simulados con la herramienta CASIMIR [44] la cual fue desarrollada en el laboratorio LIR de la Universidad Carlos III. CASIMIR es el acrónimo de Cálculos Atmosféricos para SIMulación de la transmitancia en el InfraRojo, y es una aplicación software que permite simular el efecto de absorción de los gases basándose en valores empíricos de la base de datos espectroscópica HITRAN/HITEMP [45] y en la ecuación de transferencia radiativa (véase Eq. 1.1).

Esta aplicación se basa en el concepto de linea-a-linea y calcula el espectro de radiancia a muy alta resolución. Un algoritmo de suavizado puede ser usado como postprocesado para simular espectros de menor resolución. CASIMIR necesita como

parámetros de entrada los mismos que han sido descritos para la ecuación RTE. Estos son: perfil de temperatura y la distribución espacial de las concentraciones de CO_2 y H_2O.

Descripción del conjunto de datos

La base de datos utilizada durante este estudio ha sido generado bajo las siguientes condiciones:

- El espectro sintético se corresponde con la radiancia emitida por una combustión de gases calientes (CO_2 y H_2O) de longitud Z. Los valores de temperatura y concentraciones asociados presentan gradientes a lo largo de la longitud.

- El rango espectral utilizado está localizado entre 2110cm^{-1} y 2410cm^{-1}. Este rango espectral corresponde a la banda de emisión del CO_2, que como se ha comentado es la característica espectral más importante en el espectro de emisión de una combustión. La resolución utilizada es alta con un paso de $\approx 0{,}13\text{cm}^{-1}$ lo que proporciona un descretización en 2341 números de onda, es decir $p = 2341$.

- Para poder manejar de una manera numérica los datos continuos referentes al perfil espacial de temperatura, se ha realizado una discretización de la longitud en celdas de igual tamaño. Cada una esas celdas tiene asociado un valor medio de temperatura y concentración de CO_2 y H_2O. El número total de celdas utilizadas para este estudio es de $z = 200$.

- El objetivo de este estudio está centrado en estudiar la dependencia de las diferentes distribuciones espectrales con respecto a los perfiles de temperatura. Por esta razón y por la correspondencia existente entre los perfiles de concentración de CO_2 y H_2O con los de la temperatura, los datos de concentración han sido seleccionados a partir de experimentos típicos de combustión.

- Se han realizado algunas asunciones iniciales para los perfiles de temperatura utilizados. Para una combustión de longitud z, los perfiles de temperatura

tienen siempre su valor máximo en el centro de la nube gaseosa y son simétricos con respecto a ese centro. Los perfiles se han simulado usando la siguiente parametrización

$$T(z) = t_{min} + \frac{t_{max} - T_{min}}{1 + exp^{\frac{(z-z_0)}{r}}}$$

donde t_{min} y t_{max} definen los valores mínimo y máximo de la temperatura de la llama, z_0 la distancia desde donde la temperatura empieza a disminuir y r la velocidad de decremento del perfil.

Con estos parámetros se han generado más de 3000 casos diferentes que han sido utilizados para realizar la experimentación.

Parte I

Selección de Características en Alta Dimensionalidad

Capítulo 3

Introducción

Normalmente, cuando el número de características p es lo suficientemente grande, existe un número de características l, siendo $l \ll p$ tal que l contiene virtualmente la misma información que la accesible en todas las p variables [31]. Por tanto, cuando se desea reducir la dimensionalidad de un conjunto de datos, uno de los parámetros que hay que determinar es ese número l de características a seleccionar, y decidir cual de los subconjuntos de l características es el mejor.

La selección de características, también nombrada en la literatura estadística como selección de variables o selección de subconjuntos, trata de realizar la selección óptima de ese subconjunto de características originales l de acuerdo a una función objetivo.

La primera parte de esta tesis investiga como introducir la información física dentro de una técnica de selección de características basada en el análisis de componentes principales. Para ello, en la sección 3.1 se hace una introducción a las diferentes técnicas de selección de características y su taxonomía. La sección 3.2 introduce brevemente la técnica de análisis de componentes principales y los conceptos relacionados necesarios para entender el algoritmo desarrollado. La sección 3.3 presenta los términos físicos asociados a la aplicación específica de esta tesis que van a permitir guiar la selección de características y evitar seleccionar características redundantes e irrelevantes. En el capítulo 4 se describe en detalle el algoritmo de selección de máximos locales y, el algoritmo desarrollado de selección de características haciendo uso del conocimiento sobre el problema. El capítulo 5 muestra los experimentos realizados y

sus resultados.

3.1. Técnicas de selección de características

La selección de características puede ser realizada de una manera binaria (incluyendo o no una característica) o se puede realizar una ponderación (normalmente entre 0 y 1) para indicar la importancia individual de cada característica. La mayoría de los estudios relacionados sobre la selección de características pertenecen a problemas de clasificación [21, 27, 28] y de regresión [30].

Dependiendo de si el proceso de selección se produce antes, paralelamente o simultáneamente al aprendizaje, nos encontramos con tres tipos de modelos: si es antes del aprendizaje se denomina una aproximación tipo filtro (p.e. RELIEF [46]), si se utiliza paralelamente para evaluar la selección realizada haciendo una llamada al algoritmo de aprendizaje, se llama de tipo envoltorio (–*wrapper approach*- [21],p.e. usando algoritmos genéticos para la selección [47]), y si el método de aprendizaje se incluye como parte de su algoritmia, la selección entonces será de tipo embebido (p.e. LASSO [32] y MARS [48]).

Independientemente del tipo de técnica utilizada, es necesario definir el modo de evaluación de la bondad de una característica. La información mutua, definida como $I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p_x(x)p_y(y)}$ donde $p_x(x)$ y $p_y(y)$ son las distribuciones de probabilidad marginales de x e y respectivamente, es un término muy utilizado debido a que la relación entre las entradas y las salidas debe ser alta [49]. El coeficiente de correlación de Pearson, definido como $\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$ siendo cov la covarianza entre X e Y, y σ las desviaciones estándar respectivas, es también otro de los valores utilizados frecuentemente para evaluar las características [22, 50].

Como se ha descrito en el apartado 1.3.2, los conjuntos de datos que poseen alta dimensionalidad presentan dos retos fundamentales para los problemas de regresión. Primero, en espacios de alta dimensionalidad la existencia de multicolinealidad aumenta debido a la presencia de características irrelevantes o ruido. En [51, 52] los investigadores han notado que en presencia de un gran número de variables (cientos o miles) es común que una gran parte de éstas no aporte ninguna información debido a

su irrelevancia o redundancia y por tanto que exista multicolinealidad. En problemas de regresión los efectos de la multicolinealidad van a verse reflejados en los coeficientes del modelo regresor. Estos coeficientes pasan a tener valores grandes, lo que implica que son vagamente estimados y difícilmente el modelo se va a comportar de manera adecuada frente a nuevos casos. El comportamiento del modelo únicamente será correcto cuando la nueva predicción se haga en la región del espacio donde ocurre la multicolinealidad. El segundo reto está relacionado con la dispersión de los datos. En conjuntos de datos de alta dimensionalidad, los datos tienden a estar separados unos de otros provocando así lo que comúnmente se denomina "la maldición de la dimensionalidad" [25].

La mejor manera de aliviar este problema es aumentar el número de casos disponibles de nuestra base de datos, pero en numerosas ocasiones esto no es posible debido al alto coste de etiquetado, escasez de datos reales, necesidad de implicación humana (p.e. datos médicos), etc. Recientemente en [53], han propuesto una técnica basada en la aplicación de PCA de modo supervisado para paliar las consecuencias de la alta dimensionalidad y su dispersión en el espacio. Su desarrollo está pensado para una aplicación a datos de *microarray*, y para ello realizan de manera recursiva una reducción de dimensiones tipo PCA con aquellas características originales que están más relacionadas con las salidas.

Algunas soluciones propuestas para reducir el problema de la multicolinealidad tratan de regularizar los coeficientes del modelo estimado penalizando de esta manera los coeficientes grandes. La técnica *Ridge Regression* [54] y la más reciente *LASSO* [32] son dos de las más conocidas - véase [55, 56] y sus referencias-. Estás técnicas no discriminan sobre qué variables se realiza la regularización y por tanto, aplica dicha penalización a todos los ejemplos por igual, afectando tanto a los ejemplos mejor estimados como a los peores. Esto provoca una regularización global que empeora los resultados en las estimaciones más favorables y mejora para los menos favorables. Esto es debido a la multicolinealidad y a la falta de homogeneidad.

La mayoría de estas aproximaciones son de tipo filtro o de tipo embebido. Esto se debe a que la aproximación tipo *wrapper* necesita buscar en el espacio de soluciones el mejor subconjunto de características de acuerdo al criterio de "relevancia"

establecido. Normalmente este criterio carece de monoticidad (con respecto a las características) y es necesario realizar una búsqueda combinatoria en el espacio de todos los subconjuntos de características. Esto implica, que aún sabiendo el número de características buscadas t de un conjunto con u elementos, serán necesarias $q = \binom{u}{t} = \frac{u!}{(u-t)!t!}$ evaluaciones para encontrar el subconjunto óptimo. Este número de evaluaciones es inviable en conjuntos de datos de alta dimensionalidad, y es necesario realizar asunciones heurísticas como una selección hacia delante o hacia atrás, *floating search*, u optimización con genéticos entre otras propuestas [21, 57, 58, 59].

En el contexto de regresión, la búsqueda secuencial hacia delante es conocida como *stepwise regression*. En muchos casos la inclusión o eliminación sistemática de características logra obtener modelos exitosos aunque no óptimos [60, 61]; pero si existe una multicolinealidad severa, provocará gran inestabilidad en los modelos creados. Otra desventaja de un modelo *stepwise regression* se debe a que la selección se basa en un *ranking* de características individuales, y no permite considerar sus inter-relaciones.

En esta investigación vamos a explorar un nuevo método no supervisado de selección de características en datos de alta dimensionalidad. Este método es específico de la aplicación tratada ya que se utilizará el conocimiento sobre el problema como información a priori del método. La selección de un subconjunto original de características reducirá los problemas de dispersión de los datos, mientras que la introducción de conocimiento a priori servirá para eliminar las características redundantes seleccionando aquellas que sean representativas de las diferentes cualidades físicas. Además la selección de un subconjunto de variables originales puede ser interpretado posteriormente desde un punto físico y proporcionar un mejor conocimiento sobre la aplicación. Otros beneficios de encontrar un subconjunto de características originales son la reducción del coste computacional evitando realizar cálculos sobre características irrelevantes, y reducir el coste de sensores en el caso de la teledetección.

3.2. Análisis de componentes principales (ACP o PCA)

PCA es una técnica lineal de reducción de dimensionalidad, que intenta mantener la variación presente en un conjunto de datos. Para ello se realiza una transformación a un nuevo conjunto de variables, las componentes principales (PCs), las cuales están decorreladas y ordenadas, siendo las primeras componentes las que contengan la mayor parte de la variación presente en el conjunto inicial de las variables originales.

PCA también tiene una interpretación geométrica además de la puramente estadística. La interpretación geométrica de la primera PC se corresponde con el nuevo eje de coordenadas que maximiza la variación de las proyecciones de los datos originales en el nuevo eje de coordenadas. La figura 3.2 muestra un conjunto de puntos representados en un espacio bidimensional (x_1 y x_2) y su correspondiente nueva base formada por los ejes z_1 y z_2. El eje z_1 se corresponde con el eje mayor imaginario de la elipse formada por el conjunto de puntos originales, mientras que el eje z_2 forma una base ortogonal con respecto a z_1.

Sea $\mathbf{x}_i \in \Re^{p\times 1}$ (ver notación[1]) un vector correspondiente a las diferentes longitudes de onda de un espectro de energía, y $\mathbf{X} \in \Re^{p\times m}$ su correspondiente matriz con m ejemplos. PCA es una transformación lineal de las diferentes longitudes de onda que maximiza la varianza.

Sea $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^T = \sum_i \mathbf{x}_i\mathbf{x}_i^T \in \Re^{p\times p}$ la matriz de covarianza de $\mathbf{X}$, y $\boldsymbol{\alpha}_k$ un vector columna con los coeficientes de la transformación, $\boldsymbol{\alpha}_k^T = (\alpha_{k1}, \alpha_{k2}, \cdots, \alpha_{kp})$. Para obtener la primera PC, tenemos que encontrar el valor $\boldsymbol{\alpha}_1$ que maximice $var(\sum_{j=1}^{p} \alpha_{1j}\mathbf{x}_j) = \boldsymbol{\alpha}_1^T\mathbf{\Sigma}\boldsymbol{\alpha}_1$, bajo la restricción de $\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 = 1$.

Se puede demostrar que $\boldsymbol{\alpha}_1$ es el autovector correspondiente al mayor autovalor, Λ_1, de $\mathbf{\Sigma}$, y que $var(z_1) = \Lambda_1$ [31]. En general, la k-th PC, $\mathbf{z}_k = \sum_{j=1}^{p} \alpha_{jk}\mathbf{x}_j$, puede

[1]Las letras mayúsculas en negrita denotan matrices $\mathbf{D}$, las letras minúsculas en negrita un vector columna $\mathbf{d}$. $\mathbf{d}_j$ representa el elemento columna j^{th} de la matriz $\mathbf{D}$. d_{ij} define un escalar en la fila i y la columna j de la matriz $\mathbf{D}$ y el elemento escalar i-th del vector columna $\mathbf{d}_j$. Todas las letras no negritas denotan variables escalares. $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$ define la norma euclídea de $\mathbf{x}$.

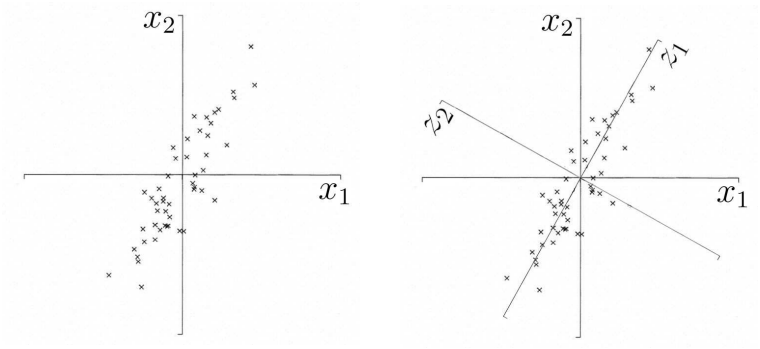

Figura 3.1: Nuevos ejes obtenidos por PCA.

ser obtenida maximizando la siguiente expresión

$$\sum_{i=1}^{n} ||\boldsymbol{\alpha}^T \mathbf{x}_i||_2^2 = ||\boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}||_F \ ,$$

con la restricción $\boldsymbol{\alpha}^T \boldsymbol{\alpha} = \mathbf{I}$, donde $\boldsymbol{\Sigma} = \boldsymbol{\alpha} \boldsymbol{\Lambda} \boldsymbol{\alpha}^T$ siendo $\boldsymbol{\alpha}_k$ el kth autovector o k-PC, y $\boldsymbol{\Lambda}_k$ su autovalor [31].

En muchas aplicaciones reales el objetivo principal es preservar la mayor parte de la variación en los datos, y para lograrlo se seleccionan las primeras dimensiones de las proyecciones $\mathbf{Z}$ en la nueva base obtenida del análisis PCA .

Algunos ejemplos de aplicaciones que utilizan estás proyecciones como variables de entrada a sus algoritmos pueden ser el reconocimiento de caras [62], o en [63] se utiliza para identificar materiales dentro de mezclas siendo las entradas espectros de alta resolución al igual que en nuestra aplicación. Este segundo ejemplo estudia como afecta la reducción de dimensionalidad con PCA al utilizarla junto con técnicas de clasificación como SVM.

3.3. Funciones de peso y su importancia en teledetección

Las funciones de peso proporcionan información sobre en qué zonas de la nube de gases existe una mayor o menor absorción para una determinada longitud de onda.

Esta mayor o menor absorción se puede interpretar como información asociada a cada característica o longitud de onda.

La ecuación 1.1 puede ser escrita como

$$R_{\overline{\nu}} = \int_0^Z B_{\overline{\nu}}\{T(z)\}K_{\overline{\nu}}(z)dz \tag{3.1}$$

donde $K_{\overline{\nu}}(z) = \frac{d\tau_{\overline{\nu}}(z)}{dz}dz$ es la función de peso [64, 65], y pondera la función de Planck $B_{\overline{\nu}}$ para la zona de emisión de radiación correspondiente z_i. Como la función de peso es la derivada del perfil de transmitancia, proporciona información espacial de qué zona de la nube gaseosa contribuye de manera más importante a la emisión para un número de onda determinado. Debido a esto, una selección cuidadosa de un conjunto de longitudes de onda puede ser elegida para disponer de información sensible a las diferentes zonas de la combustión.

Para entender mejor de una manera cualitativa por qué la funciones de peso contienen este tipo de forma podemos considerar la emisión de radiación de una combustión hacia el instrumento de medida o sensor, como una serie consecutiva de celdas a diferentes profundidades y de volumen la unidad. Entonces, la radiación emitida está determinada por estos tres factores:

- la temperatura de cada una de las celdas, que es la variable que queremos estimar.
- el número de moléculas de los gases (principalmente CO_2 en combustiones), que puede ser asumido como constante y conocido.
- la transmitancia de los gases y la atmósfera desde la celda hasta el sensor.

Esto se muestra en la figura 3.2[2] para un caso de teledetección atmosférica y para tres celdas a diferentes profundidades. Para la celda más alejada, la densidad atmosférica es alta y por eso la radiación emitida también lo es, pero la mayoría es absorbida por las celdas posteriores y una parte muy pequeña alcanza el sensor. Para la zona más cercana, ocurre lo contrario. La transmitancia es alta pero comparativamente

[2]Esta imagen ha sido tomada de [65].

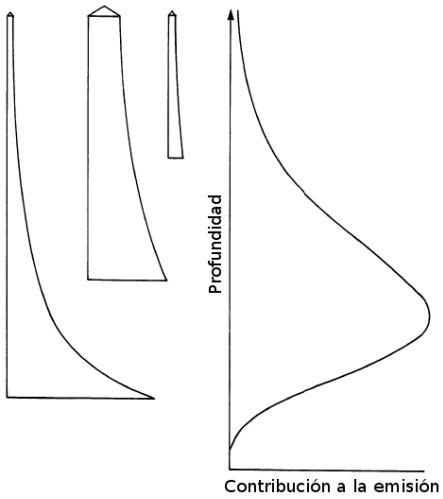

Figura 3.2: Izquierda, muestra la atenuación de la radiación emitida desde tres profundidades diferentes. Derecha, el perfil de contribución total a la emisión recibida en el sensor.

la emisión es muy pequeña porque la densidad disminuye exponencialmente con la profundidad. De este modo, existirá un punto intermedio donde la contribución a la emisión total que recibe el sensor sea máxima. La parte derecha de la figura 3.2 muestra cual es la contribución de cada celda a la radiación emitida para una longitud de onda específica. El punto más alto de la curva indica la zona de mayor influencia para esa longitud de onda.

