UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICA

Segmentación de Imágenes de Resonancia Magnética Cerebrales Aplicando Análisis de Componentes Principales y Máquinas de Soporte Vectorial

Trabajo Especial de Grado presentado ante la ilustre Universidad Central de Venezuela por el **Br. David Crespo** para optar al título de Licenciado en Matemática.

Tutora: Dra. Glaysar Castro

Caracas - Venezuela

Diciembre 2009

Contenido

Introducción

En la actualidad, las imágenes de resonancia magnética (IRMs) constituyen la principal herramienta de diagnóstico y planeación en el tratamiento de tumores cerebrales, debido a que proveen un medio no invasivo y efectivo para observar el cerebro humano. La detección y ubicación de tumores en el cerebro en estas imágenes la llevan a cabo especialistas que identifican los distintos tejidos de interés en las mismas, sin embargo, ésta no es una tarea sencilla, por lo que es necesario la ayuda de modelos matemáticos para determinar los distintos tejidos considerados, sano, líquido, entre otros. La detección de diferentes regiones de interés con características similares dentro de una imagen se conoce como segmentación de imágenes.

Se han desarrollado distintos trabajos en la segmentación de imágenes de resonancia magnética para detectar tumores cerebrales; dentro de las metodologías utilizadas se encuentran las máquinas de soporte vectorial (MSVs) [23], desarrolladas inicialmente por Vapnik en 1992, basadas en la Teoría de Aprendizaje Estadístico. Estas máquinas hallan una ley de correspondencia entre un conjunto de vectores de $\mathbb{R}^m$ y un conjunto en $\mathbb{R}$. En este trabajo se desarrolla una metodología que detecta, en imágenes de resonancia magnética multieco potenciadas en T2, el tejido invadido por tumores, el tejido sano y el líquido. Uno de los resultados mas destacados es la obtención de un patrón de comportamiento de los tejidos que se caracterizan por regiones de $\mathbb{R}^2$ obtenidas de la distribución conjunta de dos variables que resumen la información de las imágenes. Se construyen MSVs para segmentar las IRMs en base a las regiones mencionadas. Además se realiza un análisis de componentes principales con el objeto de disminuir la dimensión del problema y visualizar los datos espacialmente.

En el primer capítulo se describen el problema de segmentar imágenes de resonancia magnética cerebrales y algunos trabajos en ese sentido y se presenta la propuesta metodológica desarrollada para llevar a cabo la segmentación. El capítulo II es un resumen del análisis de componentes principales, compuesto por su base teórica y sus aplicaciones. En el tercer capítulo se exponen algunas herramientas matemáticas para la resolución de problemas de optimización. Estos se utilizan en las secciones posteriores del capítulo para describir las máquinas de aprendizaje con énfasis en las máquinas de soporte vectorial, y en particular el modelo que que se utiliza para segmentar las imágenes en este trabajo. La metodología conforma el capítulo IV y comprende la recolección de los datos, su preprocesamiento, la implantación y prueba del algoritmo de aprendizaje utilizado y la descripción de la construcción de las distintas máquinas de aprendizaje para las diferentes imágenes de resonancia magnética disponibles. Este trabajo culmina con los resultados obtenidos y las conclusiones, en el capítulo V.

Capítulo 1

Descripción del problema

En las secciones que conforman este capítulo, se expone la importancia de la segmentación de imágenes de resonancia magnética para la detección de tumores cerebrales y diversos trabajos que han sido desarrollados para tal fin.

1.1. Planteamiento del Problema.

Las tecnologías para la adquisición de imágenes de resonancia magnética han tenido un gran desarrollo en los últimos años. Estas tecnologías utilizan imanes y ondas de radiofrecuencia. Mediante los imanes se genera un campo magnético que fuerza a los núcleos de átomos de hidrógeno del organismo a alinearse de cierta manera. A su vez, se envían ondas de radiofrecuencia hacia los núcleos, que alteran su alineamiento y absorben energía de radiofrecuencia. Cuando se interrumpe el envío de ondas, los núcleos vuelven a su posición original en un proceso de relajación, liberando energía y emitiendo señales de radiofrecuencia, que son recogidas por una antena y luego se aplica la transformada de Fourier para obtener los valores con que se construyen imágenes bidimensionales de órganos y tejidos. Estas imágenes corresponden a cortes axiales, coronales o sagitales de la zona estudiada. En el estudio del cáncer en el cerebro se pueden detectar en las imágenes, zonas de alta sospecha tumoral, cuyo diagnóstico debe confirmarse finalmente mediante una biopsia. Las IRMs no son fáciles de interpretar a simple vista, por lo tanto es necesario utilizar herramientas matemáticas y computacionales para ayudar a identificar el tejido invadido por el tumor y otros tejidos de interés. El reconocimiento de los diferentes tejidos es un problema de segmentación de imágenes

que se define clásicamente como la partición de una imagen en regiones no solapadas que son homogéneas con respecto a alguna característica como intensidad o textura; Coto, E. (2003) [4]. El número de regiones se define de acuerdo a la información que se quiere extraer de la imagen.

Existen diferentes tipos de IRMs según los fenómenos que dominen en su formación y se consiguen mediante la aplicación de distintas formas de emitir las ondas de radiofrecuencias (secuencias de pulso) y la modificación de ciertos parámetros para ponderar un determinado efecto a fin de maximizar el contraste entre tejidos que se quiere estudiar. Existen tres tipos de imágenes básicas, definidas de acuerdo a la señal que se recoge en el proceso de relajación, conocidas como potenciaciones. Una de ellas es la densidad protónica donde la intensidad del pixel de la imagen es proporcional a la densidad de núcleos de hidrógeno. El Tiempo de relajación longitudinal o T1, que es el tiempo que tardan los núcleos en liberar el exeso de energía y el tiempo de relajación transversal o T2, que está relacionado con la interacción de los protones. Además, existen secuencias de pulso que permiten obtener diferentes imágenes de un mismo corte, en cuyo caso se habla de imágenes multieco, que son estudiadas en este trabajo.

Debido a las distintas modalidades de las IRMs se puede integrar la información contenida en ellas para realizar estudios robustos en las segmentaciones de estas imágenes. Uno de los retos de la comunidad científica es poder estimar con un alto grado de confianza el límite de cada tejido de interés. El ruido, los cambios graduales de intensidades o la similitud de intensidades de diferentes tejidos hace que la definición de los límites de cada tejido sea una tarea difícil. Otra componente a considerar es la presencia de artefactos, que son perturbaciones de las imágenes que disminuyen su resolución, provocadas por detalles en la obtención de las imágenes tales como problemas en el circuito de la antena de detección de señales y movimientos de los objetos durante la adquisición de las imágenes, entre otras. Es por ello que continuamente se crean diversas metodologías para segmentar este tipo de imágenes en el cerebro.

1.2. Antecedentes.

Con el objetivo de ayudar en la interpretación de las imágenes de resonancia magnética se han desarrollado una serie de trabajos fundamentados en las redes neuronales, técnicas de

umbralización y otros métodos. A continuación se describen brevemente algunas de estas técnicas y se exponen diversos trabajos desarrollados para identificar tumores cerebrales en IRMs.

Tang et al. [20] (2000) segmentaron imágenes de resonancia magnética a través de un proceso basado en métodos de umbralización y región creciente. El primero consiste en crear particiones binarias de las intensidades de la imagen a partir de un valor de intensidad, llamado umbral. El segundo se basa en extraer regiones que están conectadas según cierto criterio predefinido como intensidades y bordes de la imagen.

Martín, M. et al. [13] (2005) utilizaron distintos métodos de segmentación para evaluar e identificar tumores cerebrales a través de información obtenida con diferentes técnicas de resonancia magnética tales como espectroscopía de resonancia magnética protónica en vivo, y relaxometría, donde obtienen resultados que determinan eficientemente la localización espacial del tumor, tanto cualitativa como cuantitativamente.

Las máquinas de aprendizaje conforman otra metodología utilizada para la segmentación de imágenes, se basan en crear modelos que aprenden a hallar las relaciones entre un conjunto de entradas típicamente representadas por vectores m-dimensionales contenidos en el espacio $X = \mathbb{R}^m$ y un conjunto de salidas y_i que pueden tomar valores reales o enteros. Estas máquinas aproximan la ley de correspondencia f entre las entradas y las salidas ($y = f(x)$) y se basan en un proceso de aprendizaje que modifica los parámetros de una función hipótesis a partir de un conjunto formado por pares entrada-salida, con el fin de obtener la hipótesis que mejor se ajuste a la relación de los datos; Moreno, J. (2004) [15].

Muchas de estas máquinas se basan en un modelo de clasificación binaria que utiliza la hipótesis lineal $f(x) =< w, x > +b$, donde $<,>$ denota el producto escalar en $\mathbb{R}^m$ y $w \in \mathbb{R}^m$ y $b \in \mathbb{R}$ son los parámetros que la definen. El proceso de aprendizaje tiene como objetivo obtener un hiperplano de ecuación $< w, x > +b = 0$ que divide al espacio de entrada en dos semiespacios, donde deben encontrarse en cada uno entradas de una misma clase.

Este tipo de clasificador ha sido utilizado en estadística y por las redes neuronales artificiales (RNAs), máquinas que emulan las redes neuronales biológicas, las cuales pueden generar resultados confiables aún cuando la información sea parcialmente completa y con cierto contenido de ruido. Razones que llevaron a introducirlas en el área del procesamiento digital de imágenes como una técnica para segmentar imágenes de diferentes tipos; Reddick et al [16] (1998) y Drozdowicz, B. Bernasconi, G. et al [6] (2005).

Otro tipo de máquinas de aprendizaje que utiliza el clasificador binario descrito anteriormente son las llamadas máquinas de soporte vectorial (MSVs), cuyo algoritmo de aprendizaje obtiene los parámetros w y b que definen al hiperplano separador que divide los datos. El hiperplano que divide los datos mediante una distancia conocida como el margen se conoce como el hiperplano de máximo margen y es el que esta mas alejado de los datos de entrenamiento de ambas clases. En la mayoría de los casos no se puede separar los datos mediante un hiperplano, por lo que se introduce una función kernel que transforma los datos a un espacio de rasgos donde es posible separarlos por un hiperplano. Existen diversos algoritmos de aprendizajes para estas máquinas que resuelven el problema de clasificación, por ejemplo Jianguo Zhang y Vincent Chong [23] identificaron tumores en imágenes de RM a través de dos modelos de MSVs. El primero se basa en datos de entrenamiento que consisten sólo en vectores que contienen información de la imagen en la región del tumor. El otro modelo admite en su entrenamiento vectores con información dentro y fuera del tumor. En ambos casos se obtuvieron resultados bastantes satisfactorios, logrando detectar altos porcentajes de zonas tumorales, sin embargo, cabe destacar que el primer modelo, que utiliza solo una clase en el entrenamiento generó mejores resultados.

Capítulo 2

Análisis de Componentes Principales

En este capítulo se expone la técnica del análisis de componentes principales, la cual va a ser utilizada en el preprocesamiento de los datos para eliminar errores aleatorios y sistemáticos y para reducir la dimensión del problema en la segmentación de las imágenes de resonancia magnética. Se comienza con algunos resultados de algebra lineal necesarios para describir dicha metodología.

2.1. Algebra Lineal

La notación que se utiliza es la siguiente:

- Las matrices se denotan con letras mayúsculas, por ejemplo X y A.

- $\mathcal{M}_{t\times p}(\mathbb{R})$ denota el conjunto de matrices con entradas reales con t filas y p columnas.

- Los vectores en $\mathbb{R}^n$, con $n \in \mathbb{Z}$, son vectores columnas.

- La fila i-ésima de la matriz X se denota por X^i:

$$X^i = (x_{i1}, x_{i2}, \ldots, x_{ip}),$$

de manera que, la matriz X de dimensión $t \times p$ se escribe como:

$$X = \begin{pmatrix} X^1 \\ X^2 \\ \vdots \\ X^t \end{pmatrix}.$$

- La columna j-ésima de la matriz X se denota por X_j:

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{tj} \end{pmatrix},$$

por lo tanto la matriz X se expresa como:

$$X = (X_1, X_2, \ldots, X_p).$$

Para enunciar algunas propiedades de matrices se definen algunos espacios fundamentales asociados a $\mathcal{M}_{t\times p}(\mathbb{R})$:

1. El Espacio Nulo de X es el subespacio vectorial $\{w \in \mathbb{R}^p / Xw = 0\}$ y se denota por $EN(X)$.

2. El Espacio Fila de X es el subespacio de $\mathbb{R}^p$ generado por los vectores filas de X y se denota por $ER(X)$.

3. El Espacio columna de X es el subespacio de $\mathbb{R}^t$ generado por los vectores columnas de X y se denota por $EC(X)$.

El número de filas y columnas linealmente independientes de una matriz X es el mismo y es conocido como rango, el cual se denota por $rg(X)$. De manera que $rg(x) = dimER(X) = dimEC(X)$. A continuación se presenta un teorema que relaciona las dimensiones de los espacios fundamentales de X con su rango.

Teorema 2.1 Sea $X \in \mathcal{M}_{t\times p}(\mathbb{R})$, entonces:

1. dim $ER(X)$ + dim $EN(X) = p$

2. dim $EC(X)$ + dim $EN(X) = p$

3. dim $rg(X)$ + dim $EN(X) = p$

Ver demostración de Barbolla en [2].

Definición 2.1 Sea una matriz cuadrada $A \in \mathcal{M}_{p\times p}(\mathbb{R})$, un valor propio o autovalor de A es un escalar λ en $\mathbb{R}$ tal que existe un vector $v \neq 0$ en $\mathbb{R}^p$ que satisface $Av = \lambda v$. El vector v se conoce como vector propio de A asociado al valor propio λ.

Teorema 2.2 Sea $A \in \mathcal{M}_{p\times p}(\mathbb{R})$, una matriz simétrica con rango $r < p$ entonces A tiene exactamente r autovalores distintos de cero.

Ver demostración de Hoffman en [10].

Teorema 2.3 Sea $X \in \mathcal{M}_{t\times p}(\mathbb{R})$, entonces los autovalores de la matriz $X'X$ son reales y no negativos.

Demostración : Sea v_i el i-ésimo autovector no nulo de $X'X$ y λ_i su autovalor asociado, entonces

$$X'Xv_i = \lambda_i v_i,$$

premultiplicando la igualdad anterior por v_i' se tiene que

$$v_i'X'Xv_i = v_i'\lambda_i v_i \quad \Rightarrow \quad (Xv_i)'Xv_i = \lambda_i v_i'v_i,$$

y como $\|Xv_i\|^2 = (Xv_i)'Xv_i$ y $\|v_i\|^2 = (v_i)'v_i$ entonces

$$\|Xv_i\|^2 = \lambda_i\|v_i\|^2,$$

$v_i \neq 0$ por ser autovector, de manera que

$$\lambda_i = \frac{\|Xv_i\|^2}{\|v_i\|^2} \geq 0,$$

lo cual demuestra que los autovalores de $X'X$ son números reales no negativos.

□

Teorema 2.4 Sea $X \in \mathcal{M}_{t\times p}(\mathbb{R})$, entonces la matriz $X^{'}X$ es simétrica y sus autovectores son ortogonales.

Demostración : La matriz $X^{'}X$ es simétrica, ya que $(X^{'}X)^{'} = X^{'}(X^{'})^{'} = X^{'}X$. Ahora bien, para ver la ortogonalidad de los autovalores se considera v_i el autovector i-ésimo de $X^{'}X$ y λ_i su autovalor correspondiente, por lo que se cumple

$$X^{'}Xv_i = \lambda_i v_i. \tag{2.1}$$

Además, sea v_j un autovector de $X^{'}X$ y λ_j su autovalor asociado, donde $j \neq i$. Se premultiplica la igualdad 2.1 por $v_j^{'}$ y por ser $X^{'}X$ simétrica, $v_j^{'}(X^{'}X) = ((X^{'}X)^{'}v_j)^{'} = (X^{'}Xv_j)^{'}$, de donde se obtiene que

$$v_j^{'}X^{'}Xv_i = v_j^{'}\lambda_i v_i \quad \Rightarrow \quad (X^{'}Xv_j)^{'}v_i = \lambda_i v_j^{'}v_i,$$

y por ser v_j un autovector de $X^{'}X$, la ecuación anterior queda de la siguiente forma

$$(\lambda_j v_j)^{'}v_i = \lambda_i v_j^{'}v_i \quad \Rightarrow \quad \lambda_j v_j^{'}v_i = \lambda_i v_j^{'}v_i \quad \Rightarrow \quad (\lambda_j - \lambda_i)v_j^{'}v_i = 0.$$

Los autovalores son diferentes, por lo que sus vectores propios también lo son. De manera que la última ecuación implica que $v_j^{'}v_i = 0$, lo que quiere decir que los vectores propios de $X^{'}X$ son ortogonales.

□

Teorema 2.5 Sea $X \in \mathcal{M}_{t\times p}(\mathbb{R})$ con rango r, entonces las matrices $X^{'}X$ y $XX^{'}$ tienen los mismos autovalores no nulos.

Demostración : Las matrices $X^{'}X$ y $XX^{'}$ son simétricas de rango r, por lo tanto cada una tiene r autovalores distintos de cero. Sea $\lambda_i \neq 0$, con $i = 1 \dots r$, el autovalor i-ésimo de $X^{'}X$, entonces existe $v_i \neq 0$, un autovector asociado, tal que

$$X^{'}Xv_i = \lambda_i v_i. \tag{2.2}$$

Premultiplicando por X la ecuación anterior, se obtiene que

$$XX^{'}Xv_i = X\lambda_i v_i \quad \Rightarrow \quad (XX^{'})Xv_i = \lambda_i Xv_i.$$

Para probar que λ_i es un autovector no nulo de $X^{'}X$ basta demostrar que $Xv_i \neq 0$. Para ello supongamos lo contrario, entonces al considerar la ecuación 2.2 se obtiene que

$$\lambda_i v_i = X^{'}Xv_i = X^{'}0 = 0,$$

lo cual es una contradicción ya que λ_i y v_i son distintos de cero. Por lo tanto los autovalores no nulos de $X^{'}X$ son los mismos de $XX^{'}$. □

Teorema 2.6 Sea $X \in \mathcal{M}_{t\times p}(\mathbb{R})$ con rango r. Sean V y U las matrices cuyas columnas son los autovectores normalizados asociados a los autovalores comunes no nulos de las matrices $X^{'}X$ y $XX^{'}$ respectivamente. Entonces

$$v_i = \tfrac{1}{\sqrt{\lambda_i}} X^{'}u_i \in \mathbb{R}^p,$$
$$u_i = \tfrac{1}{\sqrt{\lambda_i}} Xv_i \in \mathbb{R}^t,$$

donde $i = 1, 2, 3, \ldots, r$.

Demostración : Por el teorema 2.5 $X^{'}X$ y $XX^{'}$ tienen los mismos autovalores. Sea λ_i el i-ésimo autovalor común de los mismos y $u_i \neq 0$ el autovector de $XX^{'}$ asociado, entonces

$$(XX^{'})u_i = \lambda_i u_i, \tag{2.3}$$

premultiplicando por $X^{'}$ a 2.3, se tiene que

$$(X^{'}XX^{'})u_i = X^{'}\lambda_i u_i \quad \Rightarrow \quad (X^{'}X)X'u_i = \lambda_i X'u_i.$$

De las dos últimas ecuaciones se deduce que los autovectores de $X'X$ asociados a λ_i son de la forma $v_i = \beta X^{'}u_i$ para algún $\beta \neq 0$. Para que los v_i tengan una norma unitaria, necesariamente $\beta = \frac{1}{\sqrt{\lambda_i}}$, ya que

$$\begin{aligned} 1 &= \|v_i\|^2 = v_i^{'}v_i = (\beta X^{'}u_i)^t(\beta X^{'}u_i) = \beta^2 u_i^{'}XX^{'}u_i \\ &= \beta^2 u_i^{'}\lambda_i u_i = \beta^2\lambda_i \|u_i\|^2 = \beta^2\lambda_i, \end{aligned}$$

además,

$$\begin{aligned} < v_i, v_j > &= v_i^{'} v_j \\ &= \left(\frac{1}{\sqrt{\lambda_i}} X^{'} u_i\right)^{'} \frac{1}{\sqrt{\lambda_j}} X^{'} u_j \\ &= \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}} u_i^{'} X X^{'} u_j \\ &= \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}} u_i^{'} \lambda_j u_j \\ &= \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}} \lambda_j < u_i, u_j > . \end{aligned}$$

Por la ortonormalidad de los u_i se tiene que $< v_i, v_j >= 0$ para $i \neq j$ y en otro caso $< v_i, v_i >= \frac{1}{\lambda_i}\lambda_i < u_i, u_i >= 1$.

Entonces los vectores normalizados de $X^{'}X$, asociados a λ_i, son $v_i = \frac{1}{\sqrt{\lambda_i}} X^{'} u_i$. Y despejando de esta expresión, se tiene que

$$\sqrt{\lambda_i} v_i = X^{'} u_i, \tag{2.4}$$

premultiplicando 2.4 por X, y por ser λ_i un autovalor de $XX^{'}$, se sigue que

$$\begin{aligned} X\sqrt{\lambda_i} v_i &= X X^{'} u_i \implies \\ \sqrt{\lambda_i} X v_i &= \lambda_i u_i \implies \\ \frac{\sqrt{\lambda_i} X v_i}{\lambda_i} &= u_i \implies \\ \frac{1}{\sqrt{\lambda_i}} X v_i &= u_i. \end{aligned}$$

□

Teorema 2.7 Sea $X \in \mathcal{M}_{t\times p}(\mathbb{R})$ con rango $r \leq min\{t, p\}$, V y U matrices cuyos vectores columnas son los autovectores normalizados asociados a los autovalores nulos y no nulos de las matrices $X^{'}X$ y $XX^{'}$, denotados por $v_i \in \mathbb{R}^p$ y $u_j \in \mathbb{R}^t$, con $i = 1, 2, 3, \ldots, p$ y $j = 1, 2, 3, \ldots, t$. Entonces X puede factorizarse como

$$X = USV^{'},$$

donde $S \in \mathcal{M}_{t\times p}(\mathbb{R})$, verifica

$$S(i,j) = \begin{cases} s_{ij} = \sigma_i = \sqrt{\lambda_i} & \text{si} \quad i = j \\ s_{ij} = 0 & \text{en otro caso} \end{cases},$$

los λ_i, $i = 1, \ldots, r$, son los autovalores comunes no nulos y los σ_i son números reales positivos que se pueden ordenar, $\sigma_1 \geq \ldots \geq \sigma_r$ y son conocidos como valores singulares. Además,

$$X = \sum_{i=1}^{r} \sqrt{\lambda_i} u_i v_i^{'}.$$

Demostración : Los autovalores de $X^{'}X$ son reales positivos y sus respectivos autovectores son ortogonales por los Teoremas 2.3 y 2.4 respectivamente. Los autovalores se ordenan de manera que $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$ y se denotan por $\{v_1, \ldots v_p\}$ a los autovectores ortonormales correspondientes, se supone que λ_r es el menor de los autovalores no nulos, es decir

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_r > 0, \quad \lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p = 0,$$

entonces,

$$XX^{'}v_i = \lambda_i v_i, \qquad y \qquad < v_i, v_j >= \begin{cases} 0 & \text{si} \quad i \neq j \\ 1 & \text{si} \quad i = j \end{cases}.$$

En particular, para $i > r$, $X^{'}Xv_i = \lambda_i v_i = 0v_i = 0$ y como $EN(X^{'}X) = EN(X)$, entonces $Xv_i = 0$. Sea V la matriz ortogonal cuyas columnas son los v_i, $V = (v_1, \ldots, v_p)$. Para $1 \leq i \leq \ldots r$, por el Teorema 2.6,

$$u_i = \frac{1}{\sqrt{\lambda_i}} X v_i, \tag{2.5}$$

son vectores ortonormales de $\mathbb{R}^t$. Completando hasta obtener una base y utilizando el proceso de ortogonalización de Gram-Schmidt se eligen $u_{r+1}, \ldots, u_t$ de manera que sean ortonormales. Sea $S \in \mathcal{M}_{t\times p}$, tal que $S = U^{'}XV$, donde $U \in \mathcal{M}_{t\times t}(\mathbb{R})$ es la matriz ortogonal cuyas columnas son los u_i, $U = (u_1, u_2, \ldots, u_t)$, entonces,

$$S = U^{'}(Xv_1, Xv_2, \ldots, Xv_p) = (U^{'}Xv_1, U^{'}Xv_2, \ldots, U^{'}Xv_p),$$

como la i-ésima fila de $U^{'}$ es la i-ésima columna de U

$$s_{ij} = u_i^{'}Xv_j =< u_i, Xv_j > .$$

Si $j > r,\ Xv_j = 0$, por lo tanto $s_{ij} =< u_i, 0 >= 0$, y en caso contrario, por 2.5 $Xv_j = \sqrt{\lambda_i}u_j$, de donde se obtiene,

$$s_{ij} =< u_i, \sqrt{\lambda_i}u_j >= \sqrt{\lambda_i} < u_i, u_j >= \begin{cases} \sigma_i = \sqrt{\lambda_i}, & \text{si} \quad i = j \quad j \leq r \\ 0 & \text{en otro caso} \end{cases} .$$

Como $S = U'XV$ y U y V ortogonales, entonces $X = USV'$, esta expresión se conoce como la descomposición en valores singulares de X. Además, al postmultiplicar una matriz por otra matriz diagonal se multiplica cada columna por el elemento de la diagonal correspondiente. Es por esto que

$$\begin{aligned} X &= USV' \\ &= \begin{bmatrix} u_1 & u_2 & \dots & u_t \end{bmatrix} \cdot \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\lambda_r} & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 \end{bmatrix} \cdot \begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_r' \\ \vdots \\ v_p' \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{\lambda_1}u_1 & \sqrt{\lambda_2}u_2 & \dots & \sqrt{\lambda_r}u_r & \dots & 0 \end{bmatrix} \cdot \begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_r' \\ \vdots \\ v_p' \end{bmatrix} \\ &= \sqrt{\lambda_1}u_1v_1' + \sqrt{\lambda_2}u_2v_2' + \ldots + \sqrt{\lambda_r}u_rv_r' \\ &= \sum_{i=1}^{r} \sqrt{\lambda_i}u_iv_i'. \end{aligned}$$

□

2.2. Análisis de Componentes Principales

El análisis de componentes principales (ACP), es una técnica estadística que fue propuesta a principios del siglo XX por Karl Pearson. La complejidad de los cálculos

implícitos retrasó su desarrollo hasta la aparición de los computadores y su utilización en la segunda mitad del siglo XX.

El objetivo principal que persigue el ACP es la representación de las medidas numéricas de varias variables en un espacio de pocas dimensiones donde nuestros sentidos puedan percibir relaciones que de otra manera permanecerían ocultas en dimensiones superiores. La técnica consiste en encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables no correlacionadas, denominadas componentes principales, que se obtienen en orden decreciente de importancia. Las componentes son combinaciones lineales de las variables originales y se espera que, solo unas pocas (las primeras) recojan la mayor parte de la variabilidad de los datos, obteniéndose una reducción de la dimensión en los mismos.

La descripción de esta técnica fue introducida por Pearson (1901) y Hotelling (1933) con dos enfoques diferentes. El primero se basa en ajustar rectas y/o planos a una nube de puntos utilizando el criterio de mínimos cuadrados, conocidos como como rectas, planos o espacios de regresión [17].

La construcción del espacio de mejor ajuste se realiza en varias etapas. Primero se obtiene la recta de regresión que es el espacio unidimensional que mejor explica los datos. Posteriormente se obtiene una recta perpendicular a la anterior que mejor ajusta los errores de la primera regresión. Los vectores directores de las rectas obtenidas forman una base ortonormal del subespacio bidimensional que mejor explica la data. Luego se ajusta una recta a los residuos de la regresión del plano obtenido y así se van incorporando sucesivamente nuevas dimensiones para obtener r vectores ortogonales que forman una base en el subespacio r-dimensional que mejor ajusta los datos.

El enfoque de Hotelling, por su parte, se basa en una transformación lineal que se aplica a un conjunto de variables aleatorias para generar nuevas variables $\{Y_i \in \mathbb{R}^t / i = 1, \ldots, p\}$ llamadas componentes principales en las que se imponen las siguientes restricciones

- Las Y_i deben ser no correlacionadas.
- $Var(Y_1) > Var(Y_2) > \ldots > Var(Y_p)$.

Si las variables originales no están correlacionadas, las componentes principales son ellas mismas, en el caso de que dicha correlación fuese muy alta, probablemente se explicará con pocas componentes la información que dichas variables contienen. La primera componente es la mas importante por ser la que explica el mayor porcentaje de la varianza de los datos, siguiéndole la segunda y así sucesivamente. Ahora bien, queda a criterio del investigador decidir cuántos componentes se elegirán en el estudio.

El ACP no necesita realizar ninguna suposición acerca de la distribución de probabilidad de las variables originales. Entre los usos más frecuentes del ACP están [22], [3]:

1) Como técnica de análisis exploratorio que permite descubrir interrelaciones entre los datos y de acuerdo con los resultados, proponer los análisis estadísticos más apropiados.

2) Reducir la dimensionalidad de la matriz de datos con el fin de evitar redundancias y destacar relaciones. En la mayoría de los casos, tomando sólo las primeras componentes, se puede explicar la mayor parte de la variación total contenida en los datos originales.

3) Construir variables no observables (componentes) a partir de variables observables.

4) Eliminar los errores aleatorios y sistemáticos presentes en un conjunto de datos.

5) Bajo ciertas circunstancias, es de gran utilidad usar estas componentes no correlacionadas, como datos de entrada para otros análisis. Por ejemplo, en el caso de la regresión múltiple cuando las variables independientes presentan alta colinealidad es preferible hacer la regresión sobre las componentes principales en lugar de usar las variables originales. Además, en otros casos, realizar un diagrama de dispersión de las primeras componentes principales puede permitir encontrar grupos en los datos o contrastar similitudes o diferencias entre los individuos, por ende adicionalmente pueden ser utilizados para alimentar máquinas de aprendizaje.

A pesar de la aparente simplicidad de la técnica, todavía se investiga en el área general del ACP. La metodología se aplica en gran variedad de áreas como lo son la agricultura, biología, química, climatología, demografía, ecología economía, geología, meteorología, psicología, control de calidad, etc. Cabe destacar que la Descomposición en Valores Singulares permite obtener las componentes principales en forma directa

2.3. Descomposición en Valores Singulares

En el ACP, las componentes principales se pueden obtener a través de la descomposición de valores singulares de la matriz de datos. En esta sección se describe como obtenerlas a través de dicha descomposición. Se considera la matriz de datos definida por vectores aleatorios columnas X_i correlacionados, $X = (X_1, X_2, \ldots, X_p)$. Se supone, sin perdida de generalidad, que las variables X_i están centradas, es decir, tienen media cero. Por el teorema 2.7, la matriz X se reescribe como

$$X = \sum_{i=1}^{r} \sqrt{\lambda_i} u_i v_i^{'},$$

y en forma matricial $X = USV^{'}$, donde r es el rango de la matriz, V y U son las matrices cuyos vectores columnas son los autovectores asociados a los autovalores de las matrices $X^{'}X$ y $XX^{'}$ respectivamente, y S es la matriz diagonal de los valores singulares de X.

Ahora bien, como V es ortogonal $XV = US$ y considerando $Y = US = XV$, cuyas columnas son variables aleatorias centradas, pues las variables originales son centradas, se tiene que

$$\begin{aligned} Cov(Y) &= E(Y^{'}Y) = E((US)^{'}(US)) = E(S^{'}U^{'}US) = S^{'}E(U^{'}U)S = S^{'}E(I)S = S^{'}IS \\ &= S^{'}S \end{aligned}$$

$$= \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & 0 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sqrt{\lambda_r} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \ldots & 0 & \ldots & 0 \end{bmatrix} \cdot \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & 0 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sqrt{\lambda_r} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \ldots & 0 & \ldots & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_r & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 & \dots & 0 \end{bmatrix}.$$

De esta manera las columnas Y_i, $i = 1, \dots, p$, no están correlacionadas y la varianza de cada Y_i es λ_i y por el Teorema 2.7, se cumple que

$$\lambda_1 > \lambda_2 > \dots \lambda_r > 0, \quad \lambda_{r+1} = \lambda_{r+2} = \dots = \lambda_p = 0,$$

por lo tanto

$$Var(Y_1) > Var(Y_2) > \dots > Var(Y_p), \quad Var(Y_{r+1}) = Var(Y_{r+2}) = \dots = Var(Y_p) = 0.$$

Las Y_i, con $i > r$, no se consideran ya que sus varianzas son cero, es decir, no aportan información relevante sobre los datos. De manera que las $\{Y_i = XV_i\}_{i=1:r}$ son las componentes principales, combinaciones lineales de las variables aleatoria originales, presentadas en orden decreciente de varianza.

Capítulo 3

Máquinas de Aprendizaje Lineales en Clasificación

En este capítulo se presentan las definiciones y características de las máquinas de aprendizaje, en especial las Máquinas de Soporte Vectorial (MSVs) con énfasis en problemas de clasificación. En las dos primeras secciones se encuentra un resumen sobre la teoría de optimización mediante la cual es posible plantear los problemas de máquinas de aprendizaje, permitiendo la caracterización de las soluciones y por ende la construcción de algoritmos de aprendizaje. Y por último se explica la técnica del ascenso del gradiente, utilizada para resolver un problema de maximización planteado en las MSVs.

3.1. Teoría de Optimización

Definición 3.1 Dadas las funciones $f,\ g_i,\ i = 1, \ldots, k\ y\ h_j,\ j = 1, \ldots, n$, definidas sobre el conjunto $\Omega \subseteq \mathbb{R}^m$, el problema de optimización asociado se define como

$$\begin{array}{lll} Minimizar & f(w), & w \in \Omega \\ sujeto\ a & g_i(w) \leq 0, & i = 1, \ldots, k; \\ & h_j(w) = 0, & j = 1, \ldots, n; \end{array} \tag{3.1}$$

donde f es la función objetivo, y $g_i\ y\ h_j$ son las restricciones de igualdad y desigualdad respectivamente. El valor óptimo de la función objetivo se conoce como el valor del problema de optimización [5].

Definición 3.2 A un problema de optimización para el cual la función objetivo y las funciones de igualdad y desigualdad son lineales se le conoce como problema lineal, mientras que si la función objetivo es cuadrática y las funciones de igualdad y desigualdad son lineales se le conoce como problema cuadrático.

Definición 3.3 Un conjunto $\Omega \subseteq \mathbb{R}$ es convexo si, para todo $w, u \in \mathbb{R}^m$ el conjunto

$$[w, u] = \{z : z = \alpha w + (1 - \alpha)u,\ 0 \leq \alpha \leq 1\}$$

esta contenido en Ω.

Definición 3.4 Una función real f es convexa en $\mathbb{R}^m$ si, $\forall w, u \in \mathbb{R}^m$ y para cualquier $\theta \in (0, 1)$ se cumple

$$f(\theta w + (1 - \theta)u) \leq \theta f(w) + (1 - \theta)f(u),$$

si la desigualdad es estricta, la función es estrictamente convexa. Geométricamente esto significa que la imagen directa mediante f del segmento de recta que une a los puntos w y u es menor o igual al segmento de recta que une a los puntos $f(w)$ y $f(u)$.

Definición 3.5 Un problema de optimización en el cual el conjunto Ω, la función objetivo y todas las restricciones son convexas, se denomina convexo, y si Ω es un conjunto convexo, la función objetivo es cuadrática y convexa y las restricciones lineales, al problema se le conoce como cuadrático convexo.

Uno de los métodos utilizados para resolver problemas de optimización es la teoría de Lagrange, basada en los conceptos de multiplicadores de Lagrange y funciones Lagrangeanas. Este método fue desarrollado por Lagrange en 1797 para problemas con solo restricciones de igualdad, extendiendo un resultado de Fermat de 1629, en el cual no considera restricción alguna. Posteriormente en 1951 Kuhn y Tucker generalizan la teoría de Lagrange introduciendo restricciones de desigualdad al problema de optimización [5].

Definición 3.6 Sea el problema de optimización 3.1 con dominio $\Omega \subseteq \mathbb{R}^m$

$$\begin{array}{lll} Minimizar & f(w), & w \in \Omega \\ sujeto\ a & g_i(w) \leq 0, & i = 1, \ldots, k; \\ & h_j(w) = 0, & j = 1, \ldots, n; \end{array}$$

se define la función Lagrangeana generalizada de la siguiente manera:

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{n} \beta_j h_j(w),$$

donde $\alpha \in \mathbb{R}^k$ y $\beta \in \mathbb{R}^n$, y α_i y β_j sus respectivas componentes, las cuales son conocidas como multiplicadores de Lagrange.

Definición 3.7 El problema dual del problema de optimización original de la definición 3.1 se define como:

$$\begin{array}{lll} Maximizar & \theta(\alpha, \beta), & \\ sujeto\ a & \alpha_i \geq 0, & i = 1, \ldots, k; \end{array} \tag{3.2}$$

donde $\theta(\alpha, \beta) = \text{ínf}_{w \in \Omega} L(w, \alpha, \beta)$.

En general el valor del problema original y el valor del problema dual no son necesariamente iguales, lo son bajo ciertas condiciones, y en tales casos el valor alcanzado por ambos enfoques es el valor del problema de optimización. A la diferencia entre ambos valores se le conoce como brecha de dualidad. Una manera de detectar si dicha brecha es cero es por la presencia de un punto de silla de la función lagrangeana generalizada.

Definición 3.8 Un punto de silla de la función Lagrangeana generalizada es una tripleta (w^*, α^*, β^*) *con* $w^* \in \Omega$ *y* $\alpha^* \geq 0$, que satisface

$$L(w^*, \alpha, \beta) \leq L(w^*, \alpha^*, \beta^*) \leq L(w, \alpha^*, \beta^*),$$

para todo $w \in \Omega$ *y* $\alpha \geq 0$.

Teorema 3.1 La tripleta (w^*, α^*, β^*) es un punto de silla de la función Lagrangeana generalizada si y solo si sus componentes son soluciones óptimas de los problemas de optimización original 3.1 y dual 3.2 respectivamente, cumpliendo $f(w^*) = \theta(\alpha^*, \beta^*)$ [5].

Teorema 3.2 Dado un problema de optimización 3.1 con dominio convexo $\Omega \in \mathbb{R}^m$

$$\begin{array}{lll} Minimizar & f(w), & w \in \Omega \\ sujeto\ a & g_i(w) \leq 0, & i = 1, \ldots, k; \\ & h_j(w) = 0, & j = 1, \ldots, n. \end{array}$$

Si g_i y h_j son funciones afines, es decir,

$$h(w) = Aw - b,$$

para alguna matriz A y algún vector b, entonces la brecha de dualidad es cero, i.e. $(\theta(\alpha^*, \beta^*) = f(w^*))$ [5].

Este teorema garantiza que la solución del problema dual y primal tienen el mismo valor para los problemas de optimización. Ahora podemos citar el teorema de Kuhn-Tucker el cual da las condiciones para la solución óptima del problema de optimización de la definición 3.1.

Teorema 3.3 (Kuhn-Tucher) Dado un problema de optimización 3.1 con dominio convexo $\Omega \in \mathbb{R}^m$

$$\begin{array}{lll} Minimizar & f(w), & w \in \Omega \\ sujeto\ a & g_i(w) \leq 0, & i = 1, \ldots, k; \\ & h_j(w) = 0, & j = 1, \ldots, n; \end{array}$$

con $f \in C^1$ convexa y g_i y h_j afines, las condiciones necesarias y suficientes para que un punto w^* sea óptimo es la existencia de α^* y β^*, tales que

$$\begin{aligned} \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial w} &= 0, \\ \frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial \beta} &= 0, \\ \alpha_i^* g_i(w^*) &= 0, \\ \alpha_i^* &\geq 0. \end{aligned}$$

Ver demostración en [18].

La relación $\alpha_i^* g_i(w^*) = 0$ se conoce como la condición complementaria de Kuhn-Tucker. Esta implica que solo las restricciones de desigualdad activas $(g_i(w^*) = 0)$ van a tener variables duales (multiplicadores de lagrange) distintas de cero, esto significa que para ciertas optimizaciones el número de variables duales envueltas en la expresión de la solución va a ser mucho menor que el número inicial de variables.

Los resultados anteriores permiten usar la descripción dual para resolver el problema original evitando trabajar con las desigualdades. Para transformar el problema original al dual, por el teorema 3.1 se debe encontrar el punto de silla de la función lagrangeana generalizada, es decir, minimizar dicha función con respecto a las variables originales, y luego maximizarla con respecto a las variables duales. Esto no es otra cosa que calcular $\theta(\alpha, \beta) = \text{ínf}_{w\in\Omega} L(w, \alpha, \beta)$ y resolver el problema dual.

3.2. Ascenso del Gradiente

En esta sección se presenta un método iterativo mediante el cual se hallan los valores donde son alcanzados los máximos de funciones de $\mathbb{R}^m$ a valores reales. Si el gradiente de la función $f : \mathbb{R}^m \to \mathbb{R}$ en el punto $x^0 \in \mathbb{R}^m$, es distinto del vector nulo, entonces es ortogonal al vector tangente a la curva de nivel $\{x \in \mathbb{R}^m / f(x) = f(x^0)\}$ que pasa por el punto x^0, además la dirección de máxima tasa de crecimiento de una función diferenciable a valores reales en el punto x^0 es ortogonal a esta curva de nivel. Por lo tanto, la función f en un punto x^0 crece en la dirección del gradiente mas que en cualquier otra. Para probar esto, se considera la derivada direccional de f en la dirección del vector unitario $d \in \mathbb{R}^m$, $\|d\| = 1$, definida como $\langle \nabla f(x), d\rangle$, que no es mas que la tasa de cambio de f en la dirección de d en el punto x. Por la desigualdad de Cauchy - Schwarz,

$$|\langle \nabla f(x), d\rangle| \leq \|\nabla f(x)\| \|d\|,$$

y como $\|d\| = 1$, se obtiene que

$$|\langle \nabla f(x), d\rangle| \leq \|\nabla f(x)\|,$$

si $d = \dfrac{\nabla f(x)}{\|\nabla f(x)\|}$, entonces

$$\left| \left\langle \nabla f(x), \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\rangle \right| = \|\nabla f(x)\|.$$

Esto prueba que la dirección del $\nabla f(x)$ es la de mayor crecimiento de f en el punto x, de manera que esta dirección es conveniente tomarla en cuenta para buscar un punto donde la función alcanza un máximo. Sea el punto $x^0 \in \mathbb{R}^m$, veamos que la función f es mayor en $x^0 + \lambda \nabla f(x^0)$

que en x^0, para ciertos $\lambda \in \mathbb{R}^+$. A continuación se presenta la definición del polinomio de Taylor, necesaria para probar lo expuesto anteriormente.