Así, sabemos que la mayoría de la información contenida en una longitud de onda determinada pertenece a un conjunto de celdas específico. De ese modo se pueden seleccionar un conjunto de longitudes de onda específicos que permitan reconstruir la información de temperaturas de la atmósfera o en nuestro caso de una combustión.

La figura 3.3 muestra las funciones de peso para cada una de las longitudes de onda de un conjunto de combustiones generadas sintéticamente. Estas funciones de peso mostradas son la media aritmética de todos los datos generados. Es decir, cada función de peso está asociada a un caso y a una determinada longitud de onda. Como nuestro objetivo es predecir un conjunto de casos variado, la función de peso final

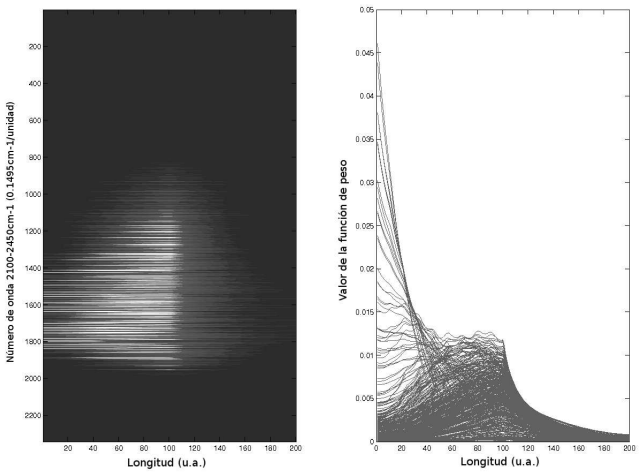

Figura 3.3: Izq. Valor de la función de peso de una combustión vista en 2D (las zonas claras indican más importancia). Dcha. Valor de la función de peso para diferentes longitudes de onda.

que se muestra en la figura se corresponde con la media aritmética de todos los casos disponibles. La imagen de la izquierda muestra con diferentes grados de luminosidad la importancia de las diferentes longitudes de onda (eje Y) para las distintas celdas a lo largo del espacio (eje X). La imagen de la derecha muestra esta misma información pero con una vista lateral, donde cada curva se corresponde con una función de peso.

Además, hay que tener en cuenta que las funciones de peso dan información sobre un rango de profundidades (un conjunto de celdas) y no sobre un punto o celda específica. Esto limita la capacidad de poder realizar estimaciones muy precisas. También, las funciones de peso se superponen unas con otras y, como consecuencia, aunque se realicen mediciones a p longitudes de onda diferentes, únicamente se obtendrán l características de información independiente.

Finalmente, añadir que para casos con temperaturas muy diferentes, las funciones de peso pueden variar mucho, y por tanto no es trivial la selección de un conjunto óptimo cuando el problema a tratar contiene una gran diversidad de perfiles. Esto provoca que no pueda utilizarse directamente este tipo de información y haya que

buscar otras alternativas. En el siguiente capítulo vamos a describir nuestra propuesta donde se introduce de una manera aproximada este tipo de información junto con la reducción de dimensionalidad por selección de características.

Capítulo 4

Selección guiada de características

Es mucho mejor dar una respuesta aproximada a la pregunta correcta,
que a menudo es vaga, que una respuesta exacta a la pregunta equivocada,
la cual siempre puede ser precisa.
–J. W. Tukey (1915-2000)

A priori la utilización de datos de alta dimensionalidad debería permitir una mejor aproximación en un problema de regresión. Por el contrario, trabajar en problemas que poseen alta dimensionalidad conlleva desventajas debido a la multicolinealidad, la redundancia y la irrelevancia de características como se mencionó en el apartado 1.3.2.

Para beneficiarnos tanto de la técnica de reducción de dimensionalidad PCA como de la información física conocida a priori, nosotros vamos a desarrollar una nueva aproximación para seleccionar características combinando ambos elementos. Esta aproximación está basada en el estudio de los coeficientes de los autovectores correspondientes al análisis PCA sobre los datos de entrada.

Debido a que no se hace uso de los datos de salida durante ninguno de los pasos de la propuesta, podemos considerarla como una técnica de selección de características no supervisada. De este modo aplicaremos PCA al conjunto de datos de entrada y posteriormente haremos una selección de características guiada examinando los coeficientes de las PCs.

El primero de los dos pasos va a aliviar los problemas relacionados con la multicolinealidad de los datos. Esto se debe a que los diferentes autovectores correspondientes

al análisis PCA están decorrelados y por tanto el estudio de los coeficientes correspondientes a distintos autovectores también contendrán información intrínsecamente decorrelada.

El segundo paso hará uso de la información conocida a priori sobre el problema para realizar una selección guiada sobre los coeficientes de los diferentes autovectores. Este proceso logra, de una manera natural e intuitiva, evitar la información redundante correspondiente a coeficientes en los autovectores con valores similares

A continuación se explica en más detalle tanto la importancia de los coeficientes de los autovectores correspondientes a PCA, como la introducción de la información física haciendo uso de esos coeficientes.

4.1. Estudio de los coeficientes del análisis PCA e introducción de la información física

Si examinamos las proyecciones $\mathbf{Z}$ de los datos originales sobre los coeficientes de los autovectores/PCs $\boldsymbol{\alpha}$ correspondientes al análisis PCA, obtenemos que la $k-$ésima componente se define como $\mathbf{z}_k = \sum_{j=1}^{p} \alpha_{jk}\mathbf{x}_j \in \Re^{p\times 1}$, donde $\mathbf{x}_j \in \Re^{p\times m}$ es el vector original de datos y $\boldsymbol{\alpha} \in \Re^{j\times k}$ contiene en sus vectores columna los autovectores correspondientes a la nueva base. Recordar que los autovectores están decorrelados y por tanto $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = \mathbf{I}$. Recordar también que el análisis PCA tiene la propiedad de que intenta maximizar la "dispersión" de los datos en el nuevo espacio $\boldsymbol{\alpha}$ de menor dimensión, lo que a efectos prácticos significa que trata de mantener los puntos en el espacio transformado tan alejados como sea posible y por lo tanto manteniendo la variación del espacio original.

En nuestro caso nos vamos a basar en los coeficientes de los autovectores $\boldsymbol{\alpha}$ en vez de utilizar las proyecciones $\mathbf{Z}$ para buscar aquellas características originales más influyentes y lograr así una selección de características eficiente.

Si estudiamos estos coeficientes $\boldsymbol{\alpha}$, se observa que un valor alto del coeficiente i-ésimo correspondiente al autovector $\boldsymbol{\alpha}_k$ implica que la característica $\mathbf{x}_i^T$ de $\mathbf{X}$ es muy dominante en ese eje o autovector. En general, seleccionando las características

correspondientes a los coeficientes más altos de cada uno de los primeros l autovectores se puede aproximar la misma proyección que la obtenida por PCA.

Para realizar una selección de características utilizando estos coeficientes es necesario definir el criterio de relevancia junto con su modo de aplicación, y también el número de características que serán seleccionadas.

El criterio de relevancia será el valor absoluto de los coeficientes de los autovectores, y existen principalmente tres métodos base de su aplicación para realizar una selección de características no supervisada [31]:

1. Asociar una característica con cada con cada uno de los últimos autovectores l_1 eliminarlas. Esto se puede realizar de una sola vez o de manera iterativa. En este último caso se realiza PCA en las $l_1^* = (p - l_1)$ características restantes, y se elimina un segundo subconjunto de características l_2, y así sucesivamente. El razonamiento de este método se basa en que los autovalores pequeños se corresponden con relaciones casi constantes entre un subconjunto de características. Por lo tanto, si una de las características que interviene en esa relación es eliminada no se perderá mucha información (es fácil de ver que se eliminará aquella con el mayor valor absoluto en el correspondiente autovector).

2. Asociar un conjunto de l^* características con los últimos l autovectores, y eliminar esas características. Un criterio para evaluar cada característica es maximizar la suma de los cuadrados de los coeficientes en los últimos l autovectores, pero esté método no ha tenido resultados satisfactorios de acuerdo a los estudios realizados en [66, 67].

3. Asociar una característica con cada uno de los primeros autovectores, seleccionando la característica con el mayor coeficiente en valor absoluto y que no haya sido seleccionada previamente. De este modo, se seleccionan l características y las restantes $l^* = p - l$ son eliminadas. Este método es complementario al descrito en el punto 1, y además selecciona una única característica para grupos que están altamente correlados disminuyendo la selección de características redundantes.

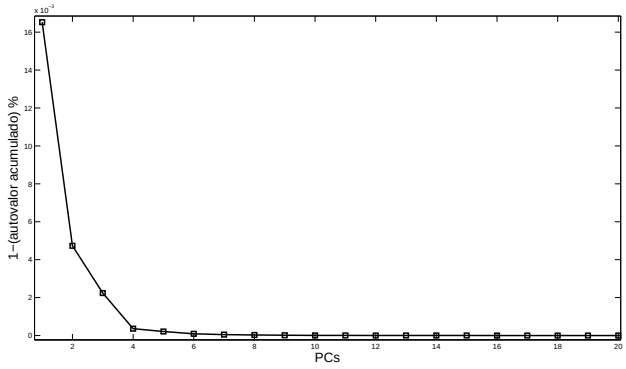

Figura 4.1: Representación "scree-graph"para la matriz de correlaciones.

A priori, tanto los métodos de eliminación de las últimas características asociadas a los últimos autovectores como los de selección de aquellas características asociadas a los primeros l autovectores poseen similares características desde un punto de vista teórico. En la práctica, debido al gran número de características utilizadas, los métodos basados en la eliminación de característica asociadas a los últimos autovectores se presentan más complejos y costosos en tiempo. Por tanto, nuestro método va a basarse en una selección de un conjunto de características correspondientes a un número k de las primeras PCs. De este modo queda definido tanto el concepto de relevancia de una características, como la manera de evaluar ese criterio de relevancia para las diferentes características del conjunto de datos.

El segundo problema planteado surge de la necesidad de determinar el número k de primeras PCs sobre el cual se desea aplicar el criterio de relevancia arriba explicado. En este caso vamos a poder aplicar un criterio similar al utilizado por la técnica PCA para lograr mantener máxima varianza de los datos utilizando únicamente las primeras k proyecciones.

Este criterio está basado en examinar la variación total acumulada por las primeras k componentes principales y establecer un valor umbral deseado, usualmente alrededor del 90 %. Entonces el número de componentes principales utilizadas será el menor número k tal que su varianza acumulada total exceda ese porcentaje. Una variación de este criterio está basado en una visualización de está varianza acumulada

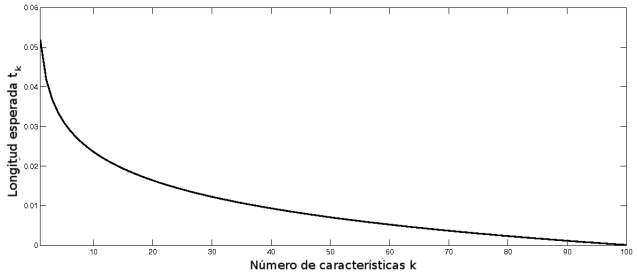

Figura 4.2: Función del modelo *broken stick model* para 100 características.

en una gráfica bi-dimensional denominada "scree-graph". La figura 4.1 muestra la representación de nuestra aplicación para la matriz de correlaciones de los datos de entrada. En ella puede verse que existe alrededor de la componente número 5 lo que se denomina "el bajo codo".

Otro criterio que extiende de un modo genérico este mismo concepto aplicado de manera visual, es el denominado modelo *broken stick*. Este modelo dice que, si se dispone de una vara con longitud la unidad y se parte en p segmentos, entonces puede ser demostrado que la longitud esperada en el k segmento más largo es

$$t_k = \frac{1}{p} \sum_{j=k}^{p} \frac{1}{j} \ .$$

Para decidir cuantas PCs seleccionar, bastaría con comparar la proporción de varianza que contiene el k autovector y ver si es mayor que el valor de t_k dependiendo del valor k correspondiente al número de PCs a seleccionar. La figura 4.1 muestra los valores correspondientes para t_k de la función descrita.

Se observa que inicialmente los valores son altos pero disminuyen rápidamente al aumentar el número de componentes k, es decir, para un pequeño número de características es muy probable que sus valores estén por encima de está función pero después está función umbral tiende a estabilizarse y será mayor que los valores de varianza las subsiguientes componentes principales. Se mantienen todas aquellas PCs cuyo valor es mayor, y el resto se eliminan.

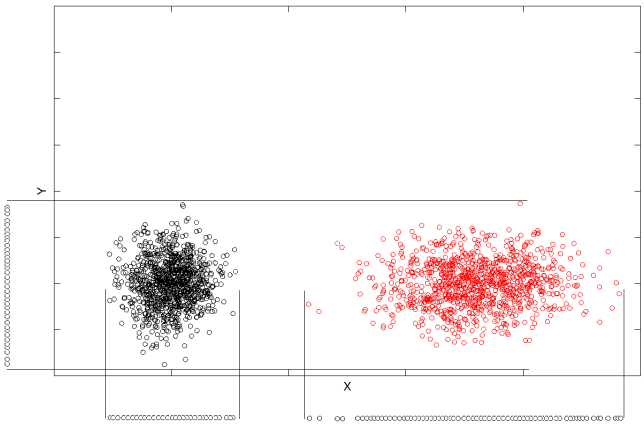

Figura 4.3: Ejemplo donde la característica y es irrelevante porque si omitimos x perdemos la información relativa a los dos grupos.

Debido a la multicolinealidad existente entre los datos, el número de PCs seleccionadas siguiendo los criterios descritos es muy bajo (aprox. 4-5). Con ese valor para k tendríamos únicamente cuatro características seleccionadas de entre las miles de posibles lo cual parece ser una reducción demasiado restrictiva y posiblemente se pierda capacidad de reconstrucción en la aplicación real de este trabajo. Empíricamente se ha comprobado que los resultados al estimar la función buscada con ese número de características están muy alejados de los deseables.

Una posible solución es aumentar el valor de k, aumentando así el número de características seleccionadas, donde cada una de ellas está asociada a una PC. Esta aproximación tiene un problema principal y es que a medida que nos acercamos a aquellas PCs de menor valor en varianza, cada vez sus características asociadas contienen menos variabilidad y por tanto su inclusión en el conjunto de características seleccionadas se hace inocuo según nos acercamos al número total de PCs existentes, que es igual al número de características originales. Incluso puede ser que algunas de estas características sean ruido o totalmente irrelevantes. Por tanto esa aproximación es una aproximación ciega y no aporta una mejora cualitativa a la metodología.

Nosotros vamos a realizar una selección guiada donde utilizaremos únicamente las primeras PCs pero vamos a incluir un buscador de máximos locales, tratando de encontrar aquellos coeficientes más relevantes en un vecindario determinado.

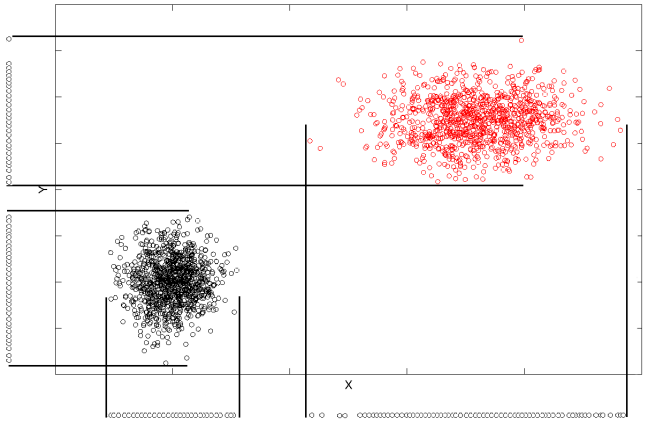

Figura 4.4: En este ejemplo las características x e y son redundantes debido a que ambas características proporcionan el mismo poder discriminante para los grupos.

Antes de poder explicar el algoritmo detallado de este modo de selección de características utilizando los coeficientes de PCA, vamos a introducir la implicación que tienen los coeficientes de las PCs y su relación con los dos problemas planteados en la sección 1.3.2 que a su vez están relacionados con dos de los problemas que ocurren en alta dimensionalidad: la redundancia y la irrelevancia de características.

La figura 4.3 muestra un ejemplo de característica irrelevante. Este tipo de características deben ser evitadas ya que pueden provocar el empeoramiento del proceso de aprendizaje. Si examinamos cómo afectan este tipo de características a los coeficientes de las PCs obtenidas por el análisis PCA en un contexto de alta dimensionalidad se observa que estas características $\mathbf{x}_i^T$ tenderán a ser constantes y cercanas a cero, y por tanto sus coeficientes asociados para las distintas componentes principales serán $\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik} \approx 0$.

Por otro lado, si examinamos la figura 4.4 se observa que la información proporcionada por las variables x e y es similar y por tanto son características redundantes. También podemos decir que la variable x puede ser obtenida a partir de y tal que $y = \alpha \cdot x$, y viceversa. Cuando existen características redundantes los coeficientes de las PCs obtenidas serán muy similares tal que, $\alpha_{ik} \approx \alpha_{jk}$ siendo $\mathbf{z}_k = \sum_{j=1}^{p} \alpha_{jk}\mathbf{x}_j$ la proyección de los datos en las PCs. Por lo tanto, durante la selección de características son deseables los siguientes criterios:

- **Criterio de relevancia**: serán relevantes aquellos valores absolutos $\boldsymbol{\alpha}_k$ correspondientes a las primeras k PCs, tal que sean distintos de 0 y lo más grandes posibles en valor absoluto.

- **Criterio de no redundancia**: se considerarán no redundantes todos los valores absolutos de $\boldsymbol{\alpha}_k$ correspondientes a cada una de las primeras k PCs, tal que dichos valores no sean similares.

Hasta ahora, hemos abordado los problemas relacionados con el número de características a seleccionar, el criterio para hacerlo, y el modo de evitar la inclusión de características redundantes. Todos estos puntos han sido tratados desde un punto de vista puramente estadístico y hemos propuesto la utilización de los coeficientes de las primeras PCs para evitar o paliar dos de los problemas presentes en alta dimensionalidad.

El último punto a tratar para poder completar la propuesta realizada está relacionado con el modo de incorporar la información física descrita en la sección 3.3 al método de selección guiada de características. Como se comentó en dicha sección, las diferentes características asociadas a cada uno de los canales del espectro (números de onda) están relacionadas de algún modo con las diferentes profundidades en la nube gaseosa del perfil de temperatura que se desea reconstruir. Recordar que la figura 3.3 mostraba las funciones de peso típicas de un proceso de combustión, donde es conocido que las zonas adyacentes x_{im} y $x_{(i+1)m}$ del espectro de energía normalmente llevan asociadas información similar o redundante desde un punto de vista físico. El reto ahora es cómo incorporar este conocimiento sobre características redundantes desde un punto de vista físico, y simultáneamente mantener el criterio que hemos establecido para eliminar las características redundantes en el proceso de selección.

En la figura 4.5 está representada la primera PC $\boldsymbol{\alpha}_1$ correspondiente al análisis PCA aplicado a los datos de entrada $\mathbf{X}$. Es decir, los valores del eje de la Y se corresponde con los coeficientes de la PC para cada una de las características de entrada $\mathbf{x}_i^T$. Si únicamente se utiliza el criterio de no redundancia como restricción para seleccionar o no una característica, entonces se seleccionará una característica de cada una de las franjas horizontales que se visualizan. Esto es debido a que, como

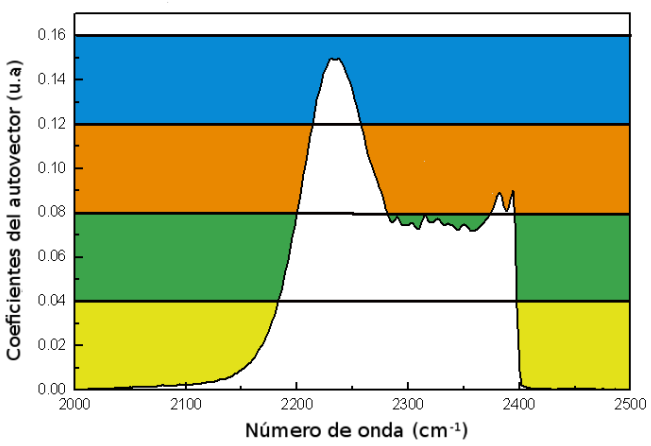

Figura 4.5: Selección de características sobre los coeficientes de una PC usando agrupamiento k-means.

se comentó anteriormente, los coeficientes de una misma PC con valores similares contienen información similar. Por tanto, para evitar la redundancia se buscarán aquellas características más representativas de cada una de las franjas horizontales. Un modo simple de poder lograr una selección de este tipo puede ser aplicar el algoritmo k-medias sobre los coeficientes de la PC y seleccionar los centroides.

Sobre el mismo tipo de representación, la figura 4.6 indica sobre los coeficientes de una PC la selección obtenida utilizando la información física conocida sobre el problema. A partir del estudio de las funciones de peso, se concluye que las zonas del espectro adyacentes contienen información similar, y por tanto es deseable elegir una característica para representar un rango espectral determinado y eliminar el resto de características que son redundantes. Esto se ve representado por las franjas verticales que separan unas características asociadas a una determinada zona del espectro de otras. Así, se seleccionan características asociadas a los diferentes rangos espectrales.

Nuestra propuesta mezcla ambas soluciones, el criterio de no redundancia y la información física, para eliminar la información redundante de una manera guiada. De este modo, sobre las primeras k PCs vamos a buscar no sólo el valor más alto correspondiente a cada una de ellas como se propuso en [31], sino los n mayores valores de los coeficientes, teniendo en cuenta que esos valores deben pertenecer a diferentes

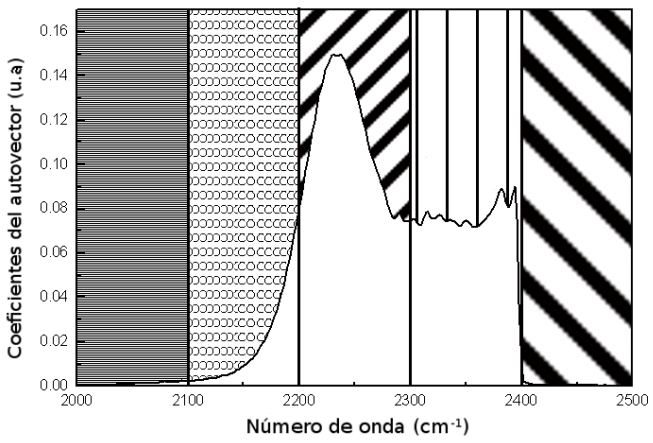

Figura 4.6: Selección de características sobre los coeficientes de las PCs agrupando por zonas de información física.

zonas del rango espectral. Por tanto, ahora redefinimos los criterios anteriormente descritos de relevancia, y de no redundancia para formalizar nuestra propuesta final:

- **Criterio de relevancia:** serán relevantes aquellas $n \cdot k$ características $\mathbf{x}_i^T$ tal que los n valores absolutos asociados a cada $\boldsymbol{\alpha}_k$ correspondiente a las primeras k PCs, sean distintos de 0 y lo más grandes posibles.

- **Criterio de no redundancia**: se considerarán no redundantes aquellas características $\mathbf{x}_i^T$ tal que los valores absolutos asociados en $\boldsymbol{\alpha}_k$ correspondientes a cada una de las primeras k PCs no sean similares.

- **Criterio de no redundancia física:** se consideran no redundantes todas aquellas características $\mathbf{x}_i^T$ tal que no exista otra característica seleccionada en el vecindario $\mathbf{x}_{i-r}^T - \mathbf{x}_{i+r}^T$ de dicha característica.

Los criterios segundo y tercero pueden ser juntados y simplificados como sigue:

- **Criterio guiado de no redundancia:** se consideran no redundantes aquellas características $\mathbf{x}_i^T$ tal que los valores absolutos asociados en $\boldsymbol{\alpha}_k$ correspondientes a cada una de las primeras k PCs no sean similares, y pertenezcan a vecindarios espectrales distintos tal que si $\mathbf{x}_i^T$ es una característica ya seleccionada, la nueva

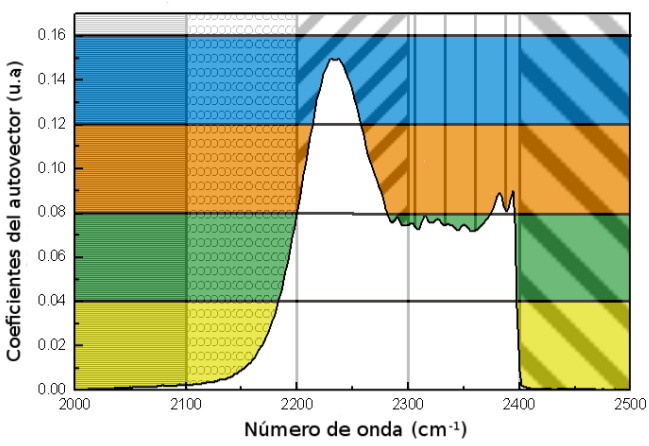

Figura 4.7: Solución propuesta para mezclar los criterios de no redundancia, relevancia y la información física.

selección $\mathbf{x}_i^{*T}$ no puede pertenecer al entorno de $\mathbf{x}_{i-r}^T - \mathbf{x}_{i+r}^T$ siendo r el rango de vecindad.

La figura 4.7 muestra este nuevo criterio guiado de no redundancia junto con el criterio de relevancia ya descrito. Cada uno de los cuadros del mallado representa dicha mezcla. De esta manera la selección de características vendrá guiada por un nuevo criterio de "relevancia guiada" que buscará, dentro de cada una de estas zonas del mallado, los coeficientes más grandes en valor absoluto. Para este fin, se ha desarrollado un algoritmo para buscar máximos locales que se ha denominado algoritmo de selección de picos [68].

El algoritmo de selección de picos va a buscar las características $\mathbf{x}_i^T$ que tengan un mayor valor absoluto en cada uno de los autovectores $\boldsymbol{\alpha}_k$, pero al mismo tiempo realiza esta búsqueda en un entorno local para evitar la selección de características adyacentes. Con esta mezcla de información estadística y física se espera que la selección de características sea más apropiada y se mejoren los resultados obtenidos en la estimación frente a otro tipo de selección, o frente a la utilización de todas las características originales.

4.2. Selector de picos en las componentes principales

El algoritmo de selección de picos se basa en la definición de una ventana deslizante que se mueve a lo largo de las diferentes características (rangos espectrales) para ir buscando los máximos locales. La ventana posee una propiedad que la determina que es su longitud. Está longitud de ventana es el equivalente a definir el parámetro descrito como rango de vecindad r en el criterio guiado de no redundancia. En la figura 4.8 puede verse dibujada esta ventana sobre uno de los máximos locales de una función sintética.

Esta ventana deslizante recorre las diferentes características $\mathbf{x}_i^T$ de cada una de las PCs seleccionadas en busca de máximos locales cumpliendo así tanto con el criterio de relevancia, valores altos de los coeficientes de las PCs, como con el criterio guiado de no redundancia evitando valores similares y cercanos en vecindad.

La descripción detallada del algoritmo se muestra en el algoritmo 1. Como entradas del algoritmo es necesario definir el tamaño horizontal de la ventana (v_anch), el número k de PCs que sobre las que se realizará la selección de características y la matriz con las PCs ($\boldsymbol{\alpha}_k$), correspondientes a los datos de entrada $\mathbf{X}$. El tamaño horizontal de la ventana está relacionado con el rango de vecindad r. Decir que el tamaño vertical de la ventana será igual al horizontal y por tanto se trata de una ventana deslizante cuadrada. Entonces, el algoritmo recorre de manera iterativa las distintas características $\mathbf{x}_i^T : \ 1 < i < p$, y cuando la distancia euclídea entre el máximo valor dentro de la ventana (max_local) y el último mínimo encontrado (min_global) sea mayor que un valor umbral definido (v_umbral) entonces se clasifica ese punto como máximo local añadiéndolo al conjunto de características seleccionadas (FS). Una vez se realiza la búsqueda de los máximos locales para las diferentes k PCs se devuelve un vector con todas las características seleccionadas.

La figura 4.8 muestra un ejemplo del resultado de aplicar el algoritmo de selección de picos a una función tipo $y = abs(sin(x) - 0{,}5)$. En la imagen, los puntos de tipo cuadrado negro se corresponden con los picos encontrados, mientras que los símbolos

Algoritmo 1 Selección de máximos locales

Entrada: v_anch ← tamaño horizontal de ventana deslizante
v_umbral ← valor umbral de máximo local
k-PCs ← número de PCs
p ← número de características totales
$\boldsymbol{\alpha}$ ← matriz de autovectores

```
1: FS ← ∅
2: for i = 1 : 1 : k-PCs do
3:    min_global = 0;
4:    for j = v_anch : 1 : p - v_anch  do
5:       min_local ← MIN(α_(j−v_anch : j+v_anch)i)
6:       if min_local < min_global then
7:          min_global = min_local
8:       end if
9:       if  (α_ji − min_global) > v_umbral)  then
10:         FS ← FS ∪ j
11:         min_global = α_ji
12:      end if
13:   end for
14: end for
15: Devolver  FS
```

en forma de cruz se corresponden con la discretización de la función. Como puede observarse el algoritmo reconoce perfectamente todos los máximos tanto globales como locales.

Algoritmo de selección guiada de características

Haciendo uso del algoritmo de búsqueda de máximos locales descrito, podemos finalizar la descripción del procedimiento de selección guiada de características propuesto en este trabajo. Este procedimiento puede resumirse en los siguientes pasos:

- Paso 1 Calcular la matriz de covarianza $\boldsymbol{\Sigma}$ de las entradas $\mathbf{X}$, $\boldsymbol{\Sigma} = \mathbf{X}\mathbf{X}^T$. En algunos casos puede ser más óptimo el cálculo de la matriz de correlaciones en vez de la matriz de covarianzas [31]. Esto es recomendable cuando las características tienen distintas escalas o pertenecen a unidades medidas diferentes.

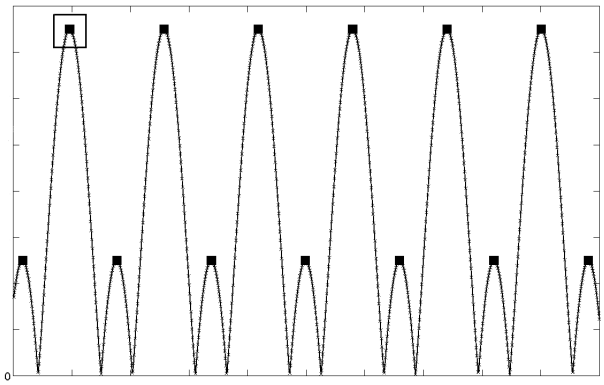

Figura 4.8: Ventana deslizante y máximos locales obtenidos por el algoritmo selección de picos para una función $y = abs(sin(x) - 0{,}5)$

- Paso 2 Calcular los autovectores y los autovalores de la matriz de covarianzas/correlaciones $\mathbf{\Sigma} = \boldsymbol{\alpha}\Lambda\boldsymbol{\alpha}^T$ utilizando la técnica PCA.
- Paso 3 Seleccionar el número de dimensiones k al que se quieren reducir los datos y construir la matriz $\boldsymbol{\alpha}_{1..k}$ a partir de $\boldsymbol{\alpha}$.
- Paso 4 Llamar al algoritmo de selección de máximos locales con la matriz $\boldsymbol{\alpha}_k$, v_umbral, p, k, y el tamaño de ventana v_anch.

Recordar que vamos a utilizar esta selección guiada de características para un problema de reconstrucción de perfiles de temperatura a partir de los datos de su espectro de energía. Entonces, una vez se han obtenido el subconjunto de características originales (FS) con este procedimiento, se construye un regresor utilizando únicamente esas características. Como se indica en el siguiente capítulo, en este trabajo hemos utilizado un regresor tipo perceptron multicapa.

Capítulo 5

Experimentos

Si buscas resultados distintos, no hagas siempre lo mismo.
–Albert Einstein (1879-1955)

Los experimentos realizados van dirigidos a mostrar la mejora de resultados obtenidos al utilizar la técnica propuesta de selección de características relevantes. Para ello se va a realizar una comparativa en el entorno de reconstrucción de temperaturas a partir de los espectros de emisión de una combustión típica.

La comparativa se basa en evaluar los resultados para estas tres situaciones:

- No se realiza ninguna selección previa y se utilizan todas las características originales para realizar la estimación. Recordar que el número de características originales es $p = 2341$.

- Método de selección B4 [31]. Se seleccionan las características originales correspondientes a los coeficientes más grandes de las primeras k PCs. Con este criterio se han seleccionado 15, 30, 45, 75 y 105 características. Mencionar que aunque el número de características sean éstas, ha sido necesario en algunos casos utilizar más PCs debido a la duplicidad de características asociadas a distintas PCs. Este método de selección ha sido elegido porque es el más conocido y el más exitoso dentro de las aproximaciones de selección de características basadas en los coeficientes de las PCs.

- Selección guiada de características utilizando nuestra aproximación basada en el algoritmo de selección máximos locales. Al igual que en la situación anterior, se han seleccionado un total de 15, 30, 45, 75 y 100 características originales. En este caso el número de PCs utilizadas ha sido de quince. Este número ha sido elegido en base a la cantidad de varianza contenida en las primeras PCs y ajustado de manera heurística con respecto al error obtenido por la estimación final del problema.

5.1. Red de neuronas

Las redes de neuronas pueden ser usadas entre otras aplicaciones, en problemas de regresión para ajustar datos experimentales. Debido a la no-linealidad del modelo inverso que estamos tratando, se ha usado una red de neuronal tipo perceptron multicapa (MLP) que posee la característica de poder realizar ajustes no lineales.

Perceptron multicapa

Ha sido demostrado que el perceptron multicapa es un un aproximador universal [69, 70], en el sentido de que cualquier función continua sobre un compacto de $\Re^n$ puede aproximarse con un perceptron multicapa, con al menos una capa oculta. Su capacidad de aprender a partir de ejemplos, aproximar funciones no lineales, filtrar ruido, etc. hace que sea un modelo adecuado para abordar problemas reales [71].

Un modelo tipo MLP está basado en tres capas: entrada, oculta y la de salida. Cada una de las capas está constituida por un conjunto de neuronas que reciben sus entradas de la capa inmediatamente anterior, y envían los valores de salida a la capa directamente siguiente. La capa de entrada se corresponde con los datos introducidos en el modelo y no realiza ningún tipo de procesamiento, únicamente transmite esos datos a la capa oculta. La salida de las otras neuronas del modelo, capa oculta y capa de salida, dependen de una función f, de las entradas ponderadas, y de un valor umbral. La salida de una neurona i será

$$out_i = f\Big(\sum_{j=1}^{u} w_{ij}x_j + b_i\Big) \tag{5.1}$$

donde x_j, w_{ij} y b_i son, respectivamente, los valores de entrada, las ponderaciones de los enlaces y el umbral asociado a la neurona i, y u es el número de neuronas de la capa anterior. La función f se denomina función de activación y la más común es la función sigmoidal, la cual se corresponde con la siguiente expresión

$$f(x) = \frac{1}{1 + e^{-x}} \ . \tag{5.2}$$

Los valores de salida out_i, son las entradas de la capa siguiente al nodo evaluado y el proceso se repite hasta obtener los valores correspondientes a las neuronas de la capa de salida.

Los valores de los pesos para cada uno de los enlaces entre las diferentes neuronas de diferentes capas tiene que ser determinado para establecer el modelo final. Este proceso de ajuste es denominado proceso de entrenamiento de la red. Para ello es necesario utilizar los ejemplos disponibles del problema a resolver. La mayoría de las reglas de aprendizaje son formuladas con un objetivo específico, p.e. mover de manera iterativa la posición de un vector hasta lograr una posición que minimiza o maximiza una función de coste particular.

Un MLP utiliza un aprendizaje supervisado que ajusta los parámetros de la red para minimizar el error entre la salida de la red y_k y la salida deseada $\hat{y}_k$ para los k-th ejemplos

$$\text{error} = \frac{1}{m}\sum_{k=1}^{m}(y_k - \hat{y}_k) \tag{5.3}$$

siendo m el número de ejemplos de entrenamiento. El método que generalmente se utiliza para el aprendizaje de los parámetros del modelo MLP es el algoritmo de retropropagación. Este algoritmo propaga hacia las capas anteriores la diferencia entre el valor de salida deseado y el obtenido por la ultima capa del MLP.