Definición 3.9 Sea una función $f : \mathbb{R}^m \to \mathbb{R}$ de clase C^n, el diferencial de f de orden n en el punto $a \in \mathbb{R}^m$, $[D^n f(a)] : (\mathbb{R}^m)^n \to \mathbb{R}$, se define como

$$[D^n f(a)]\,(h^n) = \sum_{i_1,\ldots,i_n=1}^{m} \frac{\partial^n f}{\partial x_{i_1}\partial x_{i_2}\ldots\partial x_{i_n}}(a) h_{i_1} h_{i_2}\ldots h_{i_n},$$

donde $h \in \mathbb{R}^m$ y $h^n = (h,\ldots,h) \in (\mathbb{R}^m)^n$.

Definición 3.10 Sea una función $f : U \subset \mathbb{R}^m \to \mathbb{R}$, U abierto y f de clase C^n en un entorno $B(a) \subset U$. Entonces se define el polinomio de Taylor de orden n de f en $B(a)$ como

$$P_n^a(x) = f(a) + \sum_{k=1}^{n} \frac{1}{k!}\left[D^k f(a)\right](x-a)^k,$$

y el resto de Taylor, como

$$R_n^a(x) = f(x) - P_n^a(x), \tag{3.1}$$

donde $lim_{n\to\infty}\dfrac{R_n^a(x)}{\|x-a\|^n} = 0$, es decir, que $R_n^a(x)$ tiende a cero mas rápido que $\|x-a\|^n$, $x \in B(a)$, lo que se representa de la siguiente manera

$$R_n^a(x) = o\left(\|x-a\|^n\right),$$

entonces por 3.1 $f(x) = P_n^a(x) + R_n^a(x)$ en $B(a)$. Este resultado se conoce como el teorema de Taylor [19].

Al considerar el polinomio de Taylor de grado 1 de la función f alrededor de x^0,

$$\begin{aligned} P_n^{x^0}(x) &= f\left(x^0\right) + \left[D^1 f(x^0)\right]\left(x - x^0\right) \\ &= f\left(x^0\right) + \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}\left(x^0\right)\left(x_i - x_i^0\right), \end{aligned}$$

por el teorema de Taylor,

$$f(x) = f\left(x^0\right) + \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}\left(x^0\right)\left(x_i - x_i^0\right) + R_1^{x^0}(x),$$

donde $R_1^{x^0}(x) = \frac{1}{2}\left[D^2 f(\rho)\right](x-x^0)^2 = o\left(\|x-x^0\|\right)$, $\rho \in L\left[x, x^0\right] - \{x, x^0\}$, siendo $L\left[x, x^0\right]$ el segmento de extremos x y x^0. Al evaluar $x^0 + \lambda \nabla f\left(x^0\right) \in B(x^0)$ en la última expresión de f, se obtiene que

$$\begin{aligned} f\left(x^0 + \lambda \nabla f\left(x^0\right)\right) &= f\left(x^0\right) + \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}\left(x^0\right)\left(x_i^0 + \lambda \frac{\partial f}{\partial x_i}\left(x^0\right) - x_i^0\right) + o\left(\left\|x^0 + \lambda \nabla f\left(x^0\right) - x^0\right\|\right) \\ &= f\left(x^0\right) + \lambda \left\langle \nabla f\left(x^0\right), \nabla f\left(x^0\right)\right\rangle + o\left(\lambda \left\|\nabla f\left(x^0\right)\right\|\right) \\ &= f\left(x^0\right) + \lambda \left\|\nabla f\left(x^0\right)\right\|^2 + o\left(\lambda \left\|\nabla f\left(x^0\right)\right\|\right). \end{aligned}$$

Por lo tanto, si $\nabla f\left(x^0\right) \neq 0$, para $\lambda > 0$ se tiene que

$$f\left(x^0 + \lambda \nabla f\left(x^0\right)\right) > f\left(x^0\right),$$

lo que quiere decir que $x^0 + \lambda \nabla f\left(x^0\right)$ es una mejora sobre x^0 en la búsqueda del punto donde la función f alcanza un máximo. Para formular un algoritmo iterativo que implemente esta idea, se supone que se tiene el punto x^k. Para hallar el siguiente punto x^{k+1}, se parte de x^k y se suma $\lambda_k \nabla f\left(x^k\right)$, donde λ_k es un escalar positivo, conocido como el tamaño del paso, es decir, se sigue el siguiente algoritmo iterativo,

$$x^{k+1} = x^k + \lambda_k \nabla f\left(x^k\right).$$

Este algoritmo es conocido como el ascenso del gradiente. El gradiente varía en cada iteración, tendiendo a cero mientras se aproxima al valor del problema de maximizar la función, y se detiene cuando el gradiente es cero, que es cuando se halla el punto donde f alcanza un máximo. En la práctica, el algoritmo se detiene en algún k tal que $\left\|\nabla f\left(x^k\right)\right\| < \epsilon$, para un $\epsilon \in \mathbb{R}^+$ lo suficientemente pequeño, debido a que en las computadoras frecuentemente no se obtiene el cero con una precisión del cien por ciento. El valor de α_k se puede seleccionar de diferentes maneras, dando lugar a distintos métodos que resuelven el problema. Cabe destacar que en una función que tiene varios máximos locales, mediante esta metodología se puede hallar un máximo local y no el máximo absoluto, dependiendo del punto inicial x^k que se tome, pero si la función es cuadrática convexa y está definida en un conjunto convexo, la función no tiene máximos locales distintos al global, en cuyo caso la solución obtenida por la metodología del gradiente es el punto donde es alcanzado el máximo absoluto.

3.3. Máquinas de Aprendizaje

Las máquinas de aprendizaje tienen como objetivo desarrollar técnicas que permitan a las computadoras aprender a partir de una información no estructurada suministrada. La metodología consiste en aprender una función objetivo dada, es decir, lograr que la salida de la máquina aproxime el valor de la función bajo algún criterio, para toda entrada que se le introduzca. Para ello a la máquina se le presenta un conjunto de entrenamiento formado por pares de entrada-salida deseada y se activa un proceso adaptativo de los parámetros que caracterizan una función hipótesis. La finalidad es inducir de los datos una hipótesis que logre un desempeño aceptable de la máquina en la reproducción de las salidas, dadas las entradas. Al entrenamiento descrito se le conoce como supervisado, es decir, utilizan las entradas y las salidas de los datos de entrenamiento. Existe otro tipo de aprendizaje que no se considera en este trabajo, conocido como no supervisado en el que se presentan solo los patrones de entrada, los cuales son agrupados por la máquina a partir de alguna medida de distancia, de manera que entradas similares arrojan una misma respuesta [14].

Típicamente, las entradas están formadas por vectores m-dimensionales de atributos de tal manera que el espacio de entrada es un subconjunto de $\mathbb{R}^m$. La salida deseada de cada vector de entrada es un número real, y corresponde al valor de la función objetivo. Las máquinas de aprendizaje se caracterizan por el tipo de funciones hipótesis que utilizan. Las funciones lineales, características de las máquinas de aprendizaje lineales figuran entre las más comprendidas y fáciles de aplicar. Mediante estas técnicas se pueden resolver problemas de predicción, clasificación, estimación y regresión. Nosotros nos concentraremos en las máquinas de aprendizaje lineales para resolver problemas de clasificación.

3.4. Clasificadores Lineales

El problema de clasificación consiste en asignar a un conjunto de vectores pertenecientes a un espacio de entrada $X \subseteq \mathbb{R}^m$, una clase de un conjunto de clases preestablecidas. En la clasificación multiclases (n clases) el espacio de salida, denotado por Y, es el conjunto

$\{1, 2, ..., n\}$ mientras que en la clasificación binaria, $Y = \{-1, 1\}$. Consideramos este último caso, se denota $S = ((x^1, y^1), ..., (x^p, y^p))$ al conjunto de entrenamiento, donde cada elemento o ejemplo $(x^i, y^i) \in X \times Y$, está formado por un vector de entrada y la clase a la cual pertenece, definida por 1 o -1 según sea el caso. Para resolver este problema mediante una máquina de aprendizaje lineal a menudo se utiliza una función lineal definida en X a valores reales de la siguiente forma [5, 14]: si $\in X$, entonces,

$$\begin{aligned} f(x) &= \langle w, x \rangle + b \\ &= \sum_{i=1}^{m} w_i x_i + b, \end{aligned} \tag{3.1}$$

donde w en X y b en $\mathbb{R}$ son conocidos como el vector de pesos y el valor umbral respectivamente y son los parámetros que controlan la función, mientras que la regla de decisión viene dada por signo$(f(x))$, si $f(x) \geq 0$ se le asigna la clase 1 y en caso contrario, la clase -1.

Las máquinas de aprendizaje lineales que utilizan este tipo de hipótesis son entrenadas para hallar los valores de w y b que definen un hiperplano de ecuación $\langle w, x \rangle + b = 0$ que particiona al espacio de entrada en dos semiespacios, donde deben encontrarse en cada uno entradas de una misma clase. A continuación se enuncia y demuestra un lema y un teorema que ayuda a describir el significado de los parámetros que conforman el hiperplano.

Lema 3.1 Sea w un vector de $\mathbb{R}^m$ y $P_w(x)$ la proyección ortogonal de x sobre w. Entonces

$$P_w(x) = \frac{\langle w, x \rangle}{\| w \|^2} w.$$

Demostración.

Por definición, la proyección de x sobre w es de la forma $P_w(x) = cw$, para algún $c \in \mathbb{R}$, y satisface que $\langle x - P_w(x), P_w(x) \rangle = 0$ (ver figura 3.1). Sustituyendo la primera relación en la segunda se obtiene

$$\langle x - cw, cw \rangle = 0,$$

y por propiedades de producto interno,

$$c\langle x, w \rangle - c^2 \langle w, w \rangle = 0,$$

por lo tanto,

$$c = \frac{\langle w, x \rangle}{\| w \|^2},$$

de donde se obtiene que

$$P_w(x) = \frac{\langle w, x \rangle}{\| w \|^2} w.$$

□

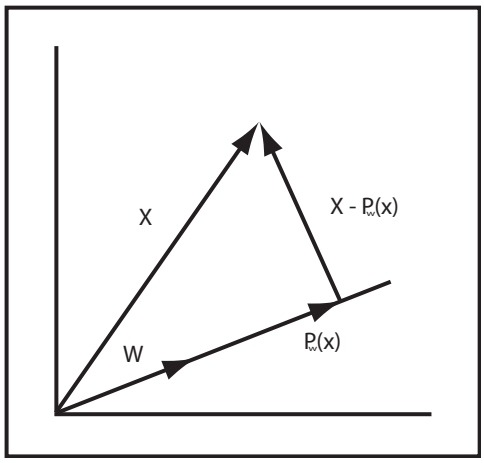


Figura 3.1: Proyección Ortogonal de x en w.

Teorema 3.4 Sea $w \in \mathbb{R}^m$ un vector no nulo, $b \in \mathbb{R}$ y $H(w, b) = \{x \in \mathbb{R}^m / \langle w, x \rangle + b = 0\}$. Entonces $x \in H(w, b)$ si, y sólo si $P_w(x) = \frac{-b}{\| w \|^2} w$.

Demostración.

($\Rightarrow$) Sea $x \in H(w, b)$, entonces

$$\langle w, x \rangle + b = 0,$$

por lo tanto,

$$\frac{\langle w, x \rangle}{\| w \|^2} w = \frac{-b}{\| w \|^2} w,$$

y por el lema 3.1

$$P_w(x) = \frac{-b}{\| w \|^2} w.$$

($\Leftarrow$) Sea $x \in \mathbb{R}^m$, por hipótesis $P_w(x) = \frac{-b}{\| w \|^2} w$, y por el lema 3.1

$$\frac{\langle w, x \rangle}{\| w \|^2} w = \frac{-b}{\| w \|^2} w,$$

entonces,

$$(\langle w, x\rangle + b)\frac{w}{\| w \|^2} = 0,$$

por ser w un vector no nulo,

$$\langle w, x\rangle + b = 0,$$

y por definición,

$$x \in H(w, b).$$

□

Del teorema anterior se deduce que el vector w es ortogonal al hiperplano H, ya que la proyección de todos los $x \in H$ en w es la misma $\left(\frac{-b}{\|w\|^2}w\right)$. Además, es de notar que la distancia del hiperplano al origen de coordenadas es $\frac{|b|}{\|w\|}$, de manera que la variación del valor umbral b traslada al hiperplano en el espacio (ver figura 3.2). Por otro lado, cabe destacar que no todos los problemas contienen datos que se puedan separar por un hiperplano, en el caso de que exista, se habla de un problema linealmente separable y en caso contrario se dice que no es separable.

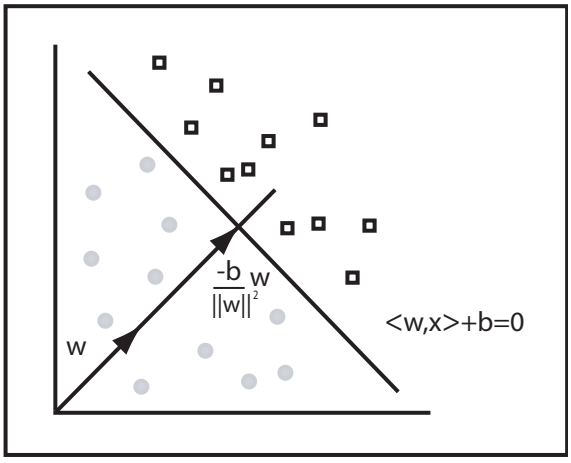


Figura 3.2: Hiperplano Separador.

3.5. Espacios de Rasgos y Funciones Kernel

Una de las desventajas de las máquinas de aprendizaje lineales es su condición de linealidad. Es bien conocido que la manera de representar muchos problemas de la vida real no es un modelo lineal. Como solución alternativa a este problema se presentan las metodologías de preprocesamiento en las cuales se mudan los datos mediante una transformación no lineal ϕ a un espacio de Hilbert $\mathcal{H}$ de mayor dimensión conocido como espacio de rasgos, y donde los datos transformados resultan separables linealmente [15]. El espacio $\mathcal{H} = \{\phi(x)/x \in X)$, es de Hilbert y se conoce como espacio de rasgos.

El objetivo que se persigue es hacer que información relevante se haga explícita y disponible para ser usada por la máquina de aprendizaje. La complejidad de la función objetivo a ser aprendida depende en gran medida de la forma en que está representada, y en consecuencia puede alterar el grado de dificultad de la tarea de aprendizaje. De manera que, posiblemente una transformación sencilla de los datos en otro espacio puede simplificar significativamente el aprendizaje.

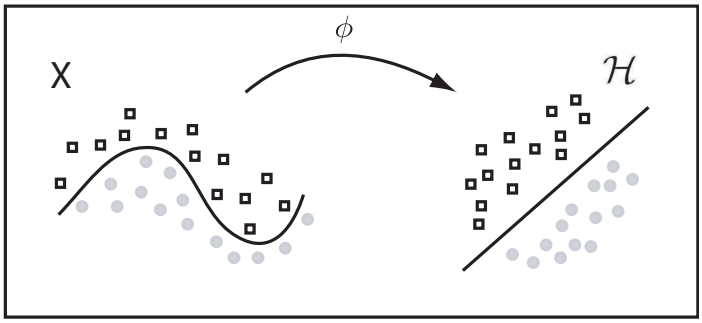


Figura 3.3: Transformación de los datos de entrada a un Espacio de Rasgos

Para poder aplicar las máquinas lineales en el espacio de entrada la condición de separabilidad lineal de los datos es fundamental, la cual no tiene que cumplirse siempre, es más, es difícil de conseguir en un problema real. Si los datos no cumplen con esta condición basta con aplicar una función a valores en un espacio de rasgos donde la imagen sea linealmente separable y poder aplicar la máquina lineal.

Como se vio en la sección 3.4, las máquinas de aprendizaje lineales utilizan unas funciones hipótesis lineales $f : X \in \mathbb{R}^m \rightarrow Y \in \mathbb{R}$ que se pueden escribir como:

$$f(x) = \langle w, x \rangle + b.$$

Ahora bien, a fin de aprender relaciones no lineales con máquinas lineales es necesario seleccionar un conjunto de características no lineales con las cuales poder reescribir la información original en una nueva representación. De ahí que el conjunto de hipótesis que se consideran son del tipo:

$$f(x) = \langle \phi(w), \phi(x) \rangle + b,$$

donde $\phi : X \rightarrow \mathcal{H}$ es una función no lineal que va del espacio de entradas a algún espacio de rasgos $\mathcal{H}$ conocida como función de rasgos. De manera que la implementación de una máquina no lineal consiste entonces de dos pasos:

- Una función de rasgos que transforma los datos a un espacio de rasgos.
- Una máquina lineal que se emplea sobre los datos transformados.

Mediante un tipo de función conocida como kernel se unifican los dos pasos involucrados en la construcción de una máquina no lineal, donde ésta calcula directamente el producto interno entre los vectores de rasgos. A continuación se define esta función y se muestran sus propiedades en el teorema de Mercer.

Una función $K : X \times X \rightarrow \mathbb{R}$, con X un conjunto compacto de dimensión m, es un kernel si satisface las siguientes propiedades:

1. $K(x, w)$ es continua.

2. $K(x, w) = K(w, x)$, es simétrica.

3. $K(x, w)$ es semidefinida positiva, es decir,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} K\left(x^i, x^j\right) c_i c_j \geq 0,$$

para cualquier secuencia $\{x^i\}_{i=1}^n$ de X y cualquier combinación de n números reales $c_1, \ldots, c_n$.

Teorema 3.5 (Mercer) Si una función K es un kernel, entonces existe un espacio de Hilbert $\mathcal{H}$ de funciones reales definidas en X y una función de rasgos $\phi : X \to \mathcal{H}$ tal que,

$$\langle \phi(x), \phi(w) \rangle = K(x, w),$$

donde $x, w \in X$ y $\langle\ ,\ \rangle$ es un producto interno en $\mathcal{H}$.

Por el teorema de Mercer, mediante una función kernel se pueden mudar los datos de entrada al espacio de rasgos de manera implícita y entrenar a la máquina en tal espacio, es decir, no es necesario representar los vectores de rasgos explícitamente, así el numero de operaciones requeridas para calcular el producto interno no es necesariamente proporcional al número de rasgos, lo cual es una gran ventaja desde el punto de vista computacional.

Cualquier máquina de aprendizaje definida en un espacio de entrada X, se puede aplicar en un espacio de rasgos simplemente al sustituir en la función hipótesis y en el algoritmo de aprendizaje que la definen, los productos internos por una función kernel. Esta metodología se conoce como el truco del kernel. Existen diversos kernel, entre los más utilizados se encuentran los polinómicos y gaussianos que son kernels no lineales.

Kernels Polinómicos: Consideremos el kernel $K(x^1, x^2) = \langle x^1, x^2 \rangle^2$, veamos que forma tienen los rasgos de la correspondencia:

$$\langle x^1, x^2 \rangle^2 = \left(\sum_{i=1}^{m} x_i^1 x_i^2 \right) \left(\sum_{j=1}^{m} x_j^1 x_j^2 \right) = \sum_{i=1}^{m} \sum_{j=1}^{m} x_i^1 x_j^1 x_i^2 x_j^2 = \sum_{(i,j)=1}^{(m,m)} \left(x_i^1 x_j^1 \right) \left(x_i^2 x_j^2 \right),$$

que es equivalente a un producto interno entre los vectores de rasgos a través de la correspondencia

$$\phi(x) = (x_i x_j)_{(i,j)}^{(m,m)}, \tag{3.1}$$

en donde hay $\binom{m+2}{2}$ rasgos correspondientes a todos los monomios de segundo grado considerados en (3.1), aunque es de notar que el rasgo $x_i x_j$ ocurre dos veces, otorgándole el doble de peso que los rasgos x_i^2. Espacios de rasgos más generales se obtienen al considerar los siguientes Kernel:

$$K(x^1, x^2) = \langle x^1, x^2 \rangle^d \quad y \quad K(x^1, x^2) = \langle x^1, x^2 + c \rangle^d,$$

con $d \geq 2$ y $c \in \mathbb{R}$. Los $\binom{m+d-1}{d}$ rasgos diferentes del primer kernel son todos los monomios de grado d, con un peso determinado por el exponente y en el segundo hay $\binom{m+d}{d}$ rasgos distintos siendo todos monomios de hasta grado d.

Kernels Gaussianos: Estos están definidos de la siguiente forma:

$$K\left(x^1, x^2\right) = \exp\left(\frac{-\left\|x^1 - x^2\right\|^2}{\sigma^2}\right),$$

donde $\sigma \in \mathbb{R}^+$, conocida también como función de base radial, ésta genera un espacio de rasgos de dimensión infinita.

3.6. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial son máquinas de aprendizaje lineales desarrolladas inicialmente en 1992 por Vapnik, basadas en la teoría de aprendizaje estadístico. Estas pertenecen a la familia de los clasificadores lineales puesto que inducen separadores lineales o hiperplanos como los descritos en la sección 3.4 en espacios de rasgos de muy alta dimensionalidad obtenidos por funciones núcleo o kernel [9].

El hiperplano que buscan estas máquinas es el que equidista de los puntos mas cercanos de cada clase; conocido como el hiperplano de máximo margen. La búsqueda del mismo se plantea como un problema de optimización cuadrático y convexo que se expresa en su forma dual a través de multiplicadores de Lagrange y se resuelve mediante diversos algoritmos numéricos. A continuación definiremos el concepto de margen, el cual juega un papel fundamental en las máquinas de soporte vectorial.

3.6.1. Margen Funcional y Geométrico

Definición 3.11 El margen funcional de un punto de entrenamiento $(x^i, y^i) \in \mathbb{R}^m \times \{-1, 1\}$ con respecto a un hiperplano de ecuación $\langle w, x\rangle + b = 0$ denotado por $((w, b))$ se define como:

$$\gamma_i = y^i \left(\left\langle w, x^i \right\rangle + b\right).$$

Se puede observar que $\gamma_i > 0$ implica clasificación correcta de los datos (x^i, y^i). Además esta propiedad es invariante ante un cambio de escala de los parámetros de la máquina.

Definición 3.12 La distribución de márgenes funcionales de un hiperplano (w, b) con respecto a un conjunto de entrenamiento S es la distribución de los márgenes funcionales de los datos en S.

Definición 3.13 El margen funcional de un hiperplano (w, b) es el mínimo de los márgenes funcionales del hiperplano con respecto a un conjunto de entrenamiento S.

Una definición equivalente al margen funcional es el margen geométrico, cuya diferencia radica en que los parámetros w y b se multiplican por $\frac{1}{\|w\|}$ como se muestra a continuación.

Definición 3.14 El margen geométrico de un punto de entrenamiento bien clasificado $(x^i, y^i) \in \mathbb{R}^m \times \{-1, 1\}$ con respecto a un hiperplano (w, b) se define

$$\gamma_{gi} = y^i \left(\left\langle \frac{w}{\| w \|}, x^i \right\rangle + \frac{b}{\| w \|} \right),$$

γ_i es siempre positivo ya que el punto es de entrenamiento y representa la distancia euclídea de los puntos en el espacio de entrada al hiperplano separador.

Definición 3.15 La distribución de márgenes geométricos de un hiperplano (w, b) con respecto a un conjunto de entrenamiento S es la distribución de los márgenes geométricos de los datos en S.

Definición 3.16 El margen geométrico de un hiperplano $(w.b)$ es el mínimo margen de los márgenes geométricos del hiperplano con respecto a un conjunto de entrenamiento S.

Definición 3.17 El margen de un conjunto de entrenamiento S es el máximo margen geométrico sobre todos los hiperplanos de separación posibles. El hiperplano al cual corresponde este margen se le conoce como hiperplano de máximo margen.

En la figura 3.4 se muestran los márgenes geométricos de dos puntos con respecto a un hiperplano, así como también al margen de un conjunto de puntos junto con el hiperplano de máximo margen.

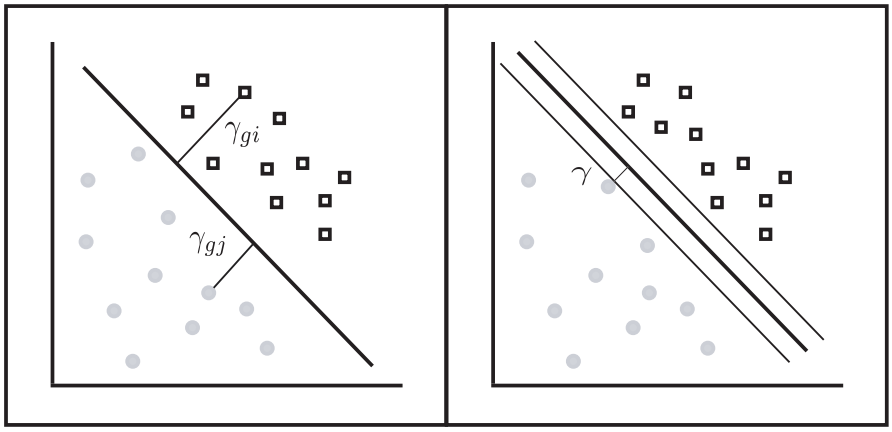


Figura 3.4: (a) Márgenes Geométricos. (b)Hiperplano de Máximo Margen.

3.6.2. Clasificador de Máximo Margen

La máquina de soporte vectorial (MSV) lineal con margen máximo (clasificador de máximo margen) es el modelo más sencillo e intuitivo de las MSVs, aunque también el que tiene condiciones de aplicabilidad mas restringidas, pues parte de la hipótesis de que el conjunto de datos es linealmente separable en el espacio de entrada. Este modelo es un clasificador lineal basado en las ideas descritas en la sección 3.4 donde se busca un hiperplano separador (w, b).

Dado un conjunto binario linealmente separable existen varios métodos que contienen diversos algoritmos para construir hiperplanos que los clasifiquen correctamente. A pesar de que esté garantizada la convergencia de todos ellos hacia un hiperplano solución, las particularidades de cada algoritmo de aprendizaje puede conducirnos a soluciones ligeramente distintas, puesto que pueden haber varios (de hecho infinitos) hiperplanos que separen el conjunto de entrenamiento. El clasificador de máximo margen consiste en seleccionar el hiperplano separador que está a la misma distancia de los patrones de entrada del conjunto de entrenamiento más cercanos de cada clase. Equivalentemente, es el hiperplano que maximiza el margen geométrico sobre todos los hiperplanos separadores posibles. De manera que el hiperplano de máximo margen es aquel que satisface

$$\max_{w} \left[\min_{(x^i, y^i)} y^i \left(\left\langle \frac{w}{\| w \|}, x^i \right\rangle + \frac{b}{\| w \|} \right) \right].$$

Al multiplicar a w y a b por una constante, el hiperplano (w, b) y el margen geométrico que determina distancia no varían. Por ello podemos reescalar w y b de manera que la distancia de

los patrones de entrada mas cercanos al hiperplano sea $\frac{1}{\|w\|}$, es decir,

$$\min_{(x^i,y^i)} \frac{y^i\left(\langle w, x^i\rangle + b\right)}{\parallel w \parallel} = \frac{1}{\parallel w \parallel}.$$

Como consecuencia, el margen funcional $y^i\left(\langle w, x^i\rangle + b\right)$ de los patrones mas cercanos es igual a 1, mientras que para el resto de los patrones es mayor que uno. De manera que el problema de encontrar al hiperplano de máximo margen se reduce a resolver el siguiente problema de optimización con restricciones de desigualdad [9]:

$$\begin{array}{lll} Maximizar & \dfrac{1}{\parallel w \parallel}, & w \in \mathbb{R}^m \\ sujeto\ a & y^i\left(\left\langle w, x^i\right\rangle + b\right) \geq 1, & i = 1, \ldots, p. \end{array}$$

Otra formulación de la MSV con margen máximo, más habitual, equivalente a la anterior, es la siguiente:

$$\begin{array}{lll} Minimizar & \dfrac{1}{2}\langle w, w\rangle, & w \in \mathbb{R}^m \\ sujeto\ a & y^i\left(\left\langle w, x^i\right\rangle + b\right) \geq 1, & i = 1, \ldots, p. \end{array} \tag{3.1}$$

Este es un problema de optimización cuadrático convexo, es decir, se trata de minimizar una función cuadrática y convexa definida en un conjunto convexo y con una restricción lineal. Este problema admite una formulación dual cuya solución es la misma que la del problema original.

3.6.3. Formulación Dual del Clasificador de Máximo Margen

La formulación dual, equivalente a la forma original 3.1 expresa la solución del problema como una combinación lineal de los patrones de entrenamiento. Su desarrollo se basa en la teoría de optimización descrita en la sección 3.1. Para resolver este problema de optimización cuadrática, de acuerdo al teorema 3.1 se tiene que encontrar el punto de silla de la función Lagrangeana

$$L(w, b, \alpha) = \frac{1}{2}\langle w, w\rangle + \sum_{i=1}^{p} \alpha_i \left(1 - y^i\left(\left\langle w, x^i\right\rangle + b\right)\right), \tag{3.2}$$

donde $\alpha_i \geq 0$ son los multiplicadores de Lagrange. Para ello se debe minimizar a L con respecto a w y b; y maximizarla sobre los $\alpha_i \geq 0$.

Según el teorema de Fermat [21] un punto donde L alcanza un mínimo satisface lo siguiente:

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \Rightarrow w - \sum_{i=1}^{p} y^i \alpha_i x^i = 0 \Rightarrow w = \sum_{i=1}^{p} y^i \alpha_i x^i. \tag{3.3}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \Rightarrow \sum_{i=1}^{p} y^i \alpha_i = 0. \tag{3.4}$$

Sustituyendo 3.3 en la función Lagrangeana 3.2 y considerando 3.4, se obtiene

$$\begin{aligned} W(\alpha) &= \frac{1}{2}\left\langle \sum_{i=1}^{p} y^i \alpha_i x^i, \sum_{j=1}^{p} y^j \alpha_j x^j \right\rangle + \sum_{i=1}^{p} \alpha_i \left(1 - y^i \left(\left\langle \sum_{j=1}^{p} y^j \alpha_j x^j, x^i \right\rangle + b\right)\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j \left\langle x^i, x^j \right\rangle + \sum_{i=1}^{p} \alpha_i - \sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j \left\langle x^i, x^j \right\rangle - \sum_{i=1}^{p} y^i \alpha_i b \\ &= \sum_{i=1}^{p} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j \left\langle x^i, x^j \right\rangle. \end{aligned}$$

La notación de $L(w,b,\alpha)$ se cambió a $W(\alpha)$ para denotar la ultima transformación; además la función obtenida es conocida como Lagrangeana Dual. Para obtener el hiperplano de máximo margen se tienen que calcular los multiplicadores de lagrange $\alpha_i \geq 0$ que maximizan la lagrangeana dual. Estos resultados nos dan una solución analítica y exacta del problema, los cuales derivan en el siguiente teorema.

Teorema 3.6 (Clasificador de Máximo Margen) Se supone que el conjunto de entrenamiento es linealmente separable en el espacio de entrada. Sea α^* una solución al problema dual

$$\begin{aligned} &Maximizar && \sum_{i=1}^{p} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j \left\langle x^i, x^j \right\rangle, \\ &sujeto\ a && \sum_{i=1}^{p} y^i \alpha i = 0, \\ & && \alpha_i \geq 0, \qquad i = 1, \ldots, p. \end{aligned} \tag{3.5}$$

Entonces, el vector

$$w^* = \sum_{i=1}^{p} y^i \alpha_i^* x^i,$$

es el vector ortogonal al hiperplano de máximo margen. La MSV lineal con margen máximo es

$$f(x) = \langle w^*, x \rangle + b^* = \sum_{i=1}^{p} y^i \alpha_i^* \left\langle x^i, x \right\rangle + b^*,$$

donde

$$b^* = \frac{1}{2}\left(\underset{y^j=-1}{\text{máx}} \left\langle w^*, x^j \right\rangle + \underset{y^j=1}{\text{mín}} \left\langle w^*, x^j \right\rangle\right),$$

y su clasificador asociado es $h(x) = signo(f(x))$.

Demostración.

La sección anterior contiene la prueba de este teorema salvo el calculo de b que se demuestra a continuación. Como el hiperplano de máximo margen es aquel que equidista de los patrones mas cercanos de cada una de las clases, se consideran dos de los patrones que están mas cerca del hiperplano en clases diferentes (x^1 *y* x^2) (ver figura 3.5), su proyección sobre w^* según el lema 3.1 es $\frac{\langle w^*, x^1\rangle}{\|w^*\|^2} w^*$ y $\frac{\langle w^*, x^2\rangle}{\|w^*\|^2} w^*$, y como la distancia de cada uno de ellos al hiperplano es la misma y la proyección de cualquier patrón que está en el hiperplano sobre w^* es $\frac{-b^*}{\|w^*\|^2} w^*$, se tiene que

$$\frac{\langle w^*, x^1\rangle}{\|w^*\|^2} w^* - \left(\frac{-b^*}{\|w^*\|^2}\right) = \frac{-b^*}{\|w^*\|^2} - \frac{\langle w^*, x^2\rangle}{\|w^*\|^2} w^*,$$

de donde

$$\left\langle w^*, x^1 \right\rangle + b^* = -b^* - \left\langle w^*, x^2 \right\rangle,$$

por tanto

$$b^* = -\frac{1}{2}\left(\left\langle w^*, x^1 \right\rangle + \left\langle w^*, x^2 \right\rangle\right),$$

y considerando que x^1 y x^2 son los patrones de cada clase que están mas cerca del hiperplano tales que la distancia de entre un patrón y el hiperplano está determinada por el producto escalar de w^* y el patrón (ver figura 3.5), se concluye que

$$b^* = \frac{1}{2}\left(\underset{y^j=-1}{\text{máx}} \left\langle w^*, x^j \right\rangle + \underset{y^j=1}{\text{mín}} \left\langle w^*, x^j \right\rangle\right).$$

□

Una de las consecuencias más importantes del teorema es la relación

$$w^* = \sum_{i=1}^{p} y^i \alpha_i^* x^i,$$

indica que el vector ortogonal al hiperplano de máximo margen se puede expresar como una combinación lineal de los p ejemplos del conjunto de entrenamiento. Por otro lado, las soluciones del problema original y dual del máximo margen satisfacen las condiciones de Kuhn-Tucker (ver teorema 3.1), entonces

$$\alpha_i^* \left(1 - y^i \left(\left\langle w^*, x^i \right\rangle + b^*\right)\right) = 0, \qquad i = 1, \ldots, p.$$

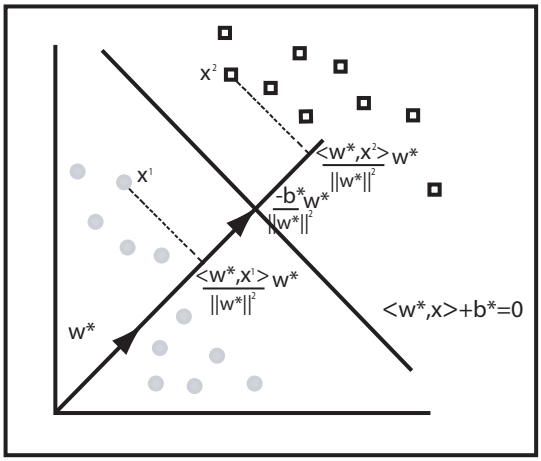


Figura 3.5: Hiperplano Separador.

Para los α_i^* distintos de cero, podemos observar de la relación anterior que el margen funcional de los vectores x^i es igual a uno. Geométricamente, dichos vectores son los mas cercanos al hiperplano de máximo margen y son conocidos como vectores soporte. Además, estos son los vectores que intervienen en la combinación lineal que define a w^*, el resto no contribuye en nada ya que sus multiplicadores de Lagrange son cero. La mayoría de las variables duales son cero, de manera que la cantidad de vectores soporte es mucho menor que el número de patrones de entrenamiento. Esta propiedad es muy importante ya que controla en alguna medida la capacidad de generalización de la máquina, mientras el conjunto de vectores soporte es más pequeño la capacidad de generalización es mayor.

3.6.4. Clasificador de Máximo Margen en El Espacio de Rasgos

La máquina de soporte vectorial lineal con margen máximo, como máquina de aprendizaje lineal se puede aplicar en un espacio de rasgos inducido por una función kernel, como se describe en la sección 3.5. Esto se logra simplemente sustituyendo en el modelo definido para el espacio de entrada, el producto interno por la función kernel. Este procedimiento es conocido comúnmente como el truco del kernel. Se tiene el siguiente resultado, bien importante ya que plantea el problema del clasificador utilizado en este trabajo.

Teorema 3.7 (Clasificador de Máximo Margen en el Espacio de Rasgos) Se supone que el conjunto de entrenamiento es linealmente separable en el espacio de rasgos definido implícitamente por una función núcleo $K(x, y)$. Sea α^* una solución al problema dual

$$\begin{aligned} Maximizar \qquad & \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j K\left(x^i, x^j\right), \\ sujeto\ a \qquad & \sum_{i=1}^{p} y^i \alpha i = 0, \\ & \alpha_i \geq 0, \qquad\qquad i = 1, \ldots, p. \end{aligned} \tag{3.6}$$

Entonces, el vector

$$w^* = \sum_{i=1}^{p} y^i \alpha_i^* \phi\left(x^i\right),$$

es el vector ortogonal al hiperplano de máximo margen en el espacio de rasgos. La MSV lineal con margen máximo en el espacio de rasgos es

$$f(x) = \langle w^*, \phi(x) \rangle + b^* = \sum_{i=1}^{p} y^i \alpha_i^* K\left(x^i, x\right) + b^*,$$

donde

$$b^* = \frac{1}{2} \left(\max_{y^j=-1} \sum_{i=1}^{p} y^i \alpha_i^* K\left(x^i, x^j\right) + \min_{y^j=1} \sum_{i=1}^{p} y^i \alpha_i^* K\left(x^i, x^j\right) \right),$$

y su clasificador asociado es $h(x) = signo(f(x))$.

Demostración.

Este teorema se obtiene se aplicar el truco del kernel al resultado del teorema 3.6 cuya demostración se mostró anteriormente.

□

3.6.5. Algoritmo de Aprendizaje Adatron

La MSV lineal con margen máximo resuelve un problema de optimización cuadrática convexa donde el número de coeficientes es igual al número de patrones de entrenamiento. Este hecho hace que para grandes cantidades de datos las técnicas numéricas de optimización existentes para resolver el problema cuadrático, no sean admisibles en términos computacionales. Es por

esto que continuamente se concentran esfuerzos en diseñar algoritmos para mejorar tanto el tiempo como el costo computacional de los algoritmos convencionales ya existentes para resolver el problema. En este sentido se considera un algoritmo de aprendizaje conocido como el Adatron que resuelve el problema de optimización cuadrática convexa 3.5 que da solución al clasificador de máximo margen, el cual es mucho más rápido, y fácil de implementar que muchos algoritmos ya existentes. El mismo se puede aplicar tanto en el espacio de entrada como se planteó inicialmente, como en el espacio de rasgos inducido por un kernel. Consideremos este último con el cual trabajaremos de ahora en adelante y donde el problema de optimización es el siguiente (forma dual 3.6)

$$\begin{aligned} Maximizar \qquad & \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j K\left(x^i, x^j\right), \\ sujeto\ a \qquad & \sum_{i=1}^{p} y^i \alpha_i = 0, \\ & \alpha_i \geq 0, \qquad\qquad i = 1, \ldots, p. \end{aligned}$$

La metodología que utiliza el Adatron para resolver este problema de optimización cuadrático y convexo es a través de la técnica del ascenso del gradiente, descrita en la sección 3.2, que consiste en la siguiente dinámica de adaptación de los α_i:

$$\delta\alpha_i = \lambda \frac{\partial W(\alpha)}{\partial \alpha_i},$$

donde λ es una constante que se conoce como tasa de aprendizaje. Considerando la condición $\sum_{i=1}^{p} y^i \alpha_i = 0$, la función Lagrangeana dual se puede escribir como

$$W(\alpha) = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j K\left(x^i, x^j\right) - \sum_{i=1}^{p} y^i \alpha_i b.$$

Por lo tanto, cada componente del gradiente de W se puede expresar de la siguiente manera

$$\begin{aligned} \frac{\partial W(\alpha)}{\partial \alpha_i} &= 1 - y^i \sum_{j=1}^{p} y^j \alpha_j K\left(x^i, x^j\right) - y^i b \\ &= 1 - y^i \left(\sum_{j=1}^{p} y^j \alpha_j K\left(x^i, x_j\right) + b \right). \end{aligned} \tag{3.7}$$

Al recordar que $\gamma_i = y^i \left(\langle w, x^i \rangle + b\right)$ y al expresar a W en su forma dual, se tiene que

$$\gamma_i = y^i \left(\left\langle \sum_{j=1}^{p} y^j \alpha_j x^j, x^i \right\rangle + b \right) = y^i \left(\sum_{j=1}^{p} y^j \alpha_j \left\langle x^i, x^j \right\rangle + b \right),$$

y al considerar el margen funcional en el espacio de rasgos

$$\gamma_i = y^i \left(\sum_{j=1}^{p} y^j \alpha_j K\left(x^i, x^j\right) + b \right).$$

Entonces, al sustituir la ecuación anterior en 3.7, se obtiene que

$$\frac{\partial W(\alpha)}{\partial \alpha_i} = 1 - \gamma_i.$$

De esta manera la dinámica de la adaptación de los α_i toma la forma

$$\delta\alpha_i = \lambda(1 - \gamma_i). \tag{3.8}$$

Para demostrar la convergencia del algoritmo del Adatron se analiza como varía la función Lagrangeana dual

$$W(\alpha) = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{p} y^i y^j \alpha_i \alpha_j K\left(x^i, x^j\right) - \sum_{i=1}^{p} y^i \alpha_i b,$$

en el proceso de aprendizaje secuencial del Adatron. Durante el aprendizaje se presenta el patrón i-ésimo y se corrige el multiplicador α_i, como consecuencia el funcional $W(\alpha)$ varía de la siguiente manera: $\triangle W = \triangle W_1 + \triangle W_2$, considerando la suma de las variaciones del término lineal y de los términos cuadráticos.