Este tipo de modelo ha sido utilizado en problemas de teledetección atmosférica [72, 73] por lo que se intuye que su uso en nuestra aplicación puede ser muy apropiada.

Queremos resaltar que, además de los problemas relacionados con la alta dimensionalidad ya explicados, para un modelo concreto como en este caso un MLP, surgen problemas de computabilidad cuando se utilizan muchas entradas. Esto se debe al aumento de la complejidad del modelo, y la consecuente gran cantidad de cálculos necesarios para realizar una iteración en el algoritmo de aprendizaje. Esto es sin duda una gran desventaja que puede hacer inviable su utilización para casos donde existen miles de entradas. La selección de un subconjunto de estas características evita también este tipo de problemas.

5.2. Experimentos y comparativa

Para medir el rendimiento de los diferentes criterios de selección de características hemos utilizado el error medio absoluto (MAE), calculado como,

$$\mathrm{MAE} = \frac{1}{z}\frac{1}{m}\sum_{k=1}^{z}\sum_{j=1}^{m}|\mathbf{y}_{kj} - \hat{\mathbf{y}}_{kj}| \tag{5.4}$$

donde z es la discretización de la longitud, m el número de ejemplos, $\mathbf{y}$ es el valor real de la temperatura, e $\hat{\mathbf{y}}$ es el valor obtenido por el modelo estimador. El MAE proporciona información sobre el error numérico y el error físico que se está cometiendo en la estimación de cada caso. Por tanto, este valor de error es medido en unidades de temperatura (K). Otra de las medidas que vamos a utilizar para medir el error es la desviación estándar (SD) de las salidas con respecto a la media.

Los experimentos realizados se corresponden con tres aproximaciones y van encaminados a mostrar que el método propuesto mejora los resultados obtenidos frente a la utilización de todas las características, o frente a otro método de selección no supervisado que también se basa en los coeficientes de las PCs como es el método B4.

Aproximación inicial: sin selección de características.

Los resultados obtenidos utilizando todas las características (2341) son mostrados en la última fila de la tabla 5.1. Para realizar las pruebas de test hemos separado el conjunto de datos en dos subconjuntos, uno para el entrenamiento de la red neuronal y el otro se utilizará únicamente para obtener los errores de test. Para el entrenamiento de la red se ha utilizado el método *cross-validation* con un valor de 10.

Algunos valores de referencia en la literatura muestran que errores del 1-5 % pueden ser considerados como válidos en este tipo de reconstrucciones de temperaturas [11, 14, 40, 41, 42]. En [74] los investigadores también han obtenido errores relativos por debajo del 1 % en el rango de 1280-1690 K, utilizando un sistema digital de imagen.

Los resultados obtenidos sin realizar una selección previa de características y utilizando un MLP como modelo estimador son de ≈ 23 K de error. Este error es $> 5\,\%$ y está por encima de los valores de referencia. Por lo tanto esta aproximación no puede ser considerada como válida.

Selección utilizando el criterio B4

Para esta segunda prueba se ha seleccionado las características utilizando el criterio de selección B4, donde se selecciona una característica por cada una de las primeras k PCs eligiendo la característica con el coeficiente más grande [67, 75].

Se ha aplicado esta selección sobras las primeras k PCs para lograr seleccionar 15, 30, 45, 75 y 105 características para realizar diferentes pruebas. Posteriormente, se han utilizado estos subconjuntos de características como entradas a un MLP para realizar las estimaciones.

Los resultados obtenidos pueden verse en la primera fila de la tabla 5.1. Estos resultados mejoran notablemente los obtenidos anteriormente sin realizar selección de características. Ahora los errores son ≈ 6 K frente a los ≈ 23 K que obteníamos antes. Estos resultados están en un error relativo de $\approx 1 - 1{,}5\,\%$ lo cual puede ser considerado como válido para algunas aplicaciones [11, 14, 40, 41, 42], pero aún existen otros métodos que alcanzan mejores resultados $< 1\,\%$ [74].

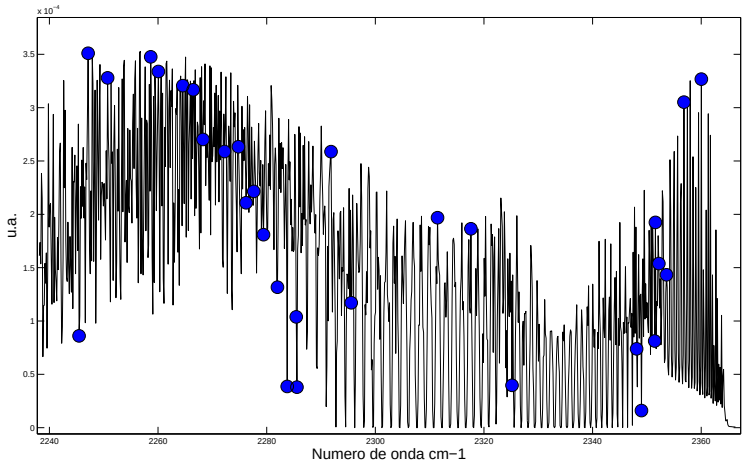

Figura 5.1: Características seleccionadas (circulos azules) usando el método (SG) para una reducción a 30 características.

Selección guiada usando el algoritmo propuesto

Para nuestra aproximación hemos usado el propuesto algoritmo de selección de máximos locales para elegir, igual que para método anterior, un subconjunto de 15, 30, 45 ,75 y 105 características. Estas características también van a ser usadas como entradas para un MLP el cual proporciona como salidas los valores de temperatura del perfil de una llama.

Un total de 15 autovectores han sido utilizados para realizar la selección y las características resultantes de esta selección son mostradas en la figura 5.1. Los errores asociados a los resultados obtenidos utilizando esta selección pueden verse en la segunda fila de la tabla 5.1. Claramente las estimaciones obtenidas mejoran las dos anteriores. El error obtenido es $\approx$ 4 K obteniendo el mejor resultado para 45 características con un error de 3.72 K. Estos resultados mejoran en 2 K los resultados obtenidos por el método B4 y en más de 18 K los obtenidos utilizando todas las características. Igualmente la desviación estándar de los resultados obtenidos es $\approx$ 2,5 K que también mejora los resultados de las otras dos aproximaciones.

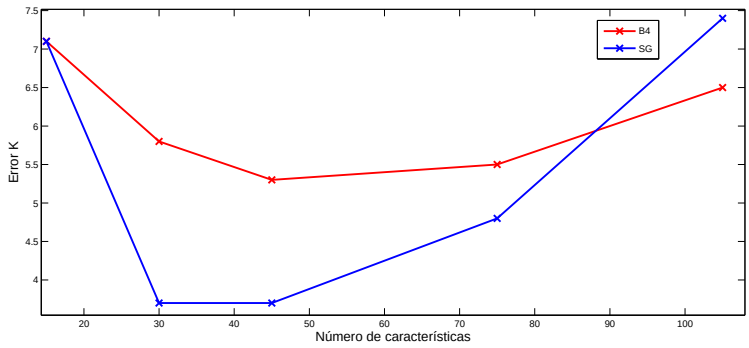

Figura 5.2: Errores obtenidos por el método B4 y la selección guiada (SG) para diferentes números de características.

Estos errores son $< 1\,\%$ lo cual puede considerarse como una precisión de reconstrucción muy alta. Cabe destacar los buenos resultados obtenidos con 30 características. Estos resultados son similares a los obtenidos con 45 características, lo cual significa que se ha realizado una reducción óptima de las características durante el proceso de selección que hemos utilizado.

La figura 5.2 muestra los errores obtenidos, tanto para el método B4 como para el método propuesto (SG), para los diferentes números de características seleccionadas. Se observa en la figura que a partir de 45 características la estimación en vez de mejorar empeora, y sucede igualmente para ambos métodos. Esto significa que las nuevas características añadidas no son relevantes para realizar la estimación, y posiblemente contienen ruido o son redundantes.

Además, queremos destacar que estos resultados reafirman lo dicho en el apartado 1.3.2 cuando decíamos de que un aumento del número de características (alta dimensionalidad) no siempre lleva asociada una mejora, sino que puede empeorar el rendimiento del proceso de aprendizaje.

5.3. Conclusiones

En esta primera parte, hemos estudiado cómo realizar una selección no supervisada de características para datos de alta dimensionalidad e introducir información

Cuadro 5.1: Errores obtenidos para los métodos de selección de características B4, nuestra propuesta y un modelo sin selección de características

Método de selección	Número de características	Temperatura Test (MAEs/SD)K	Temperatura Entrenamiento (MAEs/SD)K
B4	15	7.10/4.48	7.02/4.37
	30	5.80/4.26	5.78/4.11
	45	5.23/3.99	5.15/3.87
	75	5.48/3.93	5.36/3.77
	105	6.50/4.56	6.40/4.40
Selección guiada	15	7.10/4.48	7.02/4.37
	30	3.73/2.58	3.67/2.47
	45	3.72/2.48	3.66/3.03
	75	4.84/3.22	4.73/2.41
	105	7.42/5.26	7.38/5.18
Usando todas las características	2341	22.56/17.18	22.45/17.25

previa del problema durante el proceso. Para ello, hemos propuesto un nuevo método para seleccionar características que aprovecha la estructura de las componentes principales junto con el conocimiento previo sobre el problema para mejorar el modelo de estimación en una aplicación de teledetección. El análisis de componentes principales retiene la mayor parte de la información desde un punto de vista de las características representadas en un espacio de menor dimensión, y desde un punto de vista de minimización del error de reconstrucción. Por otro lado el conocimiento sobre el problema ha permitido dispersar la selección de las características reduciendo la redundancia en dicha selección.

Con el algoritmo propuesto se han abordado los tres problemas presentes en los datos de alta dimensionalidad:

1. Las características que son irrelevantes van a tener coeficientes en las PCs cercanos a 0 por lo que se evitan buscando aquellos coeficientes más grandes.

2. Las características redundantes tienden a tener valores similares en las PCs por lo que la introducción del conocimiento físico del problema va a dispersar la selección logrando así reducir el efecto de la redundancia en los datos.

3. En consecuencia, la multicolinealidad se va a ver reducida debido a los dos puntos ya comentados, y porque al utilizar un análisis de tipo PCA, las nuevas

PCs decorrelan los datos. Por tanto, la selección sobre sus coeficientes también ayuda a mantener esa decorrelación.

Los resultados obtenidos demuestran que nuestro algoritmo de selección es capaz de realizar una reducción de dimensionalidad seleccionando las características relevantes. Estos resultados mejoran los obtenidos previamente y son considerados como aceptables para muchas aplicaciones de teledetección.

Además, la selección de características originales tiene la ventaja de que son necesarios un menor número de cálculos para realizar la estimación o para el almacenamiento de los datos. En el caso de diseño de instrumentación, esto permite un menor número de sensores con su consecuentes beneficios tanto en coste como en complejidad. Además, la posibilidad de interpretar la selección obtenida es de gran ayuda en la mejora del conocimiento del problema tratado.

Desde el punto de vista físico podemos considerar como aceptables los resultados obtenidos con nuestra propuesta de selección guiada de características y puede concluirse que el uso de una red de neuronas tipo MLP junto con el algoritmo propuesto pueden ser aplicados para estimar perfiles de temperatura de llamas.

Parte II

Análisis Discriminante Basado en Zonas Homogéneas en Datos de Alta Dimensionalidad

Capítulo 6

Introducción

Cuando trabajo en un problema, nunca pienso en la belleza.
Sólo pienso en como resolver el problema. Pero si cuando he acabado,
la solución no es bella, entonces sé que está mal.
–R. Buckminster Fuller (1895-1983)

La extracción de características crea un conjunto pequeño de nuevas características realizando una transformación general a partir de los datos de alta dimensionalidad. Las técnicas utilizadas pueden ser supervisadas o no supervisadas. Nos referimos a técnicas no supervisadas cuando la extracción de características se realiza utilizando únicamente la matriz de datos de entrada y sin usar la información de salida.

6.1. Aproximaciones existentes y limitaciones

Aunque en la primera solución aportada en este trabajo se ha utilizado para realizar una selección de características, el análisis de componentes principales también denominado como la transformación Karhunen-Loeve o simplemente transformación KL, es posiblemente el método de extracción de características más conocido(véase sección 3.2). Se trata de una técnica no supervisada, no hace uso de las salidas, y su propósito principal es reducir la dimensionalidad manteniendo la variación en los datos y minimizando el error de reconstrucción desde un punto de visto de mínimos cuadrados.

Las nuevas características extraídas por PCA serán la proyección de los datos originales sobre las diferentes componentes principales. En muchas ocasiones el objetivo principal es preservar la mayor parte de la variación en los datos, y para lograrlo se seleccionan las primeras dimensiones de los datos proyectados en la nueva base obtenida del análisis PCA. Aunque posee desventajas, su simplicidad y potencia han hecho que sea una herramienta muy utilizada y sea considerada un estándar para reducir la dimensionalidad.

Hay muchas aplicaciones que utilizan estás proyecciones como variables de entrada a sus algoritmos. En teledetección de atmósfera ha sido utilizada de manera satisfactoria para reconstruir perfiles de temperatura usando redes de neuronas como modelo de estimación, y las proyecciones resultantes del análisis PCA como entradas a la red [72, 73]. Es común referirse a este tipo de configuraciones como PCR (*principal component regression*) donde las nuevas características son utilizadas con fines de regresión. En [63] se utiliza para identificar materiales dentro de mezclas, y las entradas son espectros de alta resolución al igual que en nuestra aplicación. En ese trabajo se estudia como afecta la reducción de dimensionalidad realizada con PCA al utilizarla junto con técnicas de clasificación como las máquinas de vectores de soporte (SVMs).

Como se ha comentado, PCA es una técnica no supervisada y por tanto no establece ningún tipo de relación entre las entradas y las salidas. Esto hace que en muchas ocasiones la reducción obtenida no sea una representación válida del modelo buscado.

Una versión supervisada de PCA es la regresión con reducción de rango (RRR) que busca un espacio de menor dimensión cuando la matriz $\mathbf{X}\mathbf{X}^T$ es de rango deficiente y por tanto no invertible. Esta aproximación tiene en cuenta la salida pero desde un punto de vista de mínimos cuadrados y no busca relaciones internas entre los datos, ni estructuras.

En ciencias aplicadas, el descubrimiento de estructuras comunes entre dos dominios de un mismo problema puede ser muy interesante para entender la naturaleza de la relación existente entre ellos. El análisis de correlaciones canónico (CCA) [76] es un método de reducción estadístico multivariante usado para identificar y cuantificar la

correlación entre dos conjuntos de características. CCA encuentra el conjunto de vectores base para dos conjuntos de variables tal que, la correlación de sus proyecciones en esas bases sea mutuamente maximizada. Este método, aunque es común en otras disciplinas como la psicología, su divulgación en el campo del aprendizaje automático es relativamente reciente. Además, tanto las técnicas PCA como CCA pueden ser fácilmente extendidas a su versión no-lineal gracias al denominado *kernel-trick* originando así las técnicas Kernel-PCA (KPCA) y Kernel-CCA (KCCA).

En [19] los investigadores han realizado un estudio de estás técnicas descritas y sus versiones no lineales KPCA y KCCA para el problema tratado en esta tesis. Las conclusiones muestran que aunque estás técnicas atenúan algunos de los problemas debidos a la alta dimensionalidad y a la multicolinealidad, se observa una gran variación en los resultados en los diferentes tipos de datos. Esto es un indicio de que la multicolinealidad en los datos sigue presente y por tanto es necesario algún otro tratamiento para lograr resultados más precisos.

Una aproximación que trata de abordar este problema realizando una agrupación de los datos en base a su correlación es descrita en [77]. Ese trabajo denominado *correlational spectral clustering* hace uso de la generalización de CCA en su forma de kernel-CCA y de ese modo trata de buscar dos subespacios nuevos, una para las entradas ψ_x y otra para las salidas ψ_y, de tal modo que los datos estén máximamente correlados en esos nuevos subespacios. Posteriormente busca agrupaciones de las proyecciones en el nuevo espacio de entrada usando un algoritmo de k-medias. Aunque este método encuentra grupos de datos homogéneos, únicamente tiene en cuenta la correlación entre los nuevos datos proyectados y no trata de mantener de ningún modo la distancia del espacio original de los datos. Además no busca un único espacio que contenga las dos representaciones de un objeto sino que busca espacios diferentes para las entradas ψ_x y las salidas ψ_y.

En resumen, tanto CCA, su extensión KCCA, o el propuesto método *correlational spectral clustering* buscan estructuras comunes en los datos, tanto en la entrada como en la salida, pero carecen de de algo importante para nuestra aplicación que es la importancia de la localidad de las relaciones entre los datos. Es decir, aquellos datos que en su espacio original se encuentran más cercanos unos de otros utilizando una

métrica euclídea tienen una relación mayor entre ellos.

Como puede observarse, la reducción de dimensionalidad es un tema ampliamente tratado tanto en las áreas de inteligencia artificial, como en minería de datos; pero, la idea de mantener la información local y buscar desde un punto de vista geométrico una la estructura de un espacio de menor dimensión que mantenga dicha información, no ha sido tratada hasta recientemente.

Las técnicas *local linear embedding* LLE [35] e ISOMAP [36] son dos estudios que se interesan por la búsqueda de esa estructura en un subespacio de menor dimensión donde posiblemente residan los datos de alta dimensionalidad. Estas técnicas se basan en la linealidad local de los datos para poder reconstruir el comportamiento de un de ellos a partir de su relación con otros datos similares (localmente cercanos de acuerdo a la métrica utilizada). De este modo logran reproducir comportamientos no lineales, y se proponen como alternativas a PCA o al escalado multi-dimensional (MDS [78]).

En [79] se desarrolla un nuevo algoritmo que busca también la estructura intrínseca de los datos pero utiliza una solución basada en conceptos bien conocidos de teoría de grafos (ver p.e. [34]). En ese mismo estudio también se establecen relaciones con métodos antes utilizados en visión por computador basados en *spectral clustering* [80].

En esta investigación hemos propuesto un modelo que busca estructuras comunes entre dos conjuntos de datos que pertenecen a distintas medidas de un mismo objeto. Para ello, vamos a seguir el mismo enfoque de las técnicas LLE e ISOMAP y de ese modo buscar conjuntos de datos homogéneos que contengan información similar en los espacios de entrada y de salida. Esto va a permitir un modelado del problema teniendo en cuenta las diferentes estructuras encontradas tanto en la salida como en la entrada.