La variación del término lineal es

$$\triangle W_1 = \left(\sum_{j=1}^{p} \alpha_j + \delta_{ij}\delta\alpha_i \right) - \sum_{j=1}^{p} \alpha_j = \sum_{j=1}^{p} \alpha_j + \sum_{j=1}^{p} \delta_{ij}\delta\alpha_i - \sum_{j=1}^{p} \alpha_j = \delta\alpha_i,$$

mientras que la variación de los términos cuadráticos se calculan como sigue:

$$\begin{aligned} \triangle W_2 \;=\; & -\frac{1}{2} \sum_{j,k=1}^{p} y^j y^k (\alpha_j + \delta_{ij}\delta\alpha_i)(\alpha_k + \delta_{ik}\delta\alpha_i) K\left(x^j, x^k\right) - b \sum_{j=1}^{p} y^j (\alpha_j + \delta_{ij}\delta\alpha_i) \\ & + \frac{1}{2} \sum_{j,k=1}^{p} y^j y^k \alpha_j \alpha_k K\left(x^j, x^k\right) + b \sum_{j=1}^{p} y^j \alpha_j \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\sum_{j,k=1}^{p} y^j y^k(\alpha_j\alpha_k + \alpha_k\delta_{ij}\delta\alpha_i + \alpha_j\delta_{ik}\delta\alpha_i + \delta_{ij}\delta_{ik}(\delta\alpha_i)^2 K\left(x^j, x^k\right) \\
&\quad +\frac{1}{2}\sum_{j,k=1}^{p} y^j y^k\alpha_j\alpha_k K\left(x^j, x^k\right) - b\sum_{j=1}^{p} y^j\alpha_j + b\sum_{j=1}^{p} y^j\alpha_j - b\sum_{j=1}^{p} y^j\delta_{ij}\delta\alpha_i \\
&= -\frac{1}{2}\sum_{j,k=1}^{p} y^j y^k(\alpha_k\delta_{ij}\delta\alpha_i + \alpha_j\delta_{ik}\delta\alpha_i + \delta_{ij}\delta_{ik}(\delta\alpha_i)^2 K\left(x^j, x^k\right) - b\sum_{j=1}^{p} y^j\delta_{ij}\delta\alpha_i \\
&= -\frac{1}{2}\delta\alpha_i y^i\sum_{k=1}^{p} y^k\alpha_k K\left(x^k, x^i\right) - \frac{1}{2}\delta\alpha_i y^i\sum_{j=1}^{p} y^j\alpha_j K\left(x^j, x^i\right) - \frac{1}{2}(\delta\alpha_i)^2 K\left(x^i, x^i\right) - by^i\delta\alpha_i \\
&= -\delta\alpha_i y^i\left(\sum_{j=1}^{p} y^j\alpha_j K\left(x^j, x^i\right) + b\right) - \frac{1}{2}(\delta\alpha_i)^2 K\left(x^i, x^i\right) \\
&= -\delta\alpha_i\gamma_i - \frac{1}{2}(\delta\alpha_i)^2 K\left(x^i, x^i\right),
\end{aligned}
$$

de tal forma que

$$
\begin{aligned}
\triangle W &= \delta\alpha_i(1-\gamma_i) - \frac{1}{2}(\delta\alpha_i)^2 K\left(x^i, x^i\right) = \delta\alpha_i\left(1 - \gamma_i - \frac{\delta\alpha_i}{2}K\left(x^i, x^i\right)\right) \\
&\geq \delta\alpha_i\left(\frac{\delta\alpha_i}{\lambda} - \frac{\delta\alpha_i}{2}K\left(x^i, x^i\right)\right) = (\delta\alpha_i)^2\left(\frac{1}{\lambda} - \frac{K\left(x^i, x^i\right)}{2}\right) > 0,
\end{aligned}
$$

para $0 < \lambda < \dfrac{2}{K\left(x^i, x^i\right)}$. Para este último calculo se utiliza que $1 - \gamma_i \geq \dfrac{\delta\alpha_i}{\lambda}$, desigualdad obtenida de la regla de corrección $\delta\alpha_i \leq \lambda(1-\gamma_i)$. Esto demuestra que el funcional W aumenta monotonamente en el proceso de aprendizaje del Adatron cuando la constante de aprendizaje está en el rango indicado. Ahora bien, el algoritmo del Adatron, que resuelve el problema del máximo margen se puede resumir en los siguientes pasos, donde se utiliza los kernel mas utilizados descritos en la sección 3.5 [7]:

1.- Inicializar $b = 0$ y $\alpha_i = 0$ para $i = 1, \ldots, p$.

2.- Para cada elemento de entrenamiento (x^i, y^i) realizar lo siguiente:

a.- Calcular $z^i = \sum_{j=1}^{p} y^j\alpha_j K\left(x^j, x^i\right)$.

b.- Calcular el margen funcional $\gamma_i = y^i(z^i + b)$.

c.- Sea $\delta\alpha_i = \lambda(1 - \gamma_i)$, entonces

$$\alpha_i = \begin{cases} 0 & \text{si } \alpha_i + \delta\alpha_i \leq 0 \\ \alpha_i + \delta\alpha_i & \text{en caso contrario} \end{cases}$$

d.- Calcular el umbral

$$b = -\frac{1}{2}\left(\max_{y^j=-1} z^j + \min_{y^j=1} z^j\right)$$

3.- Repetir el paso 2 hasta obtener un número máximo de iteraciones o que el margen funcional del hiperplano (w, b) se aproxime a uno.

Cabe destacar que el algoritmo del Adatron entrena una máquina lineal para resolver un problema de clasificación binaria en donde se halla la solución del problema dual de optimización planteado. Esta solución se utiliza para expresar la máquina de aprendizaje y clasificar cada entrada que se le presente. (ver teorema 3.6). En el caso de la clasificación multiclases (n clases) se deben entrenar $n - 1$ máquinas, en donde cada entrenamiento se realiza para discriminar una clase de las demás.

El algoritmo del Adatron se ha aplicado con éxito en los últimos años a diversos problemas reales pertenecientes a áreas como el reconocimiento y clasificación de imágenes, jerarquización de métodos de recobro de crudo, entre otros. A continuación se enuncian algunos de ellos.

1.- Frieb y colaboradores [5] lo usaron para determinar si un tumor cancerígeno dado es maligno o benigno. Los mismos también realizaron un estudio para distinguir dígitos que identifican caracteres de letras.

2. Rafael Ayestaran y Ignacio Santamaría [1] lo aplicaron en sistemas de comunicaciones digitales.

3. Bertrand Le Saux y Giuseppe Amato [11] clasificaron imágenes según objetos presentes en las mismas mediante esta metodología.

4. Cristina García y José Ali moreno [8] lo utilizaron para segmentar estructuras en imágenes de resonancia magnética.

Capítulo 4

Metodología

En este capítulo se describe la metodología utilizada para segmentar las imágenes de resonancia magnética, desde la recolección y preprocesamiento de los datos, análisis de patrones de comportamientos de los distintos tejidos considerados, caracterizados por un vector aleatorio bivariado que resume la información de las imágenes y la construcción de las máquinas de soporte vectorial.

4.1. Recolección y Preprocesamiento de los Datos

Los datos disponibles para la realización de este trabajo corresponden a IRMs cerebrales multieco potenciadas en T2, provenientes de varios cortes transversales en ocho pacientes que presentan tumores cerebrales. A cada corte de cada paciente corresponde un conjunto de ocho imágenes, las cuales están compuestas por valores distintos de T2.

4.1.1. Formato de los Datos

A cada paciente, se le asigna una matriz de datos por cada corte que se denota por $\mathbf{F}$ y tiene dimensión $n \times 8$, n es el número de pixeles que posee cada imágen del corte,

$$\mathbf{F} = (f_{i,j})_{i=1:n, j=1:8}.$$

Sus filas corresponden a los valores de intensidad de un mismo pixel en las 8 imágenes que conforman el corte, es decir, los elementos de la i-ésima fila son los tiempos de relajación

transversal del i-ésimo pixel

$$(f_{i,j})_{j=1:8}.$$

Las señales de resonancia magnética que se recogen en las ocho imágenes de un corte transversal de un paciente van disminuyendo de una imagen a otra, es decir, que para cada pixel i, $i = 1 : n$, la función $F_i = f_{i,\cdot} : \{1, 2, ..., 8\} \to \mathbb{R}$ es decreciente y su gráfico lo identificamos con una curva de relajación. Además, el decaimiento es exponencial, como se puede ver en la figura 4.16. Esto se debe a que durante el proceso de relajación la señal está asociada con la liberación de energía que va disminuyendo hasta que los núcleos de hidrógeno regresan en su posición original.

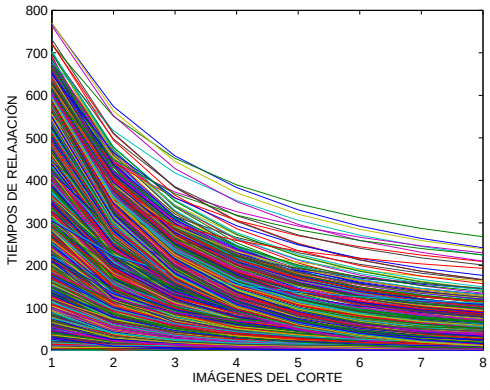


Figura 4.1: Curvas de Relajación.

Para segmentar tumores en las imágenes de resonancia magnética a partir de máquinas de soporte vectorial se expresa la información del problema en pares de entrada salida. En este caso, cada entrada es un vector formado por una fila de la matriz **F** (vector de T2) y su salida asociada corresponde al tejido al cual pertenece el pixel que representa.

Además del tejido invadido por el tumor se considera tejido sano y líquido. La identificación de los mismos en las imágenes, necesaria para entrenar y validar las máquinas de soporte vectorial se realiza con la ayuda de un trabajo del profesor Miguel Martín, et al [13], donde estos tejidos están determinados en las imágenes consideradas.

Como se mencionó en la sección 3.6.5 del Algoritmo del Adatron, el clasificador de máximo margen, máquina de soporte vectorial utilizada, es un clasificador binario, por lo que en este caso, donde se segmentan tres tejidos mas el fondo que se considera aparte (4 clases) se entrenan tres máquinas, cada una de ellas aprende a identificar cada tejido del resto y para construirlas, los datos se dividen de manera que las salidas de entradas correspondientes con el tejido a discriminar son 1 y el resto -1. Es claro que los datos de una clase que se considera en una máquina para identificarla no se considera en las máquinas que le siguen, creadas para reconocer las otras clases. De manera que la máquina general actúa como un árbol de decisión, introduciendo los datos de entrada en la primera máquina que clasifica en el primer tejido o se introduce en la segunda, y así sucesivamente hasta determinar el tejido al cual pertenece el dato. En este trabajo se considera como primer tejido el tumor, luego el tejido sano, el líquido y por último el background (ver figura 4.2). Ahora bien, de acuerdo al paciente, al corte, a los vectores que se consideran para crear cada máquina y a las clases involucradas, se seleccionan los datos de su matriz correspondiente y se agrupan en otra matriz, de donde se toman los datos para construir la máquina de aprendizaje.

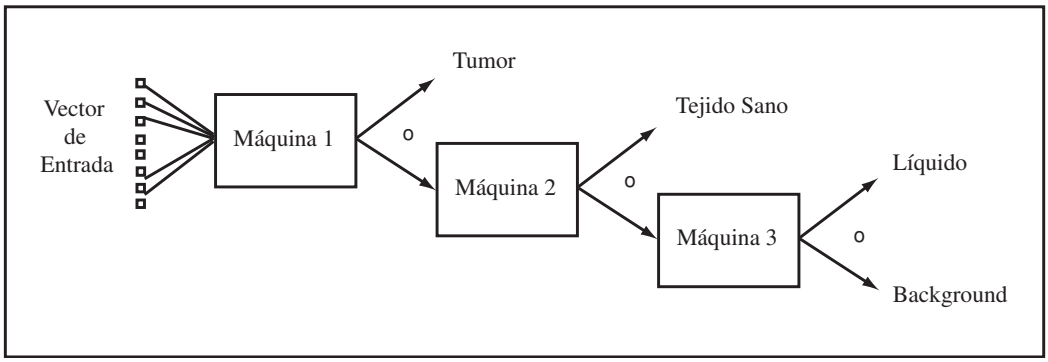


Figura 4.2: Máquina de Soporte Vectorial.

4.1.2. Ruido y Artefactos en las Imágenes de Resonancia Magnética

Como se mencionó en la sección anterior las funciones $F_i = f_{i,.} : \{1, 2, ..., 8\} \to \mathbb{R}$, $i = 1 : n$, son decrecientes, sin embargo se encontraron en todas la matrices $\mathbf{F}$ consideradas, diversas funciones F_i que no eran totalmente decrecientes, es decir, en la mismas se hallaron al menos un $0 < j < 9$ que satisface $F_i(j) > F_i(j-1)$, lo cual es teóricamente imposible por lo

expuesto anteriormente. La causa puede provenir por errores aleatorios (ruido) o sistemáticos, estos últimos probablemente se deban a algún artefacto, por ejemplo algún desperfecto de la antena de detección o movimientos producto de los latidos del corazón del paciente al que se le aplica la resonancia magnética.

Para eliminar estos errores se aplica un análisis de componentes principales a cada matriz **F**. En este análisis las columnas de cada matriz **F**, que contienen los valores de los pixeles de las 8 imágenes de cada corte, son vistas como variables aleatorias, que al aplicarles una transformación lineal se obtienen variables aleatorias no correlacionadas, conocidas como componentes principales, ordenadas en nivel de importancia, es decir, la varianza de la primera es mayor que la segunda y así sucesivamente, por lo que las primeras componentes reúnen la información relevante del problema (sección 2.2). En todos los casos las tres primeras componentes principales contienen mas del 99 % de la información(ver figura 4.3), de manera que se desechan las 5 restantes, luego se regresa al espacio original, donde se reescribe cada matriz en función de dichas componentes. Se supone que el error aleatorio y/o sistemático está contenido en el residuo, conformado por el subespacio generado por las componentes que se descartaron del análisis.

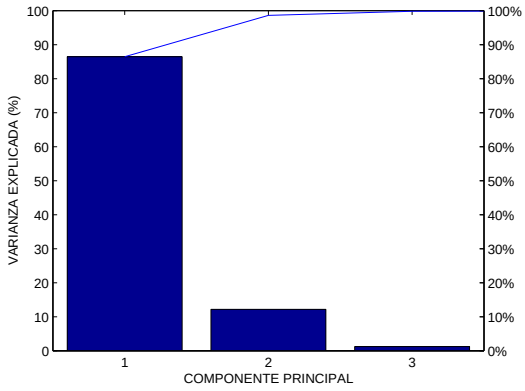
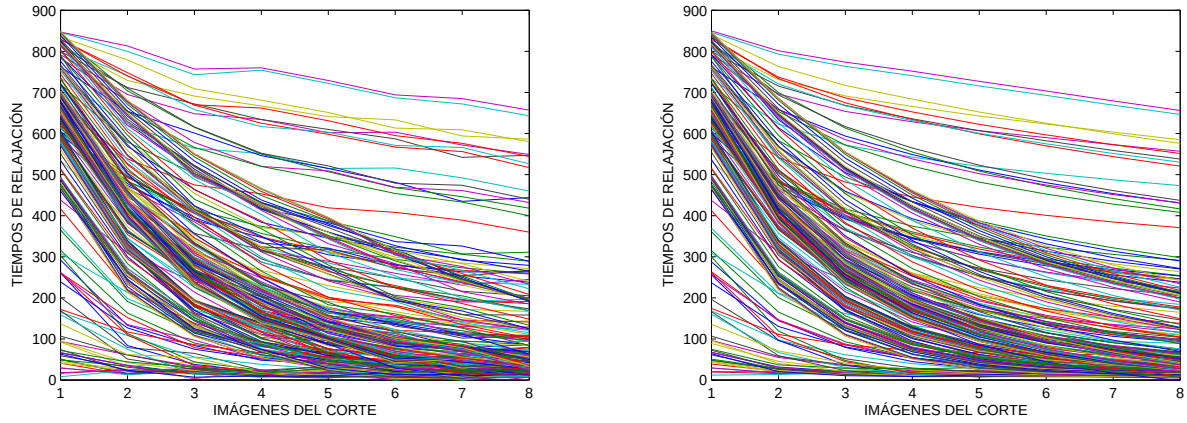


Figura 4.3: Componentes Principales.

Luego del análisis, el problema del no decrecimiento persiste en algunos puntos de distintas matrices **F**. Para solucionar este problema, cada $F_i(j)$ que presente el error, se sustituye por $R_i(j)$, donde R_i es la función obtenida de la regresión exponencial construida a partir de los puntos $(z, F_i(z))$ que no presentan el error, es decir aquellos que satisfacen $F_i(z) > F_i(k)$, $0 < z < k < 9$.



(a) (b)

Figura 4.4: (a) Curvas de Relajación Originales (b) Curvas de Relajación Obtenidas del ACP.

4.1.3. Estandarización de los Datos

Las máquinas de soporte vectorial clasifican datos de acuerdo al criterio del máximo margen, cuyo fundamento se basa en medidas de distancia que son sensibles a las diferencia de escalas y magnitudes entre variables. Es por ello que es necesaria la estandarización de los datos para evitar que las variables con una gran dispersión tengan un mayor efecto en la clasificación.

Los elementos de cada matriz son estandarizados, se calcula la media y la desviación estandar de cada columna de la matriz de datos correspondiente a valores de píxeles de cada imagen y la estandarización de cada elemento de las columnas de la matriz se realiza mediante la siguiente ecuación:

$$P_{ij}^{c} = \frac{P_{ij} - Media_j}{Desviacion_j} = \frac{P_{ij} - \frac{1}{n}\sum_{i=1}^{n} P_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_{ij} - Media_j)^2}},$$

donde

P_{ij}^{c} es el valor del píxel i-ésimo de la imagen j estandarizada presente en la matriz.
P_{ij} es el valor del píxel i-ésimo de la imagen j presente en la matriz.
$Media_j$ es la media de los valores de los pixeles de la imagen j presentes en la matriz.
$Desviacion_j$ es la desviación estándar de los valores de los pixeles de la imagen j presentes en la matriz.

Lo que se consigue con este procedimiento es eliminar las diferencias introducidas por las diferentes escalas de los valores de las distintas imágenes.

4.1.4. Selección de Datos para Entrenar y Validar

Luego de la estandarización de los datos de las diferentes matrices que contienen la información se procede a seleccionar aleatoriamente los vectores de entrenamiento para entrenar las distintas máquinas de aprendizaje y validarlas. Para entrenar, en cada matriz se eligen aleatoriamente aproximadamente 500 datos de la clase a discriminar (etiquetada con 1) y 500 vectores de la clase resto, distribuidos uniformemente sobre los grupos que la componen (etiquetada con -1). Posteriormente los datos restantes son agrupados para validar.

4.2. Implantación y Prueba del Algoritmo del Adatron

El algoritmo del Adatron se implementó en el lenguaje de programación ANSI C con las librerías estándares, de manera que los programas pudieran ser ejecutados bajo cualquier plataforma de cómputo.

El algoritmo de aprendizaje del Adatron se probó con dos problemas de clasificación binaria, donde los vectores de entrada de cada clase corresponden a puntos del plano pertenecientes a regiones distintas y disjuntas del mismo. El primer problema consiste en dos regiones circulares concéntricas convexas y el segundo en dos regiones no convexas definidas como un tablero de ajedrez, identificadas en ambos casos con colores distintos como se pueden observar en la figura 4.5. Los vectores de entrada, como puntos de $\mathbb{R}^2$, son vectores bidimensionales; la salida asociada a éstos que corresponde a la clase a la cual pertenecen se representa con 1 y -1.

En cada caso se realizaron diversos entrenamientos variando los parámetros que caracterizan al algoritmo del Adatron a fin de encontrar la máquina que clasificara la mayor cantidad de vectores del grupo de validación. Los parámetros mencionados son una tasa de aprendizaje λ y dos constantes reales, que corresponden al kernel polinómico $K(x,z) = (\langle x,z\rangle + c)^2$ y al kernel gaussiano $K(x,z) = \exp\left(\frac{-\|x-z\|^2}{\sigma^2}\right)$. Se entrenaron varias máquinas para resolver los dos

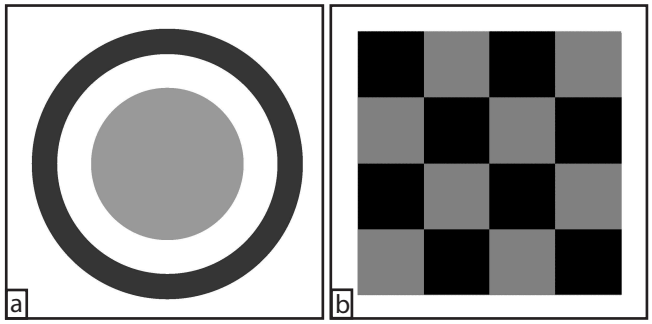


Figura 4.5: (a) Problema Circular (b) Problema del Ajedrez.

problemas, utilizando en todos los casos ambos kernels. En las tablas 4.1 y 4.2 se resumen los parámetros y resultados óptimos que se encontraron.