Resaltar que en [81] se explica la relación existente entre la utilización de una aproximación por grafos para la búsqueda de un subespacio de menor dimensión y el agrupamiento. Esto es debido a que de manera implícita el algoritmo basado en grafos enfatiza los grupos naturales contenidos en los datos. Por esto, la aproximación propuesta abre un nuevo modo de abordar problemas de regresión en alta dimensionalidad dado que la reducción de dimensionalidad propuesta da lugar a la búsqueda de manera implícita de agrupaciones de datos, que se corresponden con

distintos modelos. No existen muchos estudios en los que se intente realizar este tipo de agrupaciones en problemas de regresión. Por ejemplo en [82], los autores proponen un método heurístico e iterativo para agrupar las salidas, y posteriormente crear automáticamente sus estadísticas de primer y segundo orden del espacio de entrada. Entonces, se reasignan los ejemplos a los diferentes grupos creados a partir de la agrupación en las salidas. En dicho estudio no se habla en ningún momento de reducción de dimensiones en las salidas como modo creación de indicadores para la posterior agrupación, y tampoco se establece ningún requisito de localidad lineal siendo ambos requisitos necesarios en nuestra propuesta.

6.2. Grafos y terminología

Nuestra aproximación está basada en un entorno de grafos que va a permitir incorporar la información local de los datos. Haciendo uso del concepto de Laplaciano de un grafo, podemos calcular un espacio de baja dimensión que sea representativo de los datos y que preserva de manera óptima la información local [79]. Este tipo de análisis hace un uso explícito de las conexiones entre los vértices del grafo para poder interpretar la reducción de dimensiones de un modo geométrico. A continuación vamos a definir algunos términos necesarios para comprender mejor el desarrollo de nuestra propuesta.

Un grafo G consiste en un conjunto de vértices/nodos $V(G)$ y un conjunto de arcos/aristas $W(G)$, donde un arco es un par no ordenado de vértices distintos pertenecientes a G. Un par no ordenado se identifica como $\{i, j\} = \{j, i\}$. Si $\{i, j\}$ es un arco, entonces se dice que i y j son adyacentes, o que i es vecino de j, y se denota escribiendo $i \sim j$. En nuestro caso, este conjunto de arcos va a representar la similitud entre puntos, siendo $w_{ij} = w_{ji}$ la medida de similitud entre i y j, donde los vértices v_i y v_j representan los puntos del conjunto de datos. Así, la matriz $\mathbf{W}$ va a ser una matriz simétrica correspondiente a un grafo no dirigido donde los arcos entre dos vértices son bidireccionales. La figura 6.1 muestra un ejemplo de un grafo donde pueden verse los arcos y los vértices de un dodecaedro.

Para representar de una manera binaria las relaciones entre los distintos vértices

de un grafo se utiliza la matriz de adyacencia. La matriz de adyacencia $A(G)$ es una matriz de números enteros cuyas filas y columnas indexan los vértices de G, de modo que la entrada ij de $A(G)$ es igual al número de arcos desde i a j. La matriz $A(G)$ es simétrica, y los términos de la diagonal son ceros porque no existen bucles en un grafo. Si únicamente se dispone de la información de similitud w_{ij} entre los puntos de un conjunto de datos, entonces esta matriz $A(G)$ va a ser uno para todos sus elementos $i, j : i \neq j$. También puede utilizarse esta matriz de adyacencia para introducir información conocida sobre la localidad del problema como si se tratase de una máscara.

Por tanto, si $w_{ij} = 0$ indicará que los vértices i y j no están conectados, o de manera equivalente se puede decir que el grado de similitud entre los puntos i y j es mínimo. El grado del vértice $v_i \in V(G)$ se define como

$$deg(v_i) = \sum_{j=1}^{m} w_{ij} \ ,$$

y su matriz asociada $\mathbf{D}$ se define como una matriz diagonal con grado $deg(v_1), \ldots, deg(v_m)$ en la diagonal principal.

La matriz Laplaciana $L(G)$ es otro concepto importante y puede ser usado para encontrar muchas otras propiedades del grafo. Existen diversas definiciones de matriz Laplaciana en la literatura y para nuestro estudio se define como:

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \ .$$

Una de las propiedades importantes de esta definición de matriz Laplaciana es que es una matriz simétrica y semidefinida positiva. Una matriz $\mathbf{B}$ es semidefinida positiva si $\mathbf{u}^T\mathbf{B}\mathbf{u} \geq 0$ para cualquier vector $\mathbf{u}$. Además, si una matriz es positiva definida entonces todos sus autovalores son también positivos.

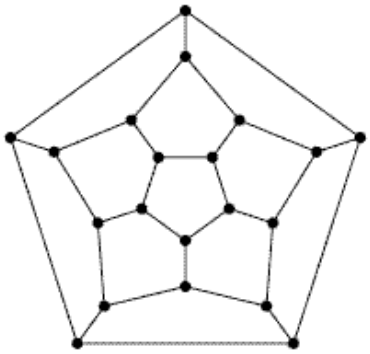

Figura 6.1: El grafo de un dodecaedro tiene 20 vértices y 30 arcos.

6.3. Una visión de grafos embebidos para reducir la dimensionalidad

Recordar que en la práctica cuando se dispone de datos de alta dimensionalidad, como es nuestro caso, es recomendable y beneficioso transformar estos datos originales a un espacio de menor dimensionalidad [83].

La tarea principal de la reducción de dimensionalidad es encontrar una función de transformación $F : x \rightarrow \hat{x}$ que transforme $\mathbf{x} \in \Re^p$ al espacio deseado de baja dimensión $\hat{x} \in \Re^{p'}$, siendo $p' \ll p$

$$\hat{x} = F(x) \ . \tag{6.1}$$

El mismo proceso es aplicable al espacio de salida $\mathbf{y} \in \Re^s$ para obtener $\hat{y} \in \Re^{s'}$ siendo $s' \ll s$. Esta función F puede ser lineal o no lineal para diferentes casos.

La búsqueda de dicha transformación F para reducir la dimensionalidad puede ser abordada desde un nuevo punto de vista basado en grafos embebidos [84].

Si introducimos el mismo problema de reducción desde un punto de visto de un grafo embebido tenemos que $G = (V, W)$ es un grafo no dirigido donde $V(G)$ son los vértices que se corresponden con los diferentes ejemplos disponibles $\{1, \ldots, m\}$, y los arcos $W(G)$ conectan los datos que están cercanos indicando su grado de similitud. Por tanto, si deseamos reducir la dimensión en el grafo G tenemos que considerar

una transformación de dicho grafo G a una línea tal que los nodos/puntos que estén conectados permanezcan tan cercanos como sea posible. Sea $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$ los nuevos datos en dicha transformación, entonces un criterio razonable para encontrar una "buena" transformación F es minimizar la siguiente función objetivo [79]

$$\sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 w_{ij} \tag{6.2}$$

bajo restricciones de ortogonalidad y escalado. Si examinamos esta función, puede verse que debido a los valores de la matriz de similitud $\mathbf{W}$ se induce una penalización cuando dos puntos vecinos x_i y x_j son proyectados lejos uno del otro. Por lo tanto, si minimizamos la expresión 6.2 estamos asegurando que si dos puntos x_i y x_j están cercanos en el espacio original, también lo estarán $\hat{x}_i$ y $\hat{x}_j$. Desarrollando la función con álgebra lineal puede verse que la solución viene dada por [81]

$$\begin{aligned} \frac{1}{2} \sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 w_{ij} &= \sum_{ij} (\hat{x}_i^2 + \hat{x}_j^2 - 2\hat{x}_i \hat{x}_j) w_{ij} \\ = \sum_{i} \hat{x}_i^2 d_{ii} + \sum_{j} \hat{x}_j^2 d_{jj} - 2 \sum_{ij} \hat{x}_i \hat{x}_j w_{ij} &= \hat{\mathbf{x}}^T \mathbf{L} \hat{\mathbf{x}} \end{aligned} \tag{6.3}$$

donde $\mathbf{L} = \mathbf{D} - \mathbf{W}$ es la matriz Laplaciana como se ha definido en el apartado anterior, y $\mathbf{D}$ es una matriz diagonal tal que $d_{ii} = \sum_j w_{ij}$.

Por tanto el problema de reducción de dimensionalidad se reduce a encontrar la transformación $\hat{\mathbf{x}}$ tal que:

$$\min_{\hat{\mathbf{x}}\mathbf{D}\hat{\mathbf{x}}=1} \hat{\mathbf{x}}^T \mathbf{L} \hat{\mathbf{x}} \tag{6.4}$$

imponiendo la restricción $\hat{\mathbf{x}}^T \mathbf{D} \hat{\mathbf{x}} = 1$ para evitar un escalado arbitrario. Esta ecuación puede ser resuelta como un problema de autovectores. De este modo la solución viene dada por la siguiente expresión

$$\mathbf{L}\hat{\mathbf{x}} = \Lambda \mathbf{D} \hat{\mathbf{x}} \ . \tag{6.5}$$

6.3.1. Estructuras locales y globales

La mayoría de los métodos tradicionales utilizan una aproximación global para realizar la reducción de la dimensionalidad. Así, PCA es óptimo para grupos de datos que estén distribuidos de una manera gaussiana, pero no es capaz de diferenciar datos que siguen otra distribución. Lo mismo sucede con el análisis lineal discriminante (LDA) [83, 85] que es el homólogo a PCA pero para problemas de clasificación.

Como se ha descrito hasta ahora, se puede considerar que los datos puede ser generados por un sistema estructurado que posiblemente contenga muchos menos grados de libertad que el que se presenta en las mediciones o espacio original. Usando la aproximación de grafos que acabamos de ver, puede utilizarse la matriz de similitud **W** para introducir la información local de los datos originales y posteriormente obtener la geometría del espacio subyacente que contiene esas estructuras locales.

Esta aproximación está muy relacionada con el aprendizaje semi-supervisado, donde se busca una función de clasificación revelada, tanto por los datos etiquetados como los que no, que sea lo suficientemente suave como para mantener la estructura interna/local [86].

El concepto clave para el aprendizaje semi-supervisado es la asunción de consistencia. En clasificación, este concepto de consistencia significa que:

1. puntos cercanos tienden a tener la misma etiqueta.

2. puntos que tienen la misma estructura tienden a tener la misma etiqueta.

Esta segunda asunción se puede interpretar desde un punto de vista de reducción de dimensionalidad como que aquellos puntos que residen en un mismo espacio de menor dimensionalidad tendrán la misma etiqueta. Una interpretación desde un punto de vista de grafos embebidos es que si dos puntos están unidos por un arco, tenderán a pertenecer a la misma clase. Además, los puntos que estén contenidos en un sub-grafo densamente conectado también tenderán a pertenecer a un mismo grupo [87]. Con estas asunciones se pueden utilizar tanto los datos etiquetados como los que no están etiquetados para mejorar el proceso de aprendizaje.

Para nuestro problema utilizaremos este tipo de aprendizaje semi-supervisado de una manera implícita debido a que al tratarse de un problema de regresión no

disponemos de información real del etiquetado de los datos, sino que disponemos de datos continuos de alta dimensión. Por tanto, podemos considerar que todos los datos están no etiquetados y aceptar como ciertas las asunciones de consistencia para poder así buscar las diferentes estructuras locales. También podemos pensar de modo contrario, que todos los datos están etiquetados pero que no existen dos etiquetas iguales y, por tanto, tenemos tantas etiquetas como casos, siendo necesario la búsqueda de estructuras locales para obtener conjuntos de puntos homogéneos. Como puede verse, nuestra propuesta está de manera implícita dentro de lo que se considera el aprendizaje semi-supervisado.

6.3.2. Grafo de similitud o matriz "kernel"

Los parámetros básicos necesarios para incluir la información local y poder descubrir las diferentes estructuras en el conjunto de datos son dos. El primero de ellos es la matriz o grafo de similitud $\mathbf{W}$. El grafo de similitud se corresponde con los valores de los arcos W de un grafo $G(V, W)$. Cada arco contiene la información referente al grado de similitud entre dos casos/vértices v_i y v_j, siendo $w_{ij} \geq 0$. Si $w_{ij} = 0$ significa que los vértices x_i y x_j no están conectados. El segundo parámetro es la matriz de adyacencia $A(G)$ que permite decidir cuando dos vértices v_i y v_j están unidos o no por un arco. Estos dos parámetros pueden ser utilizados de manera simultanea para lograr así la matriz de similitud final deseada.

Existen diferentes modos para transformar un conjunto de puntos dado $\mathbf{x}_1, \ldots, \mathbf{x}_m$, en una matriz de similitudes. Recordar que el objetivo principal es modelar el comportamiento local entre los datos. Además, la mayoría de los modelos de construcción dan lugar a matrices de tipo disperso lo cual tiene ventajas desde un punto de vista computacional. A continuación se describen las diferentes maneras de poder obtener dicha matriz de similitudes.

Grafo de ϵ-vecinos: se conectan todos los puntos cuya distancia entre pares sea más pequeña que ϵ. Debido a que las distancias entre todos los puntos conectados son de escala similar (como mucho ϵ), realizar una ponderación de los arcos no va

a incorporar información adicional sobre los datos del grafo. De este modo, el grafo ϵ-vecinos es normalmente considerado un grafo sin pesos.

Grafo de k-vecinos más cercanos: aquí, un vértice v_i estará conectado con otro v_j si v_j pertenece a los k-vecinos más cercanos de v_i. Esta definición por si sola provoca la creación de un grafo de tipo direccional, debido a que las relaciones de k-vecinos más cercanos no son simétricas. Para hacer que este grafo sea no direccionado se puede adoptar uno de los siguientes caminos. El primero consiste en ignorar las direcciones de los arcos y conectar los vértices v_i y v_j tanto si v_j está entre los k-vecinos más cercanos de v_i, o si v_i está entre los k-vecinos más cercanos de v_j. El grafo resultante es el que comúnmente es referido como grafo de k-vecinos más cercanos. La otra opción es conectar los vértices v_i y v_j, si y sólo si ambos están entre los k-vecinos más próximos del otro. A este grafo se le denomina grafo de k-vecinos mutuamente más cercanos. En ambos casos, una vez se han conectado los vértices de manera apropiada se establecen los pesos de cada uno de los arcos con los valores de similitud adecuados.

Grafo completamente conectado: simplemente se conectan todos los puntos con similitud positiva, y se establecen los valores de sus arcos con w_{ij}. Como el grafo debe modelar el comportamiento local, este tipo de construcción sólo es apropiada cuando la función de similitud utilizada contiene ese tipo de información por sí misma. Un ejemplo de función de este tipo sería una función gaussiana $w(x_i, x_j) = exp^{-\frac{||x_i - xj||^2}{2\sigma^2}}$. Para esta función, el parámetro σ controlará el tamaño de los vecinos a tener en cuenta de manera similar a como lo hace ϵ para el caso de un grafo de k-vecinos más próximos.

Otro tipo de funciones que incluyen el comportamiento local por sí mismas son las funciones de tipo kernel que se caracterizan por ser semi-definidas positivas [88]. Algunas matrices tipo kernel que pueden usarse con el fin descrito son:

$$K(x, y) = x \cdot y, \tag{6.6}$$

$$K(x, y) = (1 + x \cdot y)^p, \tag{6.7}$$

$$K(x,y) = e^{\frac{-||x-y||^2}{2\sigma^2}}. \tag{6.8}$$

Por tanto, este tipo de matrices también pueden ser utilizadas de manera equivalente como matriz de similitudes **W**.

El resto de esta segunda parte de la tesis se organiza como sigue. El capítulo 7 contiene el desarrollo de la técnica propuesta junto con la descripción del algoritmo utilizado. Los resultados de los experimentos correspondientes a esta segunda propuesta se muestran en el capítulo 8.

Capítulo 7

Análisis de datos de estructuras homogéneas

Todo tiene que ser lo más simple posible, pero no más.
–Albert Einstein (1879-1955)

Un conjunto de datos pareados es aquel donde un mismo objeto es representado en dos (o más) espacios diferentes. En ese contexto es coherente suponer que existe un modelo latente que relaciona las diferentes representaciones, las cuales pueden ser imaginadas como las diferentes formas del objeto subyacente que aparecen al ser representado en los respectivos espacios. Este pareado de conjuntos de datos puede tener su origen en los diferentes métodos de medida utilizados.

En el problema de regresión que estamos tratando, el conjunto de datos está representado por un matriz de entrada $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ donde $\mathbf{x}_i \in \Re^{p\times 1}$ y representa los datos de emisión energética a diferentes longitudes de onda de una llama, y una matriz de salida $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ donde $\mathbf{y}_i \in \Re^{s\times 1}$ y se corresponde con los perfiles de temperatura asociados a esa llama.

En nuestro caso, el objeto subyacente será la propia física y las dos representaciones serán las dos medidas que se obtienen. Por un lado, se tiene la medida obtenida usando el espectroradiómetro, lo que proporciona un espectro de energía a diferentes longitudes de onda, y por otro se tienen las medidas de la temperatura en las diferentes zonas de la llama. Ambas medidas se corresponden con la misma ley física y, por

tanto, son diferentes representaciones de un mismo objeto. Nosotros proponemos la búsqueda de ese espacio común a ambas representaciones de tal modo que en ese nuevo espacio común ambas vistas estén relacionadas en base a un criterio de homogeneidad, logrando así atenuar los problemas relacionados con la multicolinealidad explicados en la introducción (apartado 1.3.2).

De este modo, asumimos que existe un subespacio común a ambos espacios, de entrada y salida, y que ese subespacio va a contener las estructuras comunes a ambos permitiendo así identificar los diferentes grupos homogéneos. La aproximación propuesta hace uso de un procedimiento de tres pasos.

1. Dado un conjunto de datos $\mathbf{X} \in \Re^{p \times m}$ e $\mathbf{Y} \in \Re^{s \times m}$, debemos encontrar el subespacio de menor dimensión que contenga ambos conjuntos de datos y los correlacione.

2. Detectar las estructuras homogéneas que se crean en el espacio generado y agruparlas.

3. Realizar la estimación para cada una de las estructuras detectadas.

Para el diseño de este proceso hay que tratar, entre otras, las siguientes cuestiones:

- ¿Cómo definir el concepto de grupo homogéneo en el diseño de la técnica?
- ¿Qué tipo de agrupamiento debemos hacer sobre los datos obtenidos en el nuevo espacio generado?
- ¿Qué tipo de estimador será necesario utilizar para realizar la reconstrucción de las temperaturas?