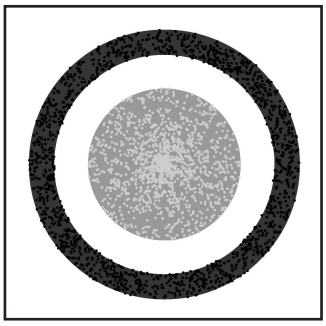


Figura 4.6: Validación del Problema Circular con el Kernel Polinómico y Gaussiano.

En la figura 4.6 se pueden observar los vectores de validación del problema circular de cada región (clase) identificados claramente por cada máquina entrenada con el kernel polinómico y el gaussiano. Este se puede considerar un problema sencillo, en donde bastaron solo 50 vectores de entrenamiento (incluso menos) para clasificar todos los datos de validación considerada y en el cual el kernel polinómico, que induce un espacio de rasgos de dimensión 6 resolvió el problema eficiente y rápidamente, y más aún con la gaussiana, la cual induce un espacio de dimensión infinita. Cabe destacar que en este último caso se varió la tasa de aprendizaje entre 0 y 2 (ver sección 3.6.5). Obteniendo los mejores resultados con los valores más grandes, es decir, alrededor de 1.9.

Kernel	λ	c	σ	Vectores Entrenamiento	Vectores Validación	Validación	Iteraciones	Vectores Soporte
Polinómico	1,0	0,1	—	50	3.000	100 %	214	6
Gaussiano	1.9	—	1,0	50	3.000	100 %	182	12

Tabla 4.1: Parámetros de Aprendizaje óptimos del Problema Circular.

El problema del ajedrez a diferencia del circular es mucho mas complicado, lo cual es de suponerse ya que los conjuntos de cada clase no son convexos, como se puede ver en la figura 4.5 b. En principio se tomaron 160 datos de entrenamiento y se utilizó el kernel polinómico variando la tasa de aprendizaje y la constante c; y en ninguno de los casos el algoritmo converge. Se observa que al fijar una tasa de aprendizaje y tomar un valor cercano a cero de la constante c, el margen es negativo y disminuye gradualmente hasta que se queda alrededor de un valor fijo; y cuando se aumenta c el margen tiende a alejarse al menos infinito, es decir disminuir exponencialmente. De igual manera ocurre cuando se aumentan los vectores de entrada.

Kernel	λ	c	σ	Vectores Entrenamiento	Vectores Validación	Validación	Iteraciones	Vectores Soporte
Polinómico	—	—	—	No Converge	—	—	—	—
Gaussiano	1.9	—	0,2	800	251.001	93.52 %	12.371	58

Tabla 4.2: Parámetros de Aprendizaje óptimos del Problema del Ajedrez.

Por el contrario, al usar el kernel Gaussiano, al variar el valor de sigma junto con la cantidad de vectores de entrenamiento, el algoritmo converge. Para validar tales entrenamientos se realiza un barrido sobre sobre todo el tablero, y la salida de cada máquina se identifica con un punto en el tablero con el color gris o negro según la salida arrojada. En la tabla 4.2 y la figura 4.7 se pueden observar los parámetros de la máquina con la cual se obtuvieron los mejores resultados y el tablero resultante de la misma respectivamente. En este último es de notar que la figura tiene la forma de los cuadros, sin embargo todavía hay muchos vectores que no se clasificaron bien. Para mejorar tal resultado se pueden aumentar la cantidad de vectores de entrenamiento.

Con estas pruebas se concluye que las máquinas de aprendizaje lineales entrenadas con el Adatron pueden aprender a clasificar diversos tipos de datos a través de un conjunto de entrenamiento que mientras más grande es, mejores resultados se obtienen. Además se pudo notar que al aumentar tanto el valor de c y del sigma de los kernels polinómicos y gaussianos respectivamente los vectores soportes tienden a disminuir, y por ende la capacidad de generalización de las máquinas es mayor, propiedad muy importante de este tipo de máquinas.

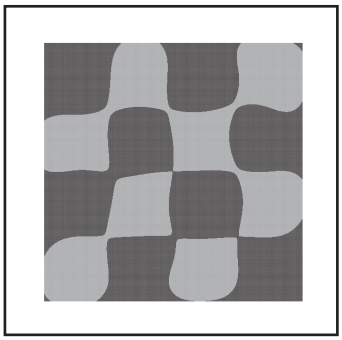


Figura 4.7: Validación del Problema del Ajedrez con el Kernel Gaussiano.

4.3. Segmentación de Imágenes de Resonancia Magnética

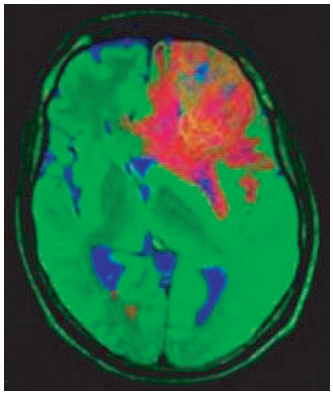
La segmentación de imágenes de resonancia magnética cerebrales se lleva a cabo en imágenes de 8 pacientes distintos. La misma se realiza para identificar el tumor, tejido sano y el líquido en cada corte transversal de cada paciente, a través de una máquina de soporte vectorial en cada caso.

Después de haber tratado el ruido y los artefactos de las imágenes, se obtiene para cada uno de los cortes una matriz de datos $\mathbf{G}$ de dimensión $n \times 8$, n es el número de pixeles de las imágenes del corte,

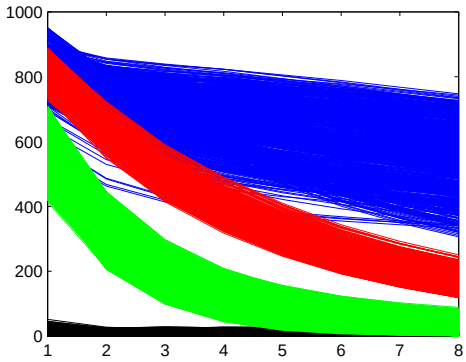
$$\mathbf{G} = (g_{i,j})_{i=1:n,j=1:8}.$$

4.3.1. Caracterización de las Curvas de Relajamiento

Para cada pixel i, $i = 1 : n$, la función $G_i = g_{i,.} : \{1, 2, ..., 8\} \rightarrow \mathbb{R}$ es decreciente y positiva. Para estudiar el comportamiento de estas funciones en las regiones diferenciadas por el Dr. Martín, et al [13], se grafican algunas funciones $(G_i)_{i \in I}$, correspondientes a la matriz **G** del corte 4 del paciente 1, donde el conjunto de subíndices I se selecciona cada vez de las distintas regiones: tumor, tejido sano, líquido y fondo, $I \in \{I_{tumor}, I_{sano}, I_{liquido}, I_{fondo}\}$ obteniéndose los resultados de la Gráfica 4.8.



(a)



(b)

Figura 4.8: (a) Selección de las curvas de distintas regiones (b) Se distinguen claramente los 4 grupos, de abajo hacia arriba son: fondo, tejido sano, tumor y líquido.

La figura 4.8 sugiere que las curvas se pueden caracterizar por su valor inicial $K_i = G_i(1)$ y una medida de velocidad de relajación muy sencilla $V_i = G_i(1) - G_i(7)$. El par (G, V) es el vector aleatorio definido de Ω en $\mathbb{R}^2$, donde Ω es el espacio de las curvas de relajación. En la figura 4.9 se gráfica el histograma de las observaciones del vector aleatorio $(K_i, V_i)_{i=1:n}$ y en la figura 4.10 sus curvas de nivel.

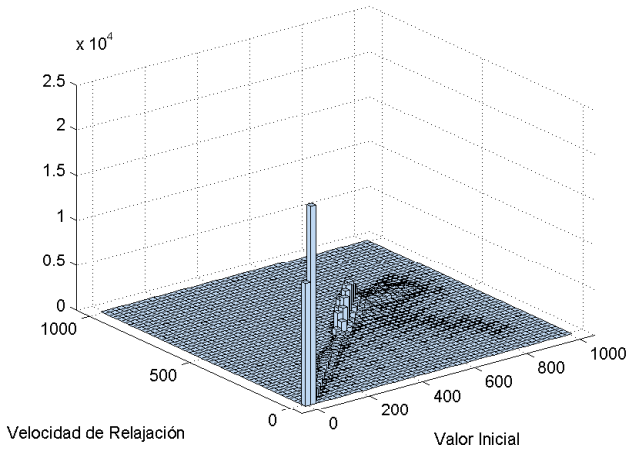
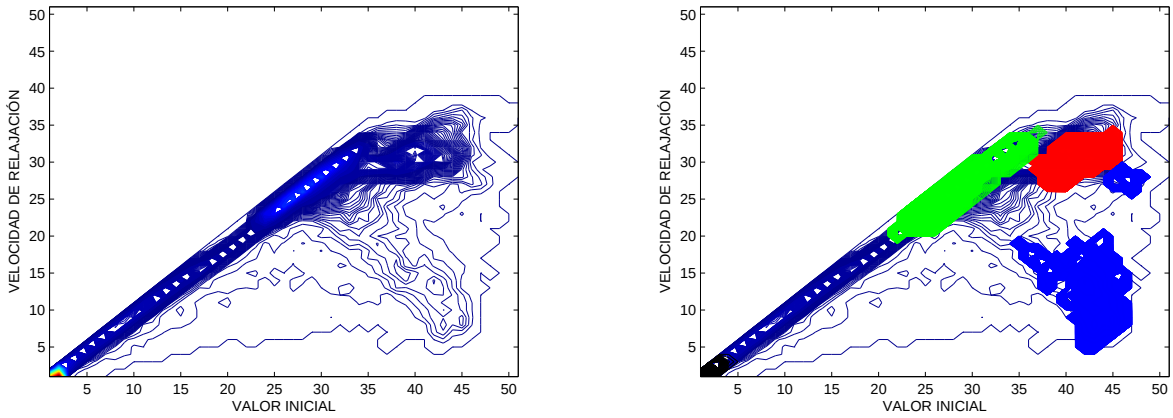


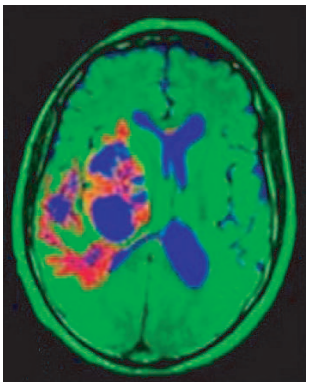
Figura 4.9: Histograma de las observaciones del vector aleatorio $(K_i, V_i)_{i=1:n}$.



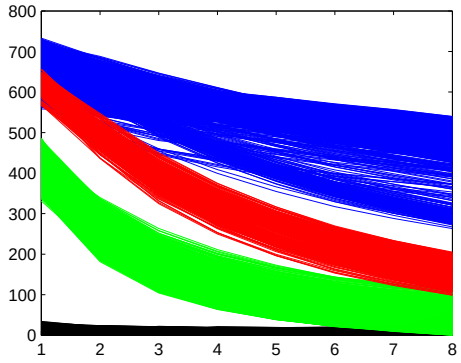
(a) (b)

Figura 4.10: (a) Curvas de Nivel del Hist. 1 (b) Regiones de los grupos en las Curvas de Nivel del Hist. 1.

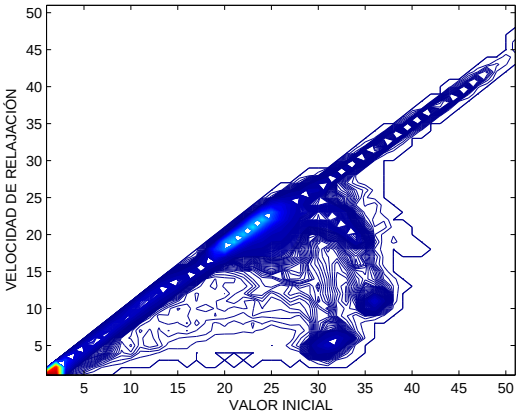
En la figura 4.10 se pueden observar las zonas de las curvas de nivel de los histogramas (CNHs) que corresponden a las curvas de los distintos grupos descritos en la figura 4.8. Claramente se distinguen las diferentes clases en regiones distintas del gráfico. Para establecer algún patrón de comportamiento se realizan los mismos gráficos correspondientes al corte 4 del paciente 2, los cuales se muestran en la figura 4.11.



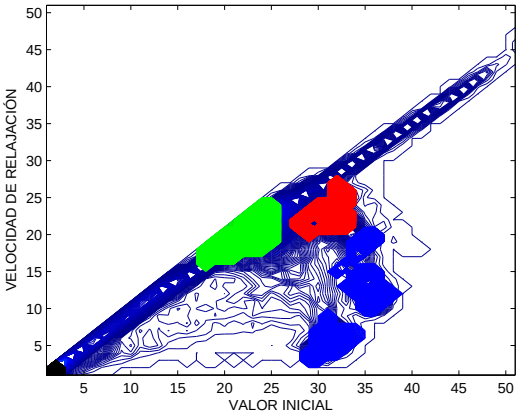
(a)



(b)



(c)

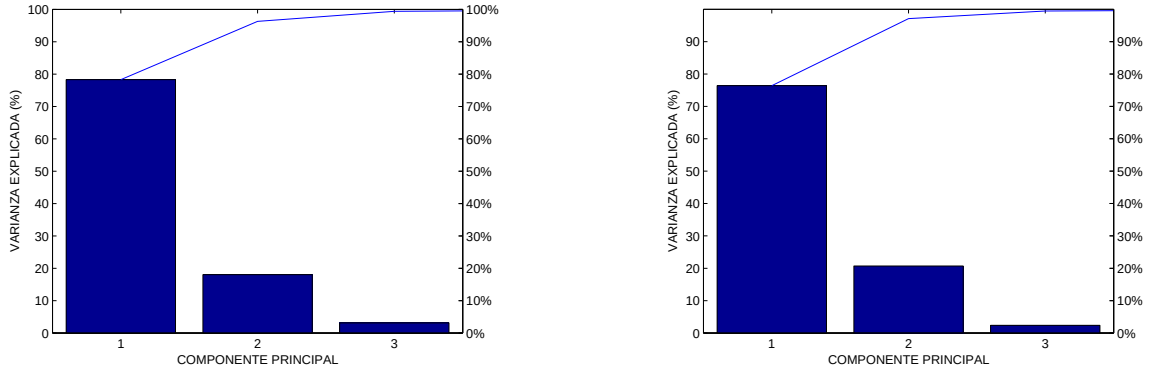


(d)

Figura 4.11: (a) Selección de las curvas de las distintas regiones del Corte 4 del Paciente 2 (b) Se distinguen claramente los 4 grupos, de abajo hacia arriba son: fondo, tejido sano, tumor y líquido (c) CNH del Corte 4 del paciente 2 (d) Regiones de los grupos en las CNH.

Al comparar las CNHs de ambos pacientes y las regiones de los distintos grupos considerados se puede observar un patrón de comportamiento y en donde la ubicación de las zonas de cada clase en ambos gráficos son similares. Ahora bien, con las curvas indexadas en I_{tumor}, I_{sano}, $I_{liquido}$ y I_{fondo} se entrena una máquina de soporte vectorial en cada caso con el fin de identificar el tumor el tejido sano y el líquido en los cortes mencionados de ambos pacientes. Dichas máquinas son alimentadas con los valores de las curvas de relajación obtenidas en el espacio de componentes principales, es decir, se realiza un análisis de componentes principales en cada caso para disminuir la dimensionalidad del problema, de ocho a tres, de manera de reducir el tiempo de entrenamiento y de visualizar los vectores de entrada espacialmente.

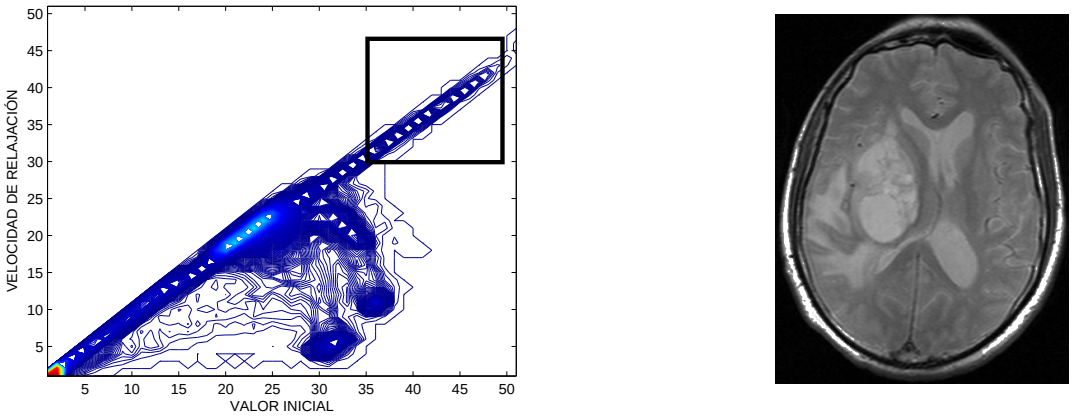
Al aplicar esta técnica en los cortes de los dos pacientes mencionados anteriormente se obtienen en cada caso los valores de su matriz **G** en un espacio donde las primeras columnas (componentes principales) contienen la mayor cantidad de información, es decir, son aquellas cuyas varianzas reúnen generalmente un 99 % o más de la variabilidad de los datos; por esta razón el resto de las componentes se desechan. En la figura 4.12 se puede observar que las tres primeras componentes principales en los dos casos estudiados están por encima del 99 %. Esto permite entonces alimentar las máquinas de soporte vectorial con vectores de $\mathbb{R}^3$ en lugar de vectores de $\mathbb{R}^8$.



(a) (b)

Figura 4.12: (a) Componentes Principales del Corte 4 del Paciente 1 (b) Componentes Principales del Corte 4 del Paciente 2.

Por otro lado, en las CNHs del corte 4 de los pacientes 1 y 2 (figuras 4.10 y 4.11 respectivamente), se observa que en el último aparece un pico en la parte superior que no existe en el primero, y el cual no está dentro de las regiones identificadas por las curvas de relajación seleccionadas. En la figura 4.13 se muestra en blanco la región de la imagen que el pico representa, se observa que corresponde a tejido alrededor del hueso, de manera que este pico se considera como tejido sano, por lo que se toman curvas de relajación del mismo para entrenar la máquina de soporte vectorial correspondiente al paciente 1, además de las mencionadas anteriormente (figura 4.11).



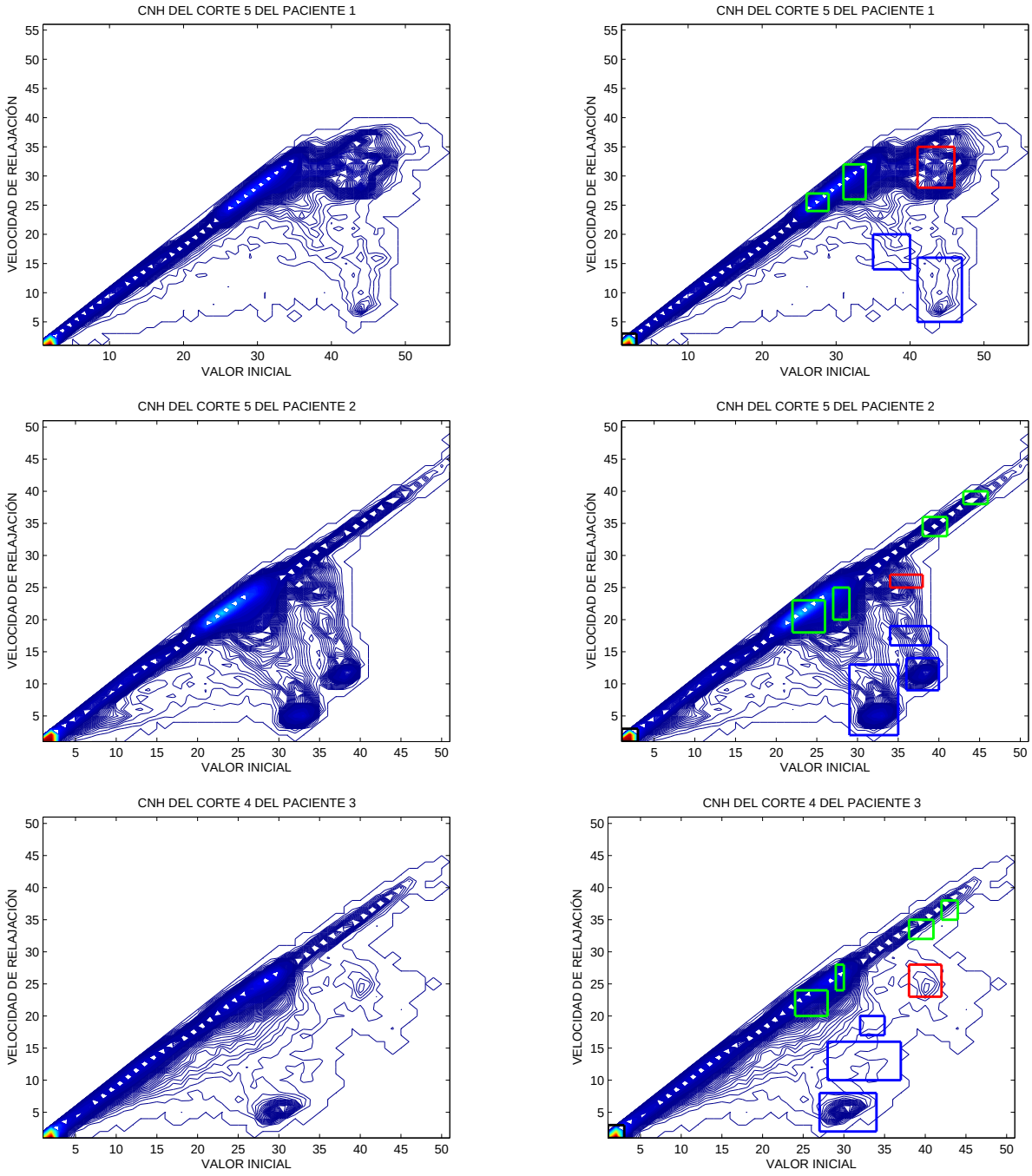
(a) (b)

Figura 4.13: (a) CNH del corte 4 del paciente 1(b) Región de la imágen del pac. 1 correspondiente a la zona considerada de las CNH.