Más concretamente, la propuesta consiste en un algoritmo de reducción de dimensionalidad que va a usar la información relacionada con las salidas para mitigar el efecto de la multicolinealidad en las entradas. El algoritmo va a explotar la geometría de los datos de salida y encontrar un subespacio de entrada que mejor preserve esa geometría. Para esto vamos a utilizar una entorno basado en grafos que nos permite introducir esa información sobre la geometría tanto global como local. A continuación

se describe la forma en la que se ha realizado esta propuesta y la descripción completa del algoritmo.

7.1. Agrupamiento por maximización de correlaciones y de distancias

En la sección 6.1 se habló del método *correlational spectral clustering* [77] el cual usa la técnica KCCA [89] para buscar dos subespacios diferentes, una para las entradas ψ_x y otra para las salidas ψ_y de tal modo que las proyecciones en los subespacios estén máximamente correladas. Este modelo asume un modelo subyacente que soporta ambos subespacios y las transformaciones son diferentes para cada uno de ellos.

Aunque este método encuentra grupos de datos homogéneos, únicamente tiene en cuenta la correlación entre los nuevos datos proyectados y no trata de mantener de ningún modo la distancia en el espacio original de los datos.

Nuestro método se basa en la asunción de un único modelo subyacente el cual contiene el subespacio común a ambas representaciones del objeto, las entradas $\mathbf{X}$ y las salidas $\mathbf{Y}$. Con este esquema, vamos a buscar la transformación que encuentre el espacio de menor dimensión y que preserve las distancias originales de ambos conjuntos de datos de manera independiente. Además, se añade la restricción de que en esos nuevos subespacio generados ψ_x y ψ_y se mantenga una relación lineal de uno frente al otro. Esta es la métrica definida como criterio de homogeneidad. Siguiendo la notación introducida en el capítulo anterior, $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_{m'}) \in \Re^{p' \times m}$ será la transformación de los datos de la entrada, y $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_{s'}) \in \Re^{s' \times m}$ de la salida.

La propuesta de este método en el contexto de una aplicación de regresión es motivada por el conocimiento de que existen un conjunto de estructuras en los datos de salida, los cuales son desconocidos en los datos de entrada debido a que es de mayor complejidad. Por lo tanto, nuestro objetivo es encontrar una transformación que permita descubrir el subespacio de las entradas que mejor represente esa información estructurada de las salidas. Siguiendo la formulación introducida en la sección 6.3

podemos construir dos grafos G y H para los datos de entrada y los datos de salida, respectivamente. Cada uno de estos grafos tiene asociado su respectiva matriz de similitud $\mathbf{U} \in \Re^{m \times m}$ y $\mathbf{W} \in \Re^{m \times m}$. Recordar que si dos vértices v_i y v_j, que se corresponden con dos ejemplos distintos de los m posibles, están conectados entonces el valor de la matriz de similitud $u_{ij} \geq 0$, y que si su valor $u_{ij} = 0$ significa que no están conectados o que su similitud es mínima. Un criterio razonable para lograr una "buena" transformación de ambos espacios preservando sus distancias locales es

$$\text{mín} \sum_{ij}^{m} (\hat{y}_i - \hat{y}_j)^2 w_{ij} \tag{7.1}$$

$$\text{mín} \sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 u_{ij} \ . \tag{7.2}$$

Además hay que añadir la restricción de que el espacio de salida ψ_y pueda ser reconstruido linealmente desde ψ_x. Esta restricción trata de encontrar en las entradas la misma estructura de las salidas. De este modo, la ecuación 7.2 queda modificada por la siguiente expresión

$$\text{mín} \sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 u_{ij} w_{ij} \tag{7.3}$$

donde se introduce la información referente a las salidas usando la matriz de similitudes $\mathbf{W}$. Esta ecuación trata de preservar las distancias en el espacio de las entradas pero únicamente cuando esas entradas están correlacionadas con las salidas. Así se logra realizar una reducción selectiva en base a homogeneidades.

Es conocido que la matriz de similitud de las salidas $\mathbf{W}$ tiene una estructura de bloque. Esta estructura de bloque vendrá determinada por el criterio elegido para crear la matriz de similitud (véase apartado 6.3.2). En nuestro caso esta matriz de similitud $\mathbf{W}$ es creada usando una función de correlación lineal. Por tanto podemos asumir sin perdida de generalidad que los datos $y_1, \ldots, y_m$ están ordenados de acuerdo

a su valor de similitud. Entonces, **W** es del tipo

$$\mathbf{W} = \begin{pmatrix} W^1 & & & \\ & W^2 & & \\ & & \ddots & \\ & & & W^c \end{pmatrix}$$

donde el superíndice c indica las diferentes estructuras de bloque internas. El objetivo entonces es intentar separar estos bloques, los cuales representan zonas homogéneas, unos de otros. Asumiendo que sólo existen dos estructuras, podemos reescribir la ecuación 7.3 como un nuevo problema de optimización donde se tratará de maximizar la distancia entre los casos que pertenezcan a estructuras diferentes, y minimizar la distancia entre los que pertenecen a una misma estructura. De este modo, obtenemos las siguientes ecuaciones

$$\text{mín} \sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 u_{ij} w_{w,ij} \tag{7.4}$$

$$\text{máx} \sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 u_{ij} w_{b,ij} \tag{7.5}$$

donde $w_{w,ij}$ indica la similitud entre los elementos que pertenecen a un mismo bloque/grupo, y $w_{b,ij}$ indica la similitud entre los elementos pertenecientes a distintos bloques. Entonces, las nuevas funciones objetivo son minimizar 7.1 y 7.4, y maximizar 7.5.

De este modo, se mantiene la información en el espacio de salida de acuerdo a la ecuación 7.1, y se buscan los conjuntos de datos de las entradas que presentan esa estructura de acuerdo a la ecuación 7.4. Simultáneamente, también se mantiene la estructura interna (información local) de las entradas, a la vez que se maximiza la distancia frente a los datos de otras estructuras en 7.5 (información global).

Asumiendo que no tenemos ninguna información a priori sobre la matriz de similitud de las entradas, podemos considerar que su estructura es del tipo $\frac{1}{m} ee^T$. Esta estructura es de tipo PCA y equivale a buscar las direcciones de mínima varianza de

$C = \frac{1}{m}\sum_i (x_i - \overline{x})(x_i - \overline{x})^T = \frac{1}{m}\mathbf{X}(\mathbf{I} - \frac{1}{m}ee^T)\mathbf{X}^T$. Desarrollando la ecuación 7.4, y teniendo en cuenta que $\sum_j u_{ij} = m$, que $\sum_{ij} u_{ij} = m^2$, y que $D_{ii} = \sum_j w_{ij}$ siguiendo la notación del apartado 6.3, entonces

$$\begin{aligned}
&\text{mín} \sum_{ij}^{m} (\hat{x}_i - \hat{x}_j)^2 u_{ij} w_{w,ij} \\
&= \text{mín} \Big(\sum_{ij} \hat{x}_i^2 u_{ij} w_{ij} + \sum_{ij} \hat{x}_j^2 u_{ij} w_{w,ij} - 2 \sum_{ij} \hat{x}_i \hat{x}_j u_{ij} w_{w,ij} \Big) \\
&= \text{mín}(2m^2\ \hat{\mathbf{x}}^T \mathbf{D} \hat{\mathbf{x}} - 2m^2\ \hat{\mathbf{x}}^T \mathbf{W}_w \hat{\mathbf{x}}) = \text{mín}\ (2m^2\ \hat{\mathbf{x}}^T \mathbf{L}_w \hat{\mathbf{x}})
\end{aligned} \tag{7.6}$$

siendo $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Como puede verse, es igual a la ecuación 6.3 salvo por el factor constante $2m^2$ que no afecta a la optimización de la ecuación. De este modo el problema de optimización queda expresado como:

$$\text{mín}\, \hat{\mathbf{x}}^T \mathbf{L}_w \hat{\mathbf{x}} \tag{7.7}$$

De manera similar se puede desarrollar la ecuación 7.5, obteniendo

$$\text{máx}\ \ \hat{\mathbf{x}}^T \mathbf{L}_b \hat{\mathbf{x}}\ . \tag{7.8}$$

Con esto demostramos que utilizando una orientación basada en grafos es simple introducir en las entradas la información referente al espacio de salida. Entonces, el problema de buscar el subespacio de entrada que mejor represente los bloques de la salida es equivalente a resolver los problemas 7.7 y 7.8.

Una vez determinado como encontrar el nuevo subespacio de entrada ψ_x , es necesario encontrar también el subespacio de la salida ψ_y correspondiente a la ecuación 7.1 Este subespacio deben mantener una dependencia lineal para poder así estimar las salidas a partir de las entradas. Desarrollando la ecuación 7.1 y teniendo en cuenta

dicha relación $\hat{\mathbf{y}}_i = \mathbf{b}^T\mathbf{x}_i$, entonces

$$\begin{aligned}
&\text{mín } \sum_{ij}^{m}(\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j)^2 w_{ij} = \text{mín} \sum_{ij}(\mathbf{b}^T\mathbf{x}_i - \mathbf{b}^T\mathbf{x}_j)^2 w_{ij} \\
&= \text{mín} \sum_{ij}(\ (\mathbf{b}^T\mathbf{x}_i)^2 + (\mathbf{b}^T\mathbf{x}_j)^2 - 2(\mathbf{b}^T\mathbf{x}_i\mathbf{b}^T\mathbf{x}_j)\ )w_{ij} \\
&= \text{mín} \sum_{i}(\mathbf{b}^T\mathbf{x}_i)^2 d_{ii} + \sum_{j}(\mathbf{b}^T\mathbf{x}_j)^2 d_{jj} - 2\sum_{ij}(\mathbf{b}^T\mathbf{x}_i\mathbf{b}^T\mathbf{x}_j)\mathbf{w}_{ij} \\
&= \text{mín}\, 2(\mathbf{b}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{b} - \mathbf{b}^T\mathbf{X}\mathbf{W}\mathbf{X}^T\mathbf{b}) \\
&= \text{mín } \mathbf{b}^T\mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^T\mathbf{b}\ . \qquad (7.9)
\end{aligned}$$

Sustituyendo $\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w$ y $\hat{\mathbf{x}}$ por su combinación lineal sobre las entradas $\hat{\mathbf{x}}_i = \mathbf{b}^T\mathbf{x}_i$, la expresión 7.7 puede ser expresada como

$$\text{mín } \mathbf{b}^T\mathbf{X}(\mathbf{D}_w - \mathbf{W}_w)\mathbf{X}^T\mathbf{b}\ . \qquad (7.10)$$

Igualmente, la ecuación 7.5 se reescribe como

$$\text{máx } \mathbf{b}^T\mathbf{X}(\mathbf{D}_b - \mathbf{W}_b)\mathbf{X}^T\mathbf{b}\ . \qquad (7.11)$$

Hay que resaltar que la matriz $\mathbf{D}_w$ proporciona una manera natural de medir la densidad que existe alrededor de un punto $\mathbf{x}_i$. Por lo tanto, cuanto más grande sea el valor d_{ii} más "importancia" tiene $\mathbf{x}_i$. Por esto, hay que añadir una restricción del tipo:

$$\mathbf{b}^T\mathbf{X}\mathbf{D}_w\mathbf{X}^T\mathbf{b} = 1$$

lo que implica que la ecuación 7.10 pasa a ser

$$\min_b(1 - \mathbf{b}^T\mathbf{X}\mathbf{W}_w\mathbf{X}^T\mathbf{b}) \qquad (7.12)$$

o de manera equivalente

$$\max_b \mathbf{b}^T\mathbf{X}\mathbf{W}_w\mathbf{X}^T\mathbf{b}\ . \qquad (7.13)$$

Si miramos las ecuaciones 7.9 y 7.13 se puede observar que son similares debido

a que la matriz $\mathbf{W}$ utilizada en 7.9 para indicar las similitudes entre los puntos y la utilizada en 7.13 para indicar la estructura de las salidas es la misma. Esto es así porque la matriz $\mathbf{W}_w$ contiene de manera intrínseca información sobre la estructura de los datos de salida, y por la restricción establecida sobre $\mathbf{D}_w$.

Por lo tanto, el objetivo final buscado queda reducido a:

$$\begin{cases} \text{máx}_b\, \mathbf{b}^T\mathbf{X}\mathbf{W}_w\mathbf{X}^T\mathbf{b} \\ \text{máx}_b\, \mathbf{b}^T\mathbf{X}(\mathbf{D}_b - \mathbf{W}_b)\mathbf{X}^T\mathbf{b} = \text{máx}_b(1 - \mathbf{b}^T\mathbf{X}\mathbf{W}_b\mathbf{X}^T\mathbf{b}) = \text{mín}_b\, \mathbf{b}^T\mathbf{X}\mathbf{W}_b\mathbf{X}^T\mathbf{b} \end{cases}$$

La función objetivo será entonces,

$$\ell = \max_b \frac{\mathbf{b}^T\mathbf{X}\mathbf{W}_w\mathbf{X}^T\mathbf{b}}{\mathbf{b}^T\mathbf{X}\mathbf{W}_b\mathbf{X}^T\mathbf{b}} \ . \tag{7.14}$$

Para el caso de búsqueda de estructuras lineales en la salida se utiliza una matriz de similitud $\mathbf{W}$ que establece las distancias en base a su correlación lineal. En este caso la matriz $\mathbf{W}_w = \mathbf{Y}^T\mathbf{Y}$ y la ecuación 7.14 pasa a ser:

$$\ell = \max_b \frac{\mathbf{b}^T\mathbf{X}\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{b}}{\mathbf{b}^T\mathbf{X}\mathbf{W}_b\mathbf{X}^T\mathbf{b}} \ . \tag{7.15}$$

Resaltar que igualmente puede utilizarse una matriz tipo kernel $\mathbf{W}_w = K(y, y) = \mathbf{K}_{yy}$ para detectar las estructuras de la salida y de eso modo permitir buscar estructuras no lineales.

Si no se dispone de información sobre los diferentes grupos de la salida no se pueden determinar los valores de la matriz $\mathbf{W}_b$ pero se puede utilizar la matriz de covarianza total como criterio único para medir la distancia entre los puntos pertenecientes a diferentes grupos. Para el caso que estamos estudiando, esto implica una matriz de covarianza lineal para las entradas. Estableciendo de ese modo $\mathbf{W}_b = \mathbf{I}$ se obtiene la función:

$$\ell = \max_b \frac{\mathbf{b}^T\mathbf{X}\mathbf{Y}^T\mathbf{Y}\mathbf{X}^T\mathbf{b}}{\mathbf{b}^T\mathbf{X}\mathbf{X}^T\mathbf{b}} \tag{7.16}$$

donde $\mathbf{b}$ es la transformación del subespacio buscado.

A continuación se describe el algoritmo completo utilizado para obtener los buscados diferentes modelos locales. Para ello se indica como se ha obtenido la transformación **b** y como se aplica un estimador para cada modelo.

7.2. Descripción del algoritmo

Con el fin de agrupar diferentes subconjuntos de datos homogéneos para así solventar el problema de la multicolinealidad, vamos a transformar el problema general de regresión en n sub-problemas de regresión donde cada uno de esos subconjuntos llevará asociado un modelo para la estimación de los perfiles de temperatura.

Para ello nos basamos en las estructuras contenidas en los datos de salida las cuales son recogidas en la matriz de Gram definida por $\mathbf{C}_{yy} = \mathbf{Y}^T\mathbf{Y}$. Posteriormente es necesario pasar esta información referente a las estructuras de la salida a los datos de entrada. Este problema es resuelto utilizando la ecuación 7.16 obtenida en el desarrollo explicado en este capítulo, y de ese modo se calcula la matriz de transformación $\mathbf{B}$.

Esta matriz $\mathbf{B}$ permite proyectar los datos de entrada $\mathbf{X}$ en un nuevo subespacio donde residen las estructuras buscadas. Posteriormente se procede a realizar una agrupación de los datos proyectados $\mathbf{P}$ donde los distintos ejemplos quedan agrupados por homogeneidad. Cada uno de estos grupos se corresponderá con un modelo local. Entonces, se calcula cada uno de estos modelos para obtener finalmente la estimación de los perfiles de temperatura.

A continuación se describe de manera detallada el algoritmo utilizado donde los datos de entrada y de salida son $\mathbf{X} \in \Re^{p\times m}$ e $\mathbf{Y} \in \Re^{s\times m}$ respectivamente, con m ejemplos, y cada uno de estos ejemplos con una dimensión de entrada p, y una dimensión de salida s:

1. Construir la matriz de Gram correspondiente a los datos de salida $\mathbf{W}_w = \mathbf{C}_{yy} = \mathbf{Y}^T\mathbf{Y} \in \Re^{m\times m}$, o su versión genérica tipo kernel $\mathbf{K}_{yy}$ usando alguno de los kernel definidos en (6.6), (6.7), o (6.8), o sus posibles combinaciones (véase [88] para ver los operadores que conservan las propiedades de un kernel).

2. Resolver el problema generalizado de autovectores correspondiente a la función

objetivo 7.16. Para ello se utilizan los siguientes pasos:

Calcular la descomposición en valores singulares (SVD) de $\overline{\mathbf{X}}$, $\overline{\mathbf{X}} = \mathbf{USV^T}$, siendo $\overline{\mathbf{X}}$ la matriz centrada de los datos de entrada, y $\mathbf{s}_1, \ldots, \mathbf{s}_r$ son los valores singulares asociados a los autovectores por la izquierda de $\mathbf{U}$, y por la derecha de $\mathbf{V}$.

Entonces, sustituyendo $\overline{\mathbf{X}}$ por su descomposición SVD, la ecuación 7.16 puede ser transformada en $\text{máx}_B \frac{tr(\mathbf{B}^T\mathbf{USV}^T\mathbf{W_wVSU}^T\mathbf{B})}{tr(\mathbf{B}^T\mathbf{USV}^T\mathbf{VSU}^T\mathbf{B})}$. Haciendo $\mathbf{C} = \mathbf{SU}^T\mathbf{B}$ se obtiene el siguiente problema de optimización, $\text{máx}_C \frac{tr(\mathbf{C}^T\mathbf{V}^T\mathbf{W}_w\mathbf{VC})}{tr(\mathbf{C}^T\mathbf{C})}$ que puede ser resuelto como un problema de autovectores sobre $\mathbf{C}$ y después resolver en $\mathbf{B}$ como $\mathbf{B} = \mathbf{US}^{-1}\mathbf{C}$.