4.3.2. Validación de los Patrones de Comportamiento

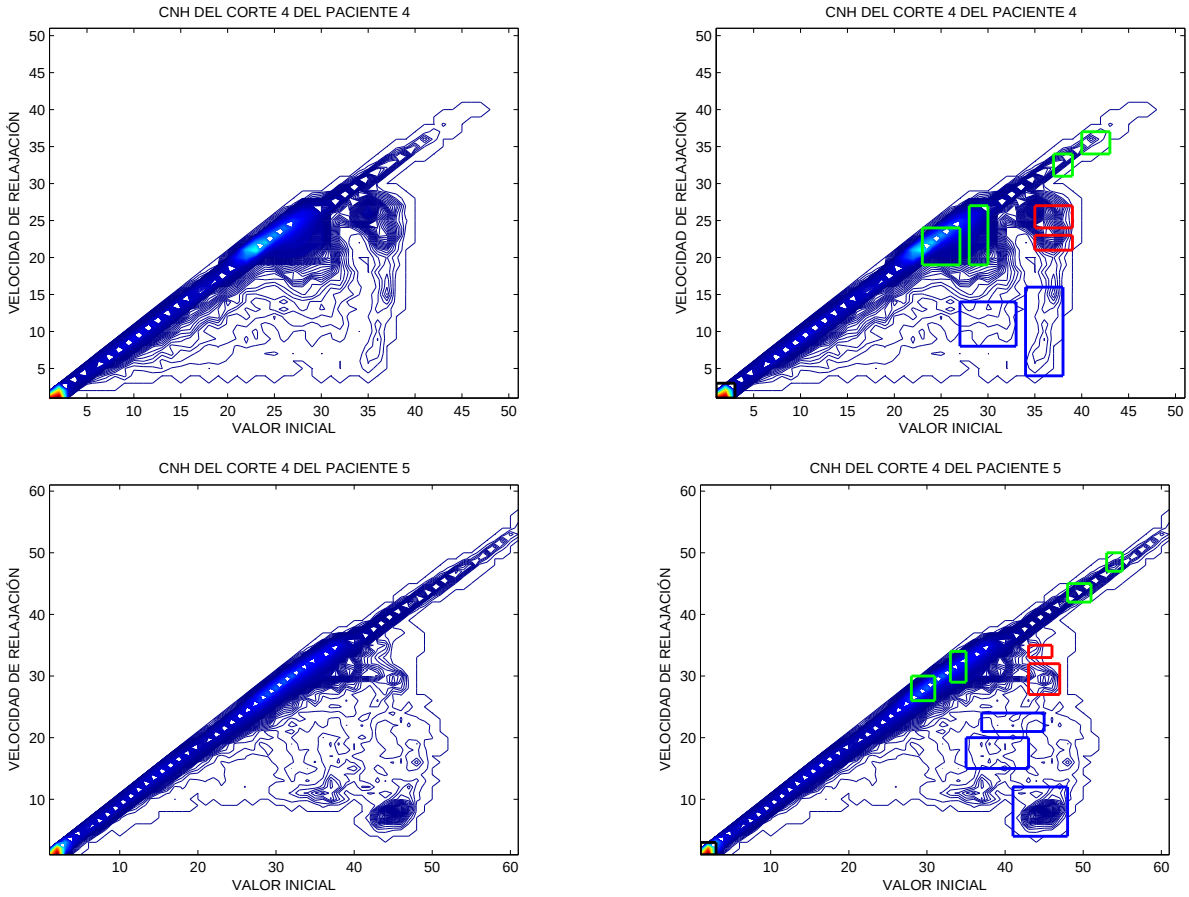
Para validar el patrón de comportamiento encontrado en la sección anterior se cuenta con las imágenes de distintos cortes de diferentes pacientes, donde se toman las curvas de relajación correspondientes a cada grupo directamente del gráfico de las CNHs siguiendo el patrón mencionado (ver figuras 4.10, 4.11 y 4.13). Las mismas, de igual forma que en la sección anterior son consideradas en el espacio de componentes principales, donde se utilizan para entrenar una máquina de soporte vectorial en cada corte considerado. Son relevantes solo las tres primeras componentes principales en todos los casos, debido a que reúnen mas del 99 % de la información, es decir que los vectores de entradas de las máquinas de aprendizaje están conformados por vectores de tres dimensiones.

En las figuras 4.14 y 4.15 se presentan las CNHs correspondientes a las imágenes del corte 5 de los pacientes 1 y 2, y del corte 4 de los pacientes 3, 4 y 5 y al lado de cada una, siguiendo el patrón mencionado anteriormente se seleccionan las regiones del tumor, el tejido sano, el líquido y el background, identificadas dentro de los rectángulos de color rojo, verde, azul y negro respectivamente.



(a) (b)

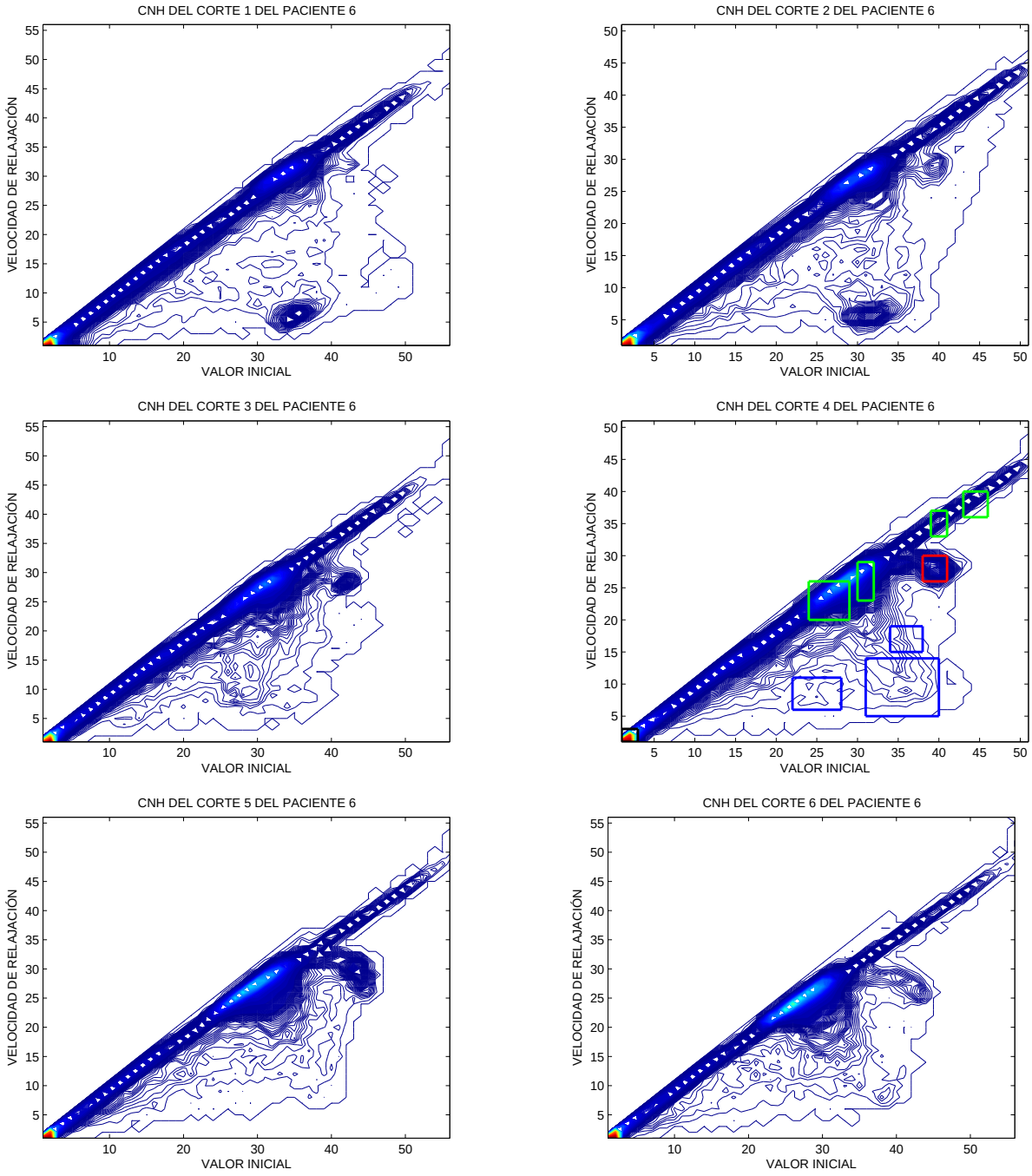
Figura 4.14: (a) CNHs de pacientes 1 y 2 del Corte 5 y del paciente 3 del Corte 4 (b) Selección de las regiones de los grupos en las CNHs.



(a) (b)

Figura 4.15: (a) CNHs de pacientes 4 y 5 del Corte 4 (b) Selección de las regiones de los grupos en las CNHs.

Del paciente 6 se disponen imágenes de 8 cortes, cuyas CNHs se pueden observar en las figuras 4.16 y 4.17, donde se nota la variación de un corte a otro. Se toma el corte 4 para entrenar una máquina de soporte vectorial con el fin de identificar los distintos tejidos en dicho corte. Los datos de entrenamiento y validación se selecciona según el patrón mencionado anteriormente en las CNH.



(a) (b)

Figura 4.16: CNHs del paciente 6 de cortes 1,2,3,4,5 y 6.

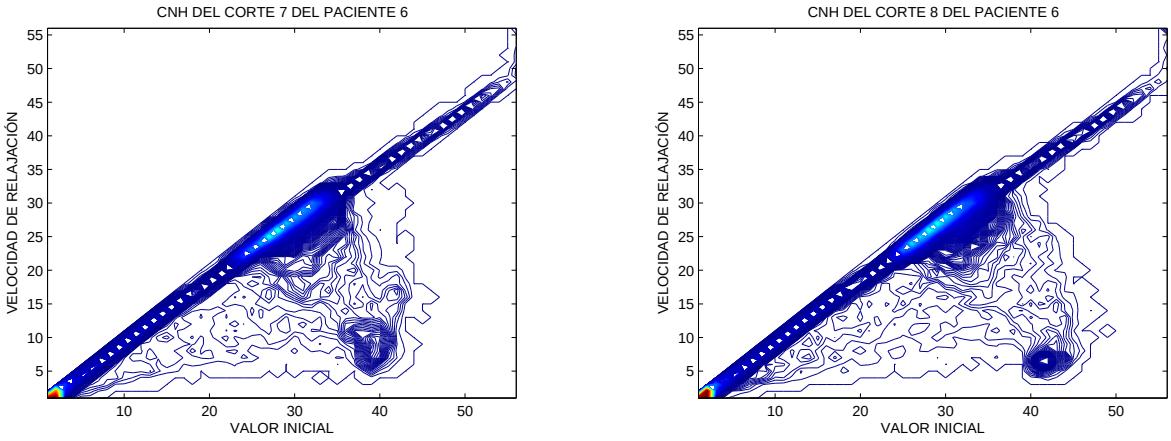


Figura 4.17: CNHs del paciente 6 de cortes 7 y 8.

Se cuenta con las imágenes de los pacientes 7 y 8 del corte 4, la figura 4.18 representa las CNHs de las imágenes y se observa que la región del tumor no está muy claro como en los casos anteriores. Con el objeto de distinguir mejor los grupos, se observan en la figuran 4.19 las curvas de relajación que se caracterizan por el vector aleatorio (K, V). En ellas se observa que el valor de la cuarta imagen es relativamente distinto para cada grupo, lo que sugiere redefinir la variable K como el valor de la cuarta imagen en lugar del valor de la primera (K_4), y construir los dos histogramas correspondientes a las observaciones del nuevo vector conjunto $(K_{4i}, V_i)_{i=1:n}$.

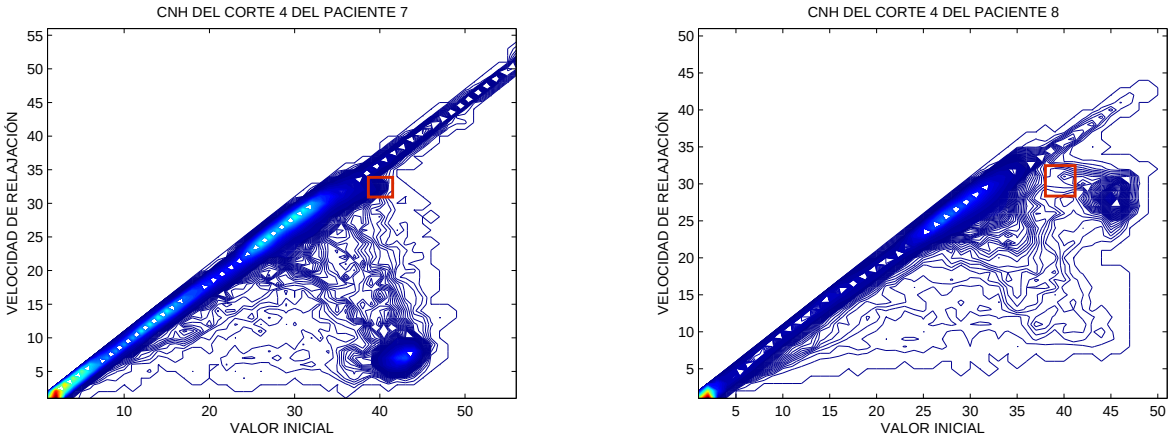


Figura 4.18: CNHs de pacientes 7 y 8 del corte 4 .

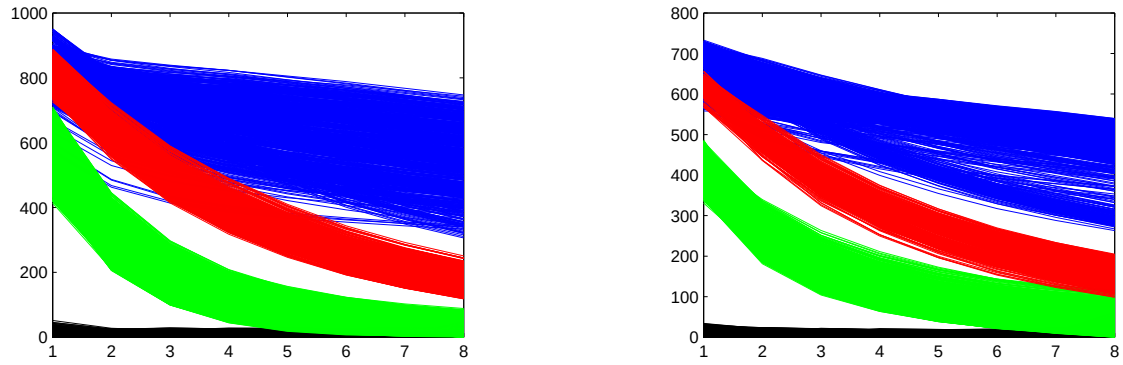


Figura 4.19: Curvas de relajación de los distintos grupos de los pacientes 1 y 2 del corte 4 .

Tomando las curvas de nivel de los nuevos histogramas se distinguen mejor las regiones del tumor y el tejido sano (ver figuras 4.20 y 4.21), las máquinas de soporte vectoriales en estos casos se construyen con los valores de las curvas de relajación tomadas en las nuevas CNHs, manteniendo el mismo patrón donde se seleccionan las regiones de los distintos grupos identificadas con los rectángulos de distintos colores que se observan en las figuras mencionadas.

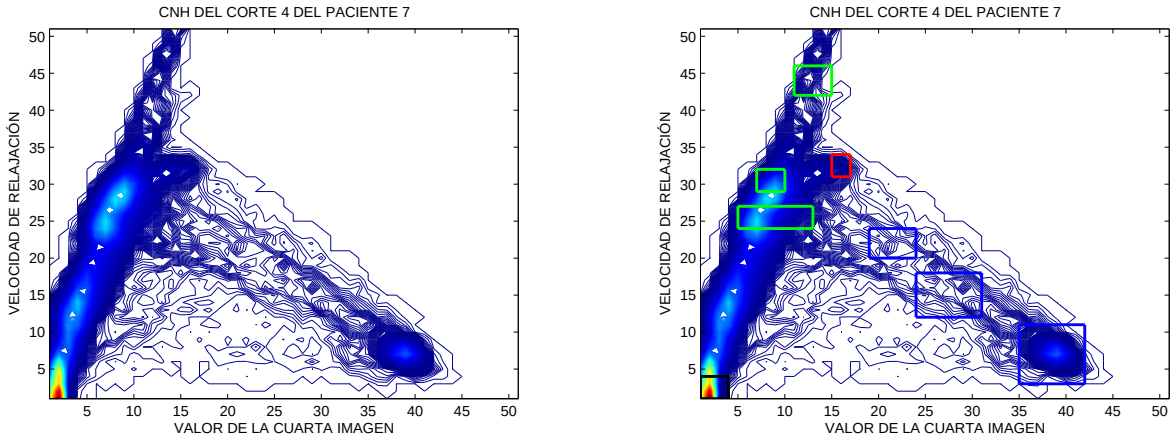
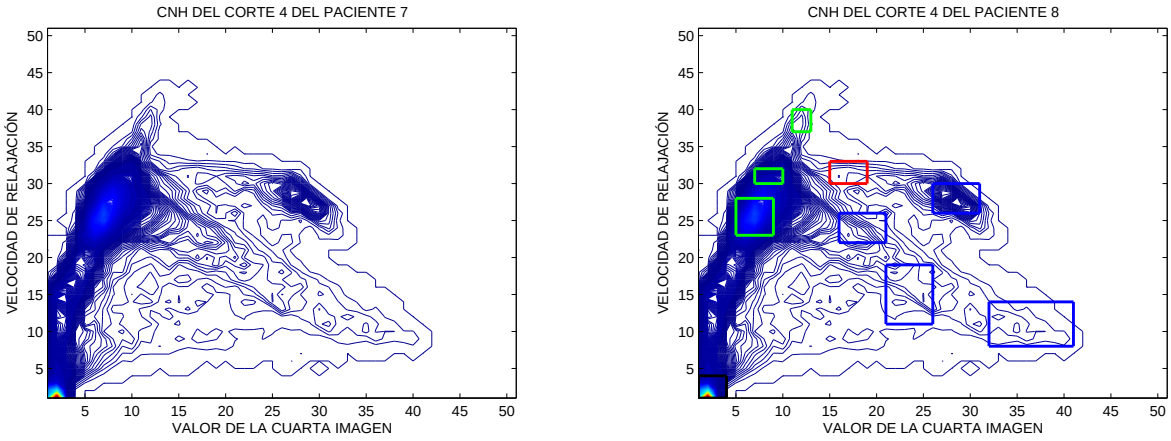


Figura 4.20: (a) CNH del corte 4 del paciente 8 (b) Selección de las regiones de los grupos en las CNH.



(a) (b)

Figura 4.21: (a) CNH del corte 4 del paciente 8 (b) Selección de las regiones de los grupos en las CNH.