3. Calcular los nuevos datos proyectados en los nuevos ejes calculados, siendo $\mathbf{P} = \mathbf{B}^T\overline{\mathbf{X}}$.

4. Aplicar un algoritmo de agrupación por densidades sobre las primeras l nuevas características de los datos proyectados $\mathbf{P} \in \Re^{l \times m}$.

5. Aplicar un modelo de regresión para cada uno de los c grupos obtenidos.

En el primer paso, la matriz de Gram ya incluye información relevante sobre la localidad de los datos por lo que no ha sido necesario utilizar una matriz de adyacencia externa. Aún así, si se dispone de información previa sobre datos que deben ser considerados vecinos se puede incluir una matriz de adyacencia indicando está información a modo de máscara.

Para calcular el número de dimensiones a utilizar en las proyecciones obtenidas $\mathbf{P}$ nos basamos en los mismos criterios que utilizamos en el capítulo 4 donde el porcentaje de varianza acumulada es el más común. Otro criterio útil en este caso es establecer su valor al número de variables desconocidas del modelo a estimar, debido a que el subespacio nuevo generado vendrá determinado en el caso ideal por ese número. Esto sucede cuando se ha logrado realizar una buena reducción de la dimensionalidad y se separan las interdependencias existentes. Aunque no disponemos de una base teórica para este supuesto, si parece que esto sucede en nuestra aplicación.

Para los casos de entrenamiento, una vez se han proyectado los datos en el nuevo espacio generado, se realiza una agrupación por densidades logrando así que todos los ejemplos que poseen homogeneidad puedan ser identificados como un único modelo.

Los algoritmos basados en densidades buscan grupos de datos teniendo en cuenta la distribución de los puntos. De este modo, se trata de encontrar zonas que tienen una alta concentración de puntos en su entorno, mientras que entre diferentes zonas aparecen áreas de baja densidad de puntos.

Existen varios algoritmos de densidades basados en diferentes criterios. Nosotros utilizamos DBSCAN [90] el cual está basado en la unión de zonas de alta densidad. Para ello, utiliza el concepto de punto central para ir construyendo de manera iterativa un esqueleto de puntos que corresponden a un mismo grupo. *Mean-Shift* [91], usa un vector de desplazamiento que indica las zonas de mayor densidad. De este modo el vector posición se desplaza indicando paso por paso las áreas de densidad alta. Otros algoritmos basados en densidades son: KNN-clust, Clique, DENCLUE, LPC, etc., donde cada uno presenta características y aplicaciones específicas.

En [37] puede encontrarse una comparativa de algunas de estas técnicas. Nosotros hemos utilizado el algoritmo DBSCAN por su simplicidad y rapidez frente a otros más costosos computacionalmente como *Mean-Shift*. Además los resultados obtenidos en la detección de los diferentes grupos buscados son satisfactorios como hemos visto en este trabajo y se demostró en [37].

Uno de los problemas comunes en este tipo de técnicas de reducción y descubrimiento de información (ISOMAP [36], LLE [35], etc.) suele ser como realizar el proceso de test. Debido a que se basan en la matriz de Gram para calcular un subespacio en base a las distancias entre pares, esta matriz también es necesaria para el test, y la única manera de poder incluir un nuevo ejemplo en el proceso es ampliar la matriz de Gram y volver a realizar el cálculo de los autovectores. Por lo tanto, estas técnicas tienen un carácter más exploratorio que de aprendizaje como tal.

Nosotros hemos desarrollado nuestro algoritmo de tal modo que al final obtenemos un conjunto de vectores que forman una base $\mathbf{B}$ y que puede utilizarse para el proceso de test. Si se quiere realizar el test sobre un nuevo caso $\mathbf{x}_{test}$, se puede obtener su proyección $\mathbf{p}_{test}$ en el nuevo espacio $\mathbf{B}$ como, $\mathbf{p}_{test} = \mathbf{B}^T\mathbf{x}_{test}$.

Cuando se proyecta un dato de test se utiliza el algoritmo del más cercano con un valor de K = 1 para su clasificación dentro de uno de los grupos detectados. Es decir, el nuevo dato $\mathbf{x}_{test}$ es proyectado obteniendo $\mathbf{p}_{test}$. Entonces, se busca el ejemplo de entrenamiento $\mathbf{p}_{train}$ más cercano a $\mathbf{p}_{test}$ y se le asigna al mismo grupo/modelo.

Capítulo 8

Experimentos

Una vez eliminado todo lo que es imposible, entonces
lo que sea que quede, aunque improbable, debe ser la verdad.
–Sherlock Holmes (El signo de los cuatro)

En este capítulo vamos a mostrar los resultados obtenidos al aplicar la técnica desarrollada al problema de estimación de la temperatura a partir de los datos de un espectro en el contexto de teledetección.

El diseño de los experimentos es el mismo que se ha descrito en el apartado 2.5 aunque los objetivos buscados difieren. Aquí vamos a comprobar que los diferentes grupos de datos encontrados van a permitir una estimación más precisa que el modelo global. Para ello se aplicará el algoritmo propuesto descrito en el apartado 7.2 y así descubrir los diferentes grupos existentes. Posteriormente se aprenderá un modelo de estimación para cada uno de ellos. Los resultados obtenidos para cada uno de los modelos serán comparados con los resultados obtenidos por un modelo global para ese mismo conjunto de ejemplos. Además, se establecerá una comparación frente a la utilización de otros métodos de extracción de características como son PCA y KCCA.

Debido a que el modelo inverso que tratamos de resolver se comporta de manera no lineal, se ha utiliza un perceptron multicapa para realizar las estimaciones en los modelos globales. Por el contrario, hemos optado por utilizar un regresor lineal (mucho más simple), para mostrar la mejora obtenida al aplicar diferentes modelos de estimación para cada uno de los diferentes conjuntos de datos descubiertos. Esto es

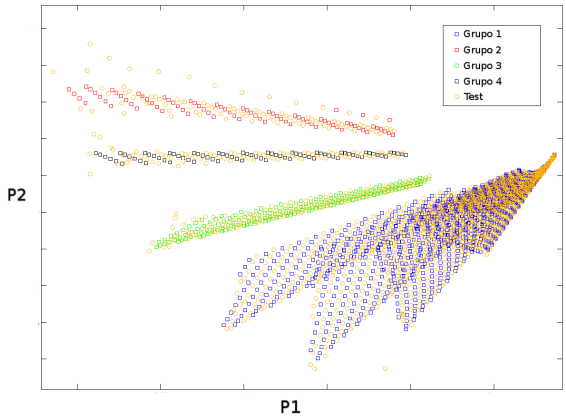

Figura 8.1: Grupos obtenidos al proyectar los ejemplos en las dos primeras dimensiones.

debido a que en el desarrollo de nuestra técnica, hemos utilizado matrices de similitud lineales y por tanto las estructuras encontradas son también lineales. Este factor nos permite poder aplicar con éxito un modelo de estimación lineal para los diferentes grupos encontrados.

8.1. Resultados

En esta sección se describen los resultados obtenidos para las cuatro pruebas realizadas. Cada una de las pruebas se corresponde con uno de los grupos encontrados, y para realizar la comparativa se obtienen las estimaciones tanto para el modelo global, como para los distintos modelos locales.

La figura 8.1 muestra las proyecciones $\mathbf{P}$ de los datos de entrenamiento y de test en las primeras dos dimensiones. Estas dos dimensiones contienen la mayor parte de la varianza acumulada, y además el número de parámetros desconocidos en la aplicación son dos, la temperatura y la longitud de la llama, lo cual coincide con uno de los criterios a utilizar para seleccionar el número de dimensiones a utilizar al realizar la reducción. Los diferentes colores que se muestran se corresponden con los diferentes grupos encontrados, y los puntos circulares de color naranja son los

puntos de test. Destacar que las diferentes estructuras mostradas no son todas iguales. Puede observarse que existe una estructura que presenta una mayor complejidad (mostrada en color azul) en la cual existen interdependencias que no han sido resueltas y probablemente contenga diferentes sub-estructuras en su interior.

La propiedad que caracteriza a estas diferentes estructuras mostradas es que se corresponden con áreas de alta densidad de puntos, frente a zonas de menor o baja densidad. Por esto, el algoritmo escogido DBSCAN realiza un buen agrupamiento de los datos. También puede observarse en la figura que los datos de test $\mathbf{p}_{test} = \mathbf{B}^T\mathbf{x}_{test}$ están perfectamente embebidos en las estructuras de los datos de entrenamiento $\mathbf{p}_{train} = \mathbf{B}^T\mathbf{x}_{train}$. Debido a esto, el algoritmo de clasificación de los K más cercanos con una valor de $K = 1$ es muy apropiado. Con este clasificador se han obteniendo resultados muy precisos, $\approx 98\,\%$ de éxito, al asignar un nuevo punto a uno de los modelos locales existentes.

Para mostrar la mejora frente a un único modelo global hemos realizado cuatro pruebas diferentes. Cada una de las pruebas se corresponde con uno de los grupos descubiertos. Para el modelo global se ha elegido un MLP como regresor debido a la no-linealidad global del problema, y para los modelos locales se ha elegido un regresor tipo lineal.

Como se ha comentado anteriormente, puede observarse en la figura 8.1, que el grupo asociado al color azul presenta una estructura mucho más compleja que los otros grupos. Por esto, es previsible que el modelo lineal sea menos favorable en este caso. La estructura de ese grupo es más compleja debido a que las distancias relativas entre sus ejemplos y el resto de grupos es muy similar. Frente a este caso puede adoptarse un modelo de estimación tipo MLP debido a que su linealidad será menor que para los otros grupos. En la tabla 8.1 hemos incluido los resultados de ambos casos para apreciar este hecho.

La arquitectura usada para el MLP es una capa oculta y el número de neuronas ha sido fijado siguiendo una aproximación tipo *greedy*. Hemos obtenido resultados satisfactorios para un valor de 30 neuronas ocultas. La tabla 8.1 muestra el error medio absoluto por perfil de temperaturas (MAEs), al igual que se hizo en la parte I. Este valor es calculado como $MAEs = \frac{1}{z}\frac{1}{m}\sum_{k=1}^{z}\sum_{j=1}^{m}|\mathbf{y}_{kj} - \hat{\mathbf{y}}_{kj}|$ donde z es la

longitud de la combustión discretizada, y m el número de ejemplos. Igualmente, la tabla recoge la información referente a la desviación estándar del error en las salidas con respecto a la media (SD). Estos errores son medidos en unidades Kelvin (K).

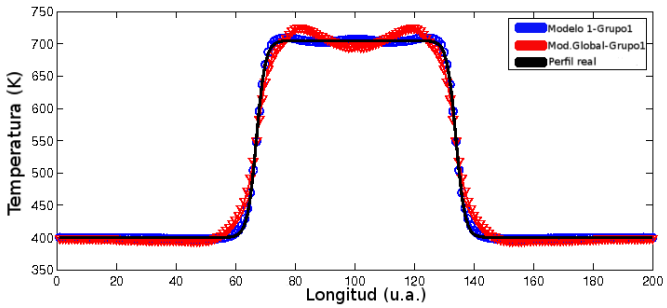

Figura 8.2: Ejemplo de una mala estimación utilizando el modelo global frente a la correcta estimación con nuestro modelo de grupos.

El error MAEs proporciona una idea del error físico obtenido. En los experimentos realizados, el error MAEs está por debajo de un 1 % de error relativo lo cual es considerado como alta precisión para la mayoría de aplicaciones prácticas [92, 74].

La tabla 8.1 está dividida en tres filas principales. A su vez, las dos primeras filas principales se dividen en otras cuatro que se corresponden con los diferentes grupos encontrados por el algoritmo. La primera fila principal muestra los resultados obtenidos para cada uno de los grupos utilizando el modelo global. Para este modelo global se ha escogido PCA como técnica de reducción de la dimensionalidad. Como se vio en el apartado 3.2, PCA es la técnica de reducción de características más utiliza y es considerada un estándar dentro de los métodos de extracción de características. Por esto, la utilizamos como base de comparación frente a nuestra propuesta. Igualmente, la segunda fila muestra los resultados correspondientes al método propuesto para cada uno de los grupos, tanto para entrenamiento como para el test.

Aunque los resultados mostrados para el modelo global son bastante precisos y reducen el error con respecto a los datos obtenidos en la primera parte[1], se observan oscilaciones no deseadas en la reconstrucción de ciertos perfiles. Para proporcionar un

[1]Recordar que en la parte I se trata una técnica de selección de características y no de extracción como en este caso y por tanto, la extracción no posee algunas de sus ventajas como la obtención de un subconjunto de características originales.

mejor entendimiento de este hecho, hemos incluido las figuras 8.2 y 8.3. La figura 8.2 muestra un ejemplo donde el modelo global obtiene peores resultados que el propuesto. Además, en la estimación obtenida por el modelo global pueden observarse algunas oscilaciones que no son deseables y son indicadores del efecto de la multicolinealidad en los datos. Estos efectos no aparecen en las estimaciones realizadas con nuestra propuesta y las oscilaciones desaparecen. Por otro lado, la figura 8.3 muestra un ejemplo donde los dos modelos, tanto el global como el de agrupamiento, obtienen resultados satisfactorios y su error de estimación es muy bajo.

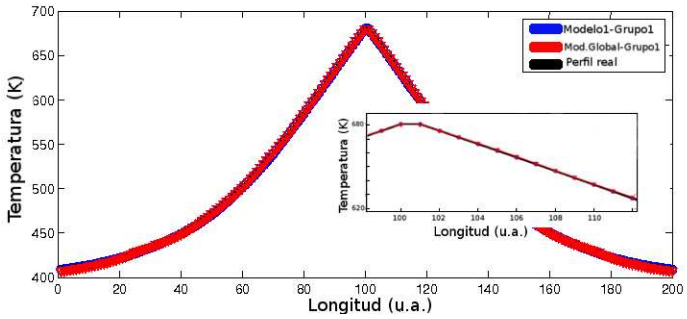

Figura 8.3: Ejemplo de una correcta estimación tanto para el modelo global como para el modelo de grupos.

Como se ha comentado, la tabla 8.1 incluye para el grupo 1 los resultados tanto para un estimador lineal como para un estimador MLP. Puede verse que debido a que se trata de un grupo de complejidad mayor el estimador lineal obtiene perores resultados, y por eso se ha utilizado un estimador no lineal que es capaz de mejorarlos.

Finalmente, la tercera de las filas principales incluye los resultados obtenidos para un único modelo global para diferentes tipos de estimadores. Se han elegido tres estimadores diferentes, los cuales son representativos del entorno de extracción de características y de esta aplicación. Estos modelos son de tipo lineal, PCA y KCCA. En el modelo lineal se utilizan las variables originales y se ha incluido para enriquecer la comparativa frente al resto de modelos. PCA y KCCA son dos técnicas de extracción de características que permiten reducir el número de dimensiones como ya se comentó en la sección 6.1. Por eso, también se han incluido en la comparativa debido

Cuadro 8.1: Error absoluto medio de temperaturas por ejemplo (MAEs), y su desviación estándar (SD).

Método	Grupo No.	Temperatura **Test** (MAEs/SD)K	Temperatura **Entrenamiento** (MAEs/SD)K
Modelo global	Grupo 1(MLP)	2.04/2.84	2.32/3.36
	Grupo 2(MLP)	3.47/3.32	2.38/3.24
	Grupo 3(MLP)	0.90/1.12	0.35/0.49
	Grupo 4(MLP)	0.97/1.32	0.41/0.51
	Media total	1.65/1.94	1.11/1.52
Modelo 1	Grupo 1(MLP)	0.61/1.29	0.52/0.60
Modelo 1	Grupo 1(lineal)	2.50/3.09	2.40/2.70
Modelo 2	Grupo 2(lineal)	0.40/0.45	0.31/0.26
Modelo 3	Grupo 3(lineal)	0.39/0.86	0.19/0.16
Modelo 4	Grupo 4(lineal)	0.42/0.46	0.24/0.19
	Media total	0.44/0.64	0.29/0.26
Modelo global	Lineal	4.74/5.45	4.52/5.15
	PCA+MLP	2.38/17.33	2.26/16.42
	KCCA+MLP	2.27/4.65	2.20/4.41

a que pertenecen al mismo contexto de nuestra propuesta, y además se trata de técnicas utilizadas con éxito en diversos tipos de aplicaciones. Observar que la media de error $\approx$ 0,4 K obtenida por nuestra propuesta mejora también significativamente los resultados de estos tres estimadores, donde el mejor de ellos (KCCA) tiene un error $\approx$ 2,4 K.

En general, los resultados obtenidos por la aproximación propuesta de agrupamiento+regresión mejoran notablemente los resultados que logra el modelo global. Esta mejora es tanto cualitativa como cuantitativa. Esto se observa en la tabla 8.1 donde para todos los casos los modelos específicos mejoran los resultados del modelo global tanto en error absoluto como en la desviación estándar, y en la figura 8.2 donde se observa que el modelo específico resuelve el problema de las oscilaciones.

8.2. Conclusiones

Toma un poco más de té –ofreció solícita la Liebre de Marzo.
–Hasta ahora no he tomado nada –protestó Alicia en tono ofendido–,
de modo que no puedo tomar más. –Quieres decir que no puedes tomar menos
–puntualizó el Sombrerero–. Es mucho más fácil tomar más que nada.
–Lewis Carroll (Alicia en el país de las maravillas)

En esta segunda parte de la tesis hemos estudiado el problema de reducción de dimensionalidad desde un punto de vista de extracción de características. Más concretamente hemos presentado un método que busca la estructura intrínseca común a los datos de entrada y de salida, preservando simultáneamente su información local. Para ello, hemos desarrollado un algoritmo basado en grafos que preserva la información de la entrada e incluye la estructura de las salidas a modo de guía. Hemos aplicado esta técnica al problema específico de teledetección donde se quiere estimar los perfiles de temperatura a partir de los datos pertenecientes a espectros de energía.

Los resultados obtenidos después de realizar una división del problema global en varios modelos lineales más simples ha mejorado los resultados obtenidos por un único modelo. Esto demuestra la hipótesis inicial donde se preveía la falta de homogeneidad de los datos y, por tanto, una influencia de la multicolinealidad en los resultados de las estimaciones para los casos de test. Con nuestra técnica esto no sucede y se ha visto que los datos de test quedan implícitamente embebidos en las estructuras descubiertas logrando así una clasificación casi perfecta de los nuevas casos y mitigando los efectos de la alta dimensionalidad de los datos. Además, los errores obtenidos están por debajo del 1 % lo cual es considerado como muy alta precisión [74].

Una de las características importantes de las técnicas de descubrimiento de información es la posibilidad de interpretación. En este sentido queremos resaltar la relación que existe entre los grupos encontrados y el tamaño de la llama. Los diferentes grupos identificados se corresponden en algún modo con diferentes longitudes y por tanto existen determinados umbrales de longitud donde el comportamiento de la combustión cambia significativamente, mientras que dentro de unos rangos determinados su relación es lineal. Este tipo de información es muy útil a la hora de

caracterizar nuevos problemas y determinar la capacidad para resolverlos.

También queremos destacar que aunque los resultados son satisfactorios, de entre las estructuras encontradas hay algunas más complejas que otras. Esto se ve reflejado en que el error asociado a esas estructuras es ligeramente mayor, o en nuestro caso, ha sido necesario la utilización de un modelo no lineal para lograr errores similares a los obtenidos en otros grupos. Este hecho también puede apreciarse visualmente en las proyecciones sobre el nuevo espacio (véase figura 8.1).

Esto indica que esa estructura puede ser motivo de un segundo análisis o realizar un análisis recursivo, logrando así una búsqueda de estructuras lineales jerarquizadas. Aunque es muy posible que realizando ese tipo de análisis se obtengan nuevas estructuras lineales en el grupo de mayor complejidad, no hemos realizado esas pruebas porque los resultados obtenidos son óptimos para la aplicación presentada. A pesar de esto, queremos indicar que para otras aplicaciones pudiera ser interesante buscar estas jerarquías y obtener una taxonomía de grupos homogéneos, tanto para obtener información descriptiva como para fines de estimación como los estudiados en esta tesis.

Capítulo 9

Conclusiones y resumen de contribuciones

No he fallado. Únicamente he encontrado 10.000 maneras que no funcionan. –Thomas A. Edison 1847-1931

En esta tesis hemos estudiado algunos de los problemas relacionados con la alta dimensionalidad de los datos en el contexto de una aplicación de teledetección. En este contexto, la reducción de la dimensionalidad es un proceso necesario para lograr obtener resultados más precisos y más robustos ante nuevos casos que deseen ser estimados.

Esta reducción ha sido estudiada desde dos puntos de vista: la selección y la extracción de características. En el apartado de selección de características se ha utilizado la estructura de las componentes principales junto con información física del problema para desarrollar un nuevo algoritmo que permite encontrar un subconjunto de características originales relevantes. Con este subconjunto de características seleccionadas se han obtenido mejores resultados que con el método B4, o frente a la utilización de todas las características.

Los errores obtenidos han sido $< 1\,\%$ lo cual puede considerarse como una precisión alta en este tipo de aplicaciones. Además, se ha logrado una reducción significativa del número de características originales logrando una reducción de factor setenta y ocho.

Esto va a permitir diseñar sistemas específicos utilizando únicamente ese subconjunto de longitudes de onda lo que simplifica los sistemas y reduce su coste.

Con respecto a las principales contribuciones de esta parte, han sido:

- Desarrollo de un algoritmo de selección de características no supervisado que permite incluir información previa sobre el problema específico de estimación de temperaturas en teledetección.

- Con el algoritmo desarrollado se ha introducido el modo de resolver algunos de los problemas relacionados con la alta dimensionalidad: características irrelevantes y/o redundantes, y la multicolinealidad, haciendo uso de los coeficientes de las estructuras de las componentes principales.

- Obtención de un subconjunto de características originales, las cuales están asociadas a un subconjunto de longitudes de onda específicos. Esto no sólo ha mejorado los resultados en obtenidos en precisión, sino que además permite el diseño de sensores por bandas con el consiguiente ahorro tanto en coste como en tiempo de computación.

A pesar de los buenos resultados obtenidos en la primera parte, se ha observado que algunos de los perfiles estimados contienen oscilaciones indeseadas. Estas oscilaciones son debidas principalmente a los efectos de la multicolinealidad, la cual es muy severa en la aplicación tratada, y a la heterogeneidad de los datos.

Por eso, la segunda parte de este trabajo se ha basado en la reducción de la dimensionalidad bajo una aproximación de extracción de características y con el principal objetivo de resolver este efecto. Desde un punto de vista de regresión, los efectos de la multicolinealidad son visibles cuando los datos donde ocurre son heterogéneos. Para evitar sus efectos y simultáneamente reducir la dimensionalidad de los datos, se ha desarrollado un algoritmo basado en grafos el cual intenta preservar la localidad de las estructuras en un espacio de menor dimensión logrando así agrupar zonas homogéneas. De este modo se ha divido el problema global de regresión en varios modelos más simples, convirtiendo así el problema inicial en un problema de clasificación+regresión. Con esta aproximación se ha logrado una mejora en los resultados

obtenidos con un error medio de $\approx$ 0,4 K, y se ha resuelto de manera exitosa el problema de las oscilaciones en los perfiles estimados.

Las principales contribuciones de esta segunda parte han sido:

- Se ha definido y desarrollado el concepto de grupos homogéneos como solución al problema de la multicolinealidad en datos de alta dimensión.
- Se ha desarrollado una técnica basada en grafos para descubrir un subespacio de menor dimensión que el original y que contenga los grupos de datos homogéneos correspondientes a dos representaciones de un mismo objeto.
- La técnica desarrollada permite aprender el subespacio que soporta las diferentes estructuras homogéneas, por lo que el proceso de test puede realizarse sin problemas frente a otras técnicas similares que carecen de esta posibilidad.
- Se introduce una nueva perspectiva para abordar problemas de regresión en alta dimensionalidad basada en la búsqueda de estructuras en el espacio de salidas (taxonomía en las estructuras de salida) y transferirlas a las entradas.
- Se han mejorado en precisión los resultados obtenidos hasta ahora, y se puede considerar que las estimaciones realizadas son de una precisión muy alta para la aplicación de teledetección en llamas.

Como lineas de trabajo futuras se proponen la búsqueda de taxonomías de estructuras en el espacio de salidas de una manera recursiva y, la utilización de modelos no-lineales en el análisis de estructuras homogéneas. La extensión recursiva permitirá ir descubriendo los diferentes modelos subyacentes a los datos de una manera iterativa, y posteriormente aplicar modelos de estimación específicos para cada uno de ellos obteniendo resultados más precisos.

En cuanto a la utilización de modelos no-lineales permitiría ampliar el número de posibles aplicaciones debido a que en muchos casos las asunciones de linealidad pueden no ser válidas.

Bibliografía

[1] K. Howard, "Special section: The bioinformatics gold rush.," *Scientific American*, July 2000.

[2] E. P. Xing, M. I. Jordan, and R. Karp, "Feature selection for high-dimensional genomic microarray data," *International Conference on Machine Learning*, 2001.

[3] H. Simon, "Why should machines learn?," *Machine learning: An Artificial Intelligence Approach in R.S. Michalski, J. G. Carbonell and T. M. Mithcell, eds.*, vol. 1, 1983.

[4] T. G. Dietterich and J. W. Shavlik, *Readings in Machine Learning.* Morgan Kaufmann Publishers, Inc., 1990.

[5] S. C. Basak and G. J. Niemi, "Determining structural similarity of chemicals using graph-theoretic indices," *Discrete Applied Mathematics*, vol. 19, pp. 17–44, 1988.

[6] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proceedings of the 11th annual conference on Computational Learning Theory*, pp. 92–100, 1998.

[7] D. Zhou, B. Schölkopf, and T. Hofmann, "Semi-supervised learning on directed graphs.," *Advances in neural information processing systems*, vol. 17, 2005.

[8] G. Camps-Valls, T. V. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 3044–3054, 2007.

[9] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge discovery and data mining: Towards a unifying framework," *Proceedings of the International Conference on Knowledge Discovery and Data Mining, Portland, Oregon*, 1996.

[10] U. von Luxburg, "A tutorial in spectral clustering," *Technical Report No. TR-149 Max Plank Institute for Biological Cybernetics*, August 2006.

[11] C. Romero, K. S. Li, X., and R. Rossow, "Spectrometer-based combustion monitoring for flame stoichiometry and temperature control," *Appl. Therm. Eng.*, vol. 25, pp. 659–676, 2005.

[12] L. H. Liu and J. Jiang, "Inverse radiation problem for reconstruction of temperature profile in axisymmetric free flames," *J. Quant. Spectrosc. Radit. Transfer*, vol. 70, pp. 207–215, 2001.

[13] Y. Deguchi and et. al., "Industrial applications of temperature and species concentration monitoring using laser diagnostics," *Meas. Sci. Technol.*, vol. 13, pp. 103–115, 2002.

[14] N. Docquier and S. Candel, "Combustion control and sensors: a review," *Progress in Energy and Combustion Science*, vol. 28, pp. 107–150, 2002.

[15] N. Afgan, M. Carvalho, P. Pilavachi, A. Tourlidakis, G. Olkhonskiand, and N. Martins, "An expert system concept for diagnosis and monitoring of gas turbine combustion chambers," *Applied Therma Engineering*, vol. 26, pp. 766–771, 2006.

[16] R. M. Goody and Y. Yung, *Atmospheric Radiation. Theoretical basis (Chap.2).* New York: Oxford University Press, 1989.

[17] C. Rodgers, "Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation," *J. Geophys. Res.*, vol. 14, pp. 609–624.

[18] W. J. Blackwell, "Retrieval of atmospheric temperature and moisture profiles from hyperspectral sounding data using a projected principal components transform and a neural network.," *IEEE International Geoscience and Remote Sensing Symposium Proceedings*, June 2003.

[19] E. García-Cuesta, F. de la Torre, and A. J. de Castro, "Advances in computational algorithms and data analysis: Machine learning approaches for the inversion of the radiative transfer equation," *Lectures Notes in Electrical Engineering*, vol. 14, pp. 319–333, 2008.

[20] D. Koller and M. Sahami, "Toward optimal feature selection," in *In Proceedings of the 13th International Conference on Machine Learning*, p. 284–292, 1996.

[21] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence, special issue on relevance*, vol. 97, no. 1-2, pp. 273–324, 1997.

[22] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," *In Proc. 17th International Conference on Machine Learning*, pp. 359–366, 2000.

[23] A. Hinneburg, C. Aggarwal, and D. Keim, "What is the nearest neighbor in high dimensional spaces?," in *In Proceedings 26th International Conference on Very Large Data Bases (VLDB)*, pp. 506–515, Morgan Kaufmann, 2000.

[24] B. Kevin and et. al., "When is nearest neighbor meaningful?," in *In Proceedings of 7th International Conference on Database Theory (ICDT)*, pp. 217–235, 1999.

[25] R. Bellman, "Adaptive control processes: A guided tour," *Princeton University Press.*, 1961.

[26] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.

[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[28] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 153–157, February 1997.

[29] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–859, 2004.

[30] A. Miller, *Subset Selection in Regression*. London: Chapman - Hall, 2002.

[31] I. T. Jollife, *Principal Component Analysis (2nd Ed.)*. New York: Springer Series in Statistics Springer-Verlag (Chap.8), 2002.

[32] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[33] I. Cohen, Q. Tian, X. Sean Zhou, and T. S. Huang, "Feature selection using principal feature analysis," *Proceedings of the 15th international conference on Multimedia*, pp. 301–304, 2007.

[34] F. Chung, "Spectral graph theory," *Regional Conference Series in Mathematics*, no. 2, 1997.

[35] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, pp. 2323–2326, December 2000.

[36] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, December 2000.

[37] M. Gallardo-Campos, *Aplicación de técnicas de clustering para la mejora del aprendizaje*. Carlos III University PFC, 2009.

[38] A. Lecuona, *La contaminación atmosférica y la combustión, Curso de Contaminación Atmosférica*. Universidad Carlos III de Madrid, 1994.

[39] J. Griffiths and J. Bannard, *Flame and Combustion*. Londres: Ed. Blackie Academic & Profesional, 1995.

[40] M. Webber, J. Wang, S. Sanders, D. Baer, and R. Hanson, "In situ combustion measurements of co, co2, h2o and temperature using diode laser absorption sensors," *Proceedings of the Combustion Institute*, vol. 28, pp. 407–413, 2000.

[41] M. Thakur, A. Vyas, and C. Shakher, "Measurement of temperature and temperature profile of an axisymmetric gaseous flames using lau phase interferometer with linear gratings," *Optics and Lasers in Engineering*, vol. 36, pp. 373–380, 2001.

[42] X. Zhou, J. Jeffries, R. Hanson, G. Li, and J. Gutmark, "Wavelength-scanned tunable diode laser temperature measurements in a model gas turbine combustor," *AIAA Journal*, vol. 45, pp. 420–425, 2007.

[43] N. Chigier, *Combustion Measurements Encyclopaedia of Energy Technology and the Environment.* New York 2: Ed. Wiley, 1995.

[44] E. García-Cuesta, *CASIMIR: Cálculos Atmosféricos y Simulación de la Transmitancia en el Infrarrojo.* Madrid (in Spanish): University Carlos III L/PFC 01781, 2003.

[45] L. S. Rothman, "The hitran molecular spectroscopic database: edition of 2000 including updates through 2001," *J. Quant. Spectrosc. Radiat. Transfer*, 2003.

[46] K. Kira and L. Rendell, "The feature selection problem: Traditional methods and a new algorithm," *In Proc. of the 10th National Conference on Artificial Intelligence*, pp. 129–134, 1992.

[47] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation.*

[48] J. Friedman, "Multivariate adaptive regression splines," *Annals of Statistics*, vol. 19, pp. 1–67, March 1991.

[49] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550, July 1994.

[50] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," *In Proc. 20th International Conference on Machine Learning*, pp. 856–863, 2003.

[51] Y. Yang and J. O. Pederson, "A comparative study on feature selection in text categorization," in *In Proceedings of the 14th International Conference on Machine Learning*, p. 412–420, 1997.

[52] E. Xing, M. Jordan, and R. Karp, "Feature selection for high-dimensional genomic microarray data," in *In Proceedings of the 18th International Conference onMachine Learning*, p. 601–608, 2001.

[53] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, "Prediction by supervised principal components," *J. Am. Stat. Assoc.*, vol. 101, pp. 119–137, 2006.

[54] A. E. Hoerl and R. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[55] L. Meier and P. Buhlmann, "Smoothing l1-penalized estimators for high.dimensional time-course data," *Electronic Journal of Statistics*, pp. 597–615, 2007.

[56] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther, "Forward stagewise regression and the monotone lasso," *Electronic Journal of Statistics*, pp. 1–29, 2007.

[57] R. Caruana and D. Freitag, "Greedy attribute selection," in *In Proc. 11th International Conference on Machine Learning*, pp. 28–36, Morgan Kaufmann, 1994.

[58] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems*, vol. 13, pp. 44–49, 1998.

[59] P. Pudil, J.Ñovovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, 1994.

[60] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3.

[61] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, pp. 2429–2437, 2004.

[62] P. Phillips and H. Moon, "Analysis of PCA-based face recognition algorithms," in *AVBPA99*, 1999.

[63] M.-L. O. C. Tom Howley, Michael G. Madden and A. G. Ryder, "The effect of principal component analysis on machine learning accuracy with high dimensional spectral data," *Proceedings of AI-2005, 25th International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge*, 2005.

[64] C. D. Rogers, "Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation," *Review of Geophysics and Space Science*, 2005.

[65] J. R. Eyre, *Inversion methods for satellite sounding data. Lecture Notes NWP Course.* European Centre for Medium-Range Weather Forecasts (ECMWF), 2004.

[66] I. T. Jollife, *Redundant Varibales in Multivariate Analysis.* Unpusblished D. Phil. thesis. University of Sussex.

[67] I. T. Jollife, "Discarding variables in a principal component analysis 1: Artificial data," *Applied Statistics*, vol. 21, pp. 160–173.

[68] E. García-Cuesta, I. M. Galvan, and A. J. de Castro, "Multilayer perceptron as inverse model in a ground-based remote sensing temperature retrieval problem," *J. Eng. Appl. Artif. Intell.*, vol. 21, pp. 26–34, February 2008.

[69] G. Cybenko, "Approximation by superposition of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 303–314, 1989.

[70] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[71] P. Isasi Viñuela and I. Galván León, *Redes de neuronas artificiales: un enfoque práctico*. Monographs and Textbooks on Probability and Mathematical Statistics, Madrid: Prentice Hall, 2004.

[72] F. Aires, A. Chedin, N. Scott, and W. B. Rossow, "A regularized neural net approach for retrieval of atmospheric and surface temperatures with the iasi instrument," *Journal of Applied MEteorology*, vol. 41, pp. 144–159, 2001.

[73] W. J. Blackwell, "A neural-network technique for retrieval of atmospheric temperature and moisture profiles from high spectral resolution sounding data," *IEEE Trans. Geosci. Remote Sens*, vol. 43, pp. 2535–2546, 2005.

[74] G. Lu, Y. Yan, and M. Colechin, "A digital imaging based multifuncional flame monitoring system," *IEEE T. Instrum. Meas.*, vol. 53, pp. 1152–1158, 2004.

[75] I. T. Jollife, "Discarding variables in principal component analysis ii: Real data," *Applied Statistics*, vol. 22, 1973.

[76] H. Hotelling, "Relations between two sets of variants," *Biometrika*, vol. 28, pp. 321–377, 1936.

[77] M. Blaschko and C. Lampert, "Correlational spectral clustering," *CVPR*, 2008.

[78] I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications*. New York: Springer-Verlag, 2005.

[79] M. Belkin and P.Ñiyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing Systems*, pp. 585–591, 2001.

[80] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[81] M. Belkin and P.Ñiyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[82] W. Hwang and J.Weng, "Hierarchical discriminant regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1277–1293, 2000.

[83] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1990.

[84] Y. Shuicheng, D. Xu, B. Zhang, and H.-J. Zhang, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, January 2007.

[85] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, February 2001.

[86] D. Zhu, O. Bousquet, T.Ñ. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[87] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," *ICCV*, 2007.

[88] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* United Kingdom: Cambridge University Press, 2004.

[89] T. Melzer, M. Reiter, and H. Beschof, "Appearance models based on kernel canonical correlation analysis," *Pattern Recognition*, vol. 36, pp. 1961–1973, 2003.

[90] M. Ester, H.-P. Kriegel, S. Jörg, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD'96*, pp. 226–231, 1996.

[91] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[92] G. Lu, Y. Yan, and M. Colechin, "A digital imaging based multifuncional flame monitoring system," *IEEE T. Instrum. Meas.*, vol. 53.