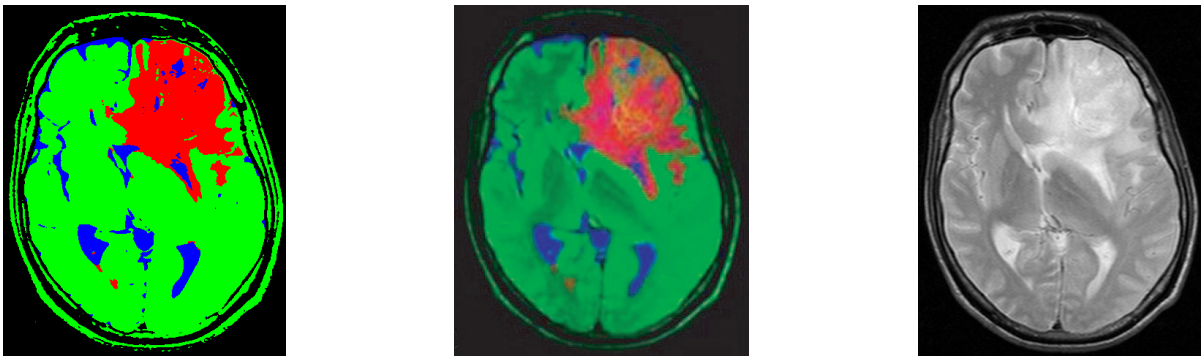
Capítulo 5

Resultados y Conclusiones

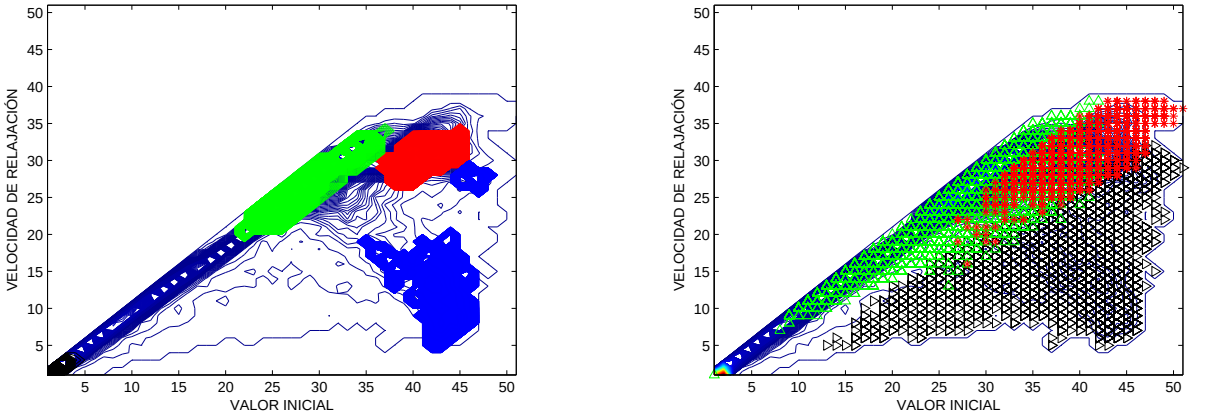
En este último capítulo se muestran y analizan las segmentaciones de las imágenes de resonancia magnética obtenidas de las distintas máquinas de soporte vectorial en base a la información extraída de los procesos de entrenamiento y validación. Posteriormente se presentan las conclusiones que derivan del análisis y algunas recomendaciones que se consideran convenientes.

5.1. Resultados

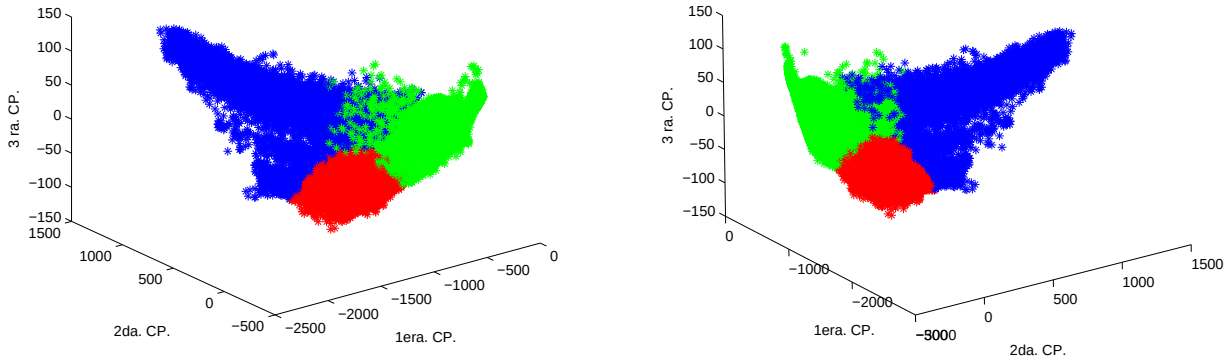
Las segmentaciones de los distintos cortes de los diferentes pacientes obtenidas de las máquinas de soporte vectorial descritas en el capítulo anterior se presentan a continuación. Cada segmentación se acompaña de la primera imagen original del corte considerado, las CNH, el mapeo de las distintas regiones generadas por la MSV en las CNH y los vectores de T2 del corte en el espacio de componentes principales. En la mayoría de los casos se muestran las segmentaciones obtenidas por el Dr. Miguel Martín, et al. [13], con el fin de visualizar los resultados. Por otro lado, la cantidad de datos utilizados para diseñar cada máquina de aprendizaje y los parámetros de entrenamiento óptimos de las mismas se resumen en tablas presentadas junto con las demás figuras. Cabe destacar que en la construcción de las máquinas se utiliza el kernel gaussiano.



(a) (b) (c)



(d) (e)

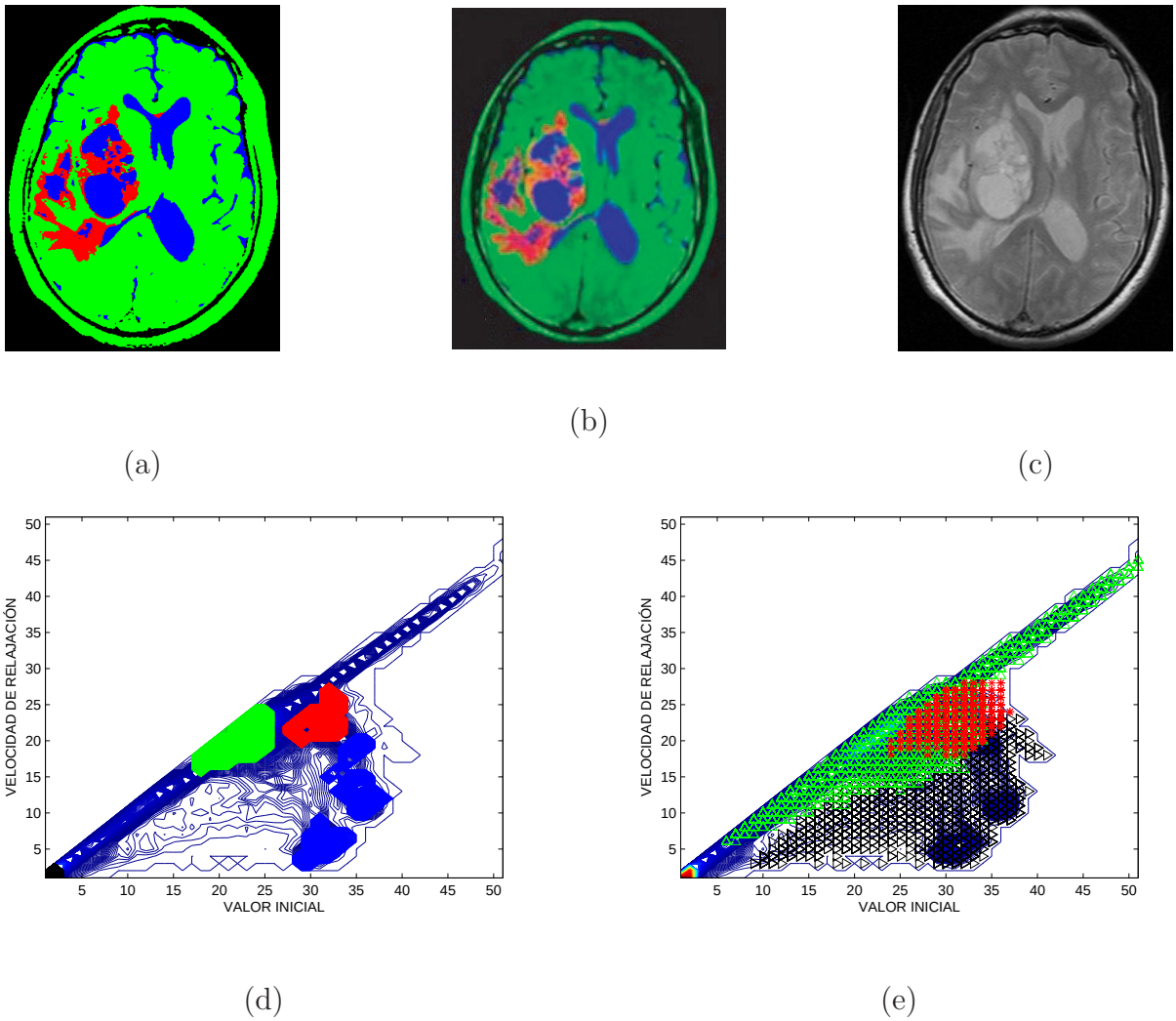


(f) (g)

Figura 5.1: (a) Segmentación del corte 4 del paciente 1 con la MSV diseñada (b) Segmentación del profesor Martín, et al. (c) Imagen original (d) CNH (e) Mapeo de las regiones obtenidas por la MSV en las CNH (el color negro representa al líquido) (f)(g) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	1.0	1000	10380	0	100 %	100 %	100 %	19	21
Tejido Sano vs R.	1.9	3.0	1000	9331	0	100 %	100 %	100 %	35	29
Liquido vs R.	1.9	3.0	1000	1494	0	100 %	100 %	100 %	3	9

Tabla 5.1: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 1.



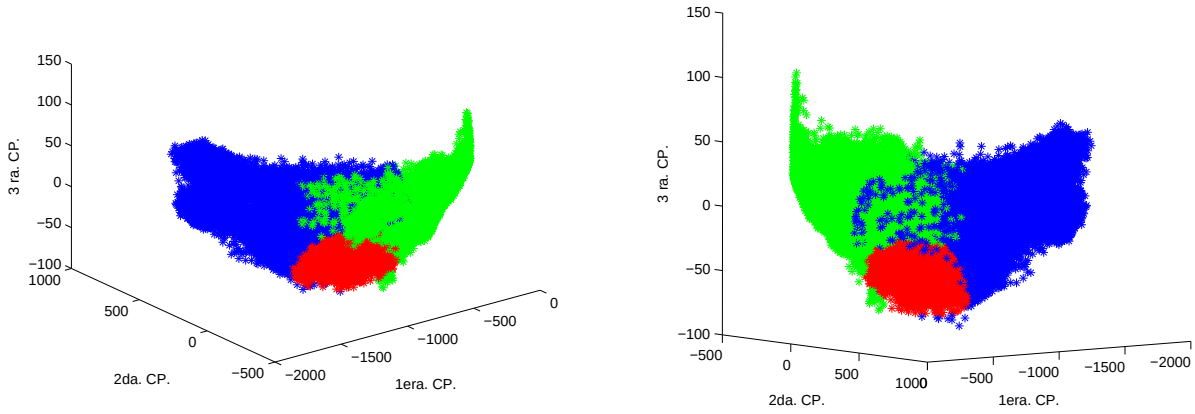
(a) (b) (c)

(d) (e)

Figura 5.2: (a) Segmentación del corte 4 del paciente 2 con la MSV diseñada (b) Segmentación del profesor Martín, et al. (c) Imagen original (d) CNH (e) Mapeo de las regiones obtenidas por la MSV en las CNH .

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	0.4	460	9106	0	100 %	100 %	100 %	60	91
Tejido Sano vs R.	1.9	15.0	1000	8250	0	100 %	100 %	100 %	336	24
Liquido vs R.	1.9	10.0	1000	2807	0	100 %	100 %	100 %	25	2

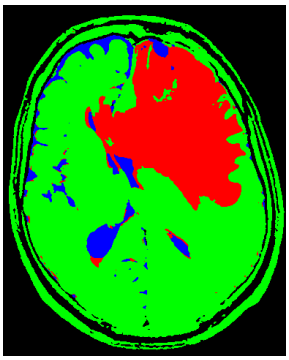
Tabla 5.2: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 2.



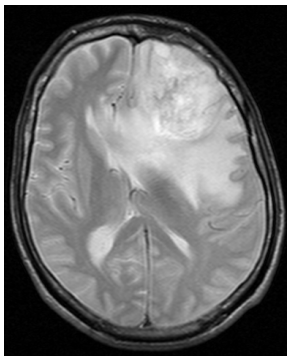
(a) (b)

Figura 5.3: (a)(b) Vectores de T2 del corte 4 del paciente 2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

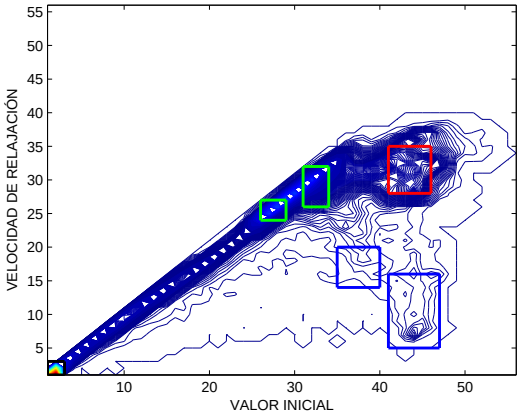
En las figuras anteriores se observan las segmentaciones del corte 4 de los pacientes 1 y 2, muy similares a las obtenidas por el profesor Martín, et al [13]. Las regiones activadas de cada grupo en las CNHs siguen el patrón descrito en el capítulo anterior. En las tablas de los parámetros óptimos de las máquinas de aprendizaje se observa que se obtuvo un porcentaje de patrones bien clasificados del 100 %. El uso del ACP permite visualizar los vectores de T2 de cada corte en sectores distintos de $\mathbb{R}^3$ según las clases generadas por las MSVs (figuras 5.1 f, g y 5.3). Los resultados del corte 5 de los pacientes 1 y 2 se observan en las figuras 5.4 y 5.5, donde los vectores de entrenamiento se tomaron directamente de las regiones de las CNH según el patrón mencionado, a diferencia de los casos anteriores que se tomaron de acuerdo a las segmentaciones del profesor Martín, et al [13].



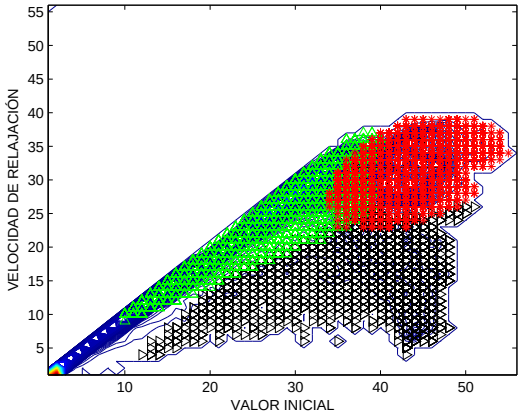
(a)



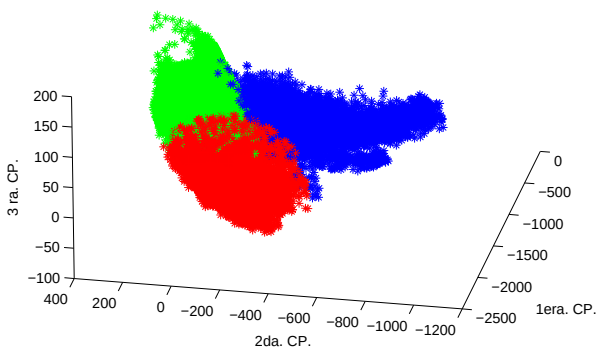
(b)



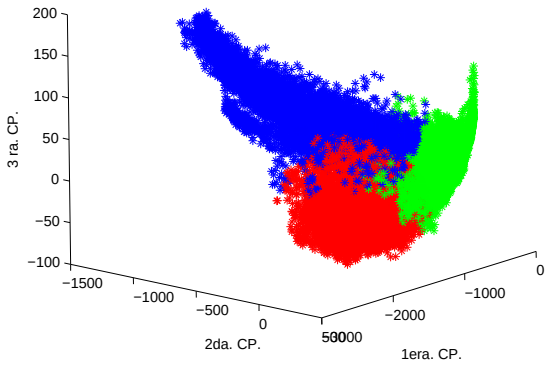
(c)



(d)



(e)



(f)

Figura 5.4: (a) Segmentación del corte 5 del paciente 1 con la MSV diseñada (b) Imagen original (c) CNH (d) Mapeo de las regiones obtenidas por la MSV en las CNH (e)(f) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

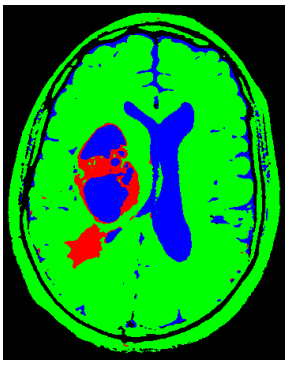
Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	4.0	1000	111533	9	99.99 %	100 %	99.99 %	181	17
Tejido Sano vs R.	1.9	10.0	1000	97243	0	100 %	100 %	100 %	80	16
Liquido vs R.	1.9	3.0	1000	49301	0	100 %	100 %	100 %	7	13

Tabla 5.3: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 5 del paciente 1.

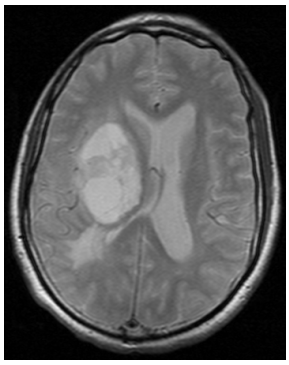
	1	**2**	**3**	**4**	**% Clase**
1	0	0	0	0	100 %
2	9	0	0	0	99.98 %
3	0	0	0	0	100 %
4	0	0	0	0	100 %
% de Clasificación General					99.99 %

Tabla 5.4: Matriz de rendimiento de la MSV del corte 5 del paciente 1.

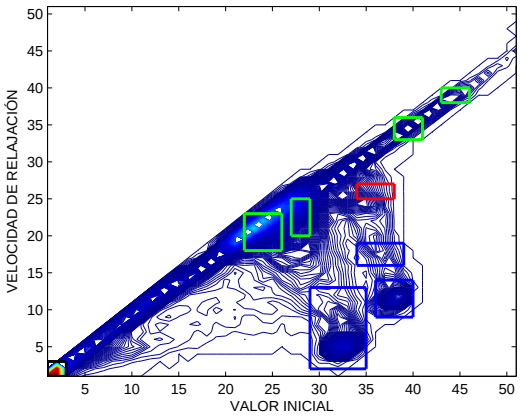
Las MSVs del corte 5 de los pacientes 1 y 2 correspondientes al tumor (ver tablas 5.3 y 5.5) clasifican correctamente los patrones de validación en un 99.99 % en ambos casos y un 100 % las máquinas restantes que identifican a los otros tejidos y generan las segmentaciones mostradas en cada caso en las figuras 5.4 y 5.5 respectivamente. Ambas segmentaciones son ligeramente distintas a las obtenidas en el corte 4, en el paciente 1, la región del tumor es mas grande. Además, se confirma el patrón descrito en el capítulo anterior en las CNH, donde se activan las mismas regiones para cada grupo. Por otro lado, en las tablas mencionadas se puede observar el número de errores que cometen las MSVs en cada caso. Para conocer a que clase pertenecen los patrones involucrados en los errores y la clase errónea que genera cada máquina de aprendizaje, se crea la matriz de rendimiento, cuya entrada ubicada en la fila i y columna j corresponde al número de patrones de la clase i que la máquina clasifica en la clase j (las clases 1, 2, 3 y 4 corresponden al tumor, tejido sano, líquido y fondo respectivamente). La matriz de rendimiento se construye al introducir en la MSV los datos cuyas clases están identificadas. En el paciente 1 se observa que la máquina clasificó 9 vectores de T2 de tejido sano como tumor.



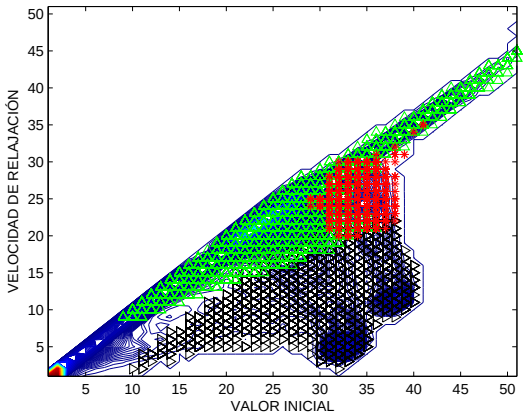
(a)



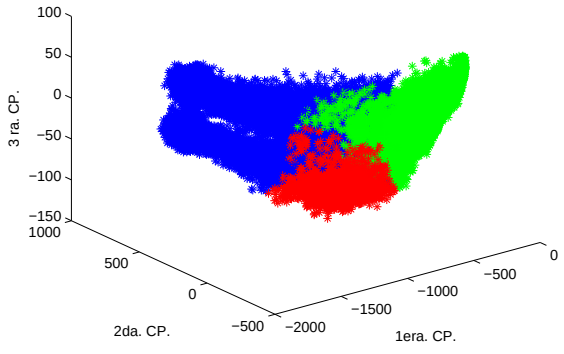
(b)



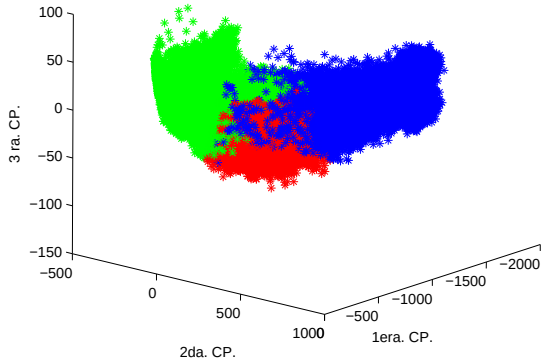
(c)



(d)



(e)



(f)

Figura 5.5: (a) Segmentación del corte 5 del paciente 2 con la MSV diseñada (b) Imagen original (c) CNH (d) Mapeo de las regiones obtenidas por la MSV en las CNH (e)(f) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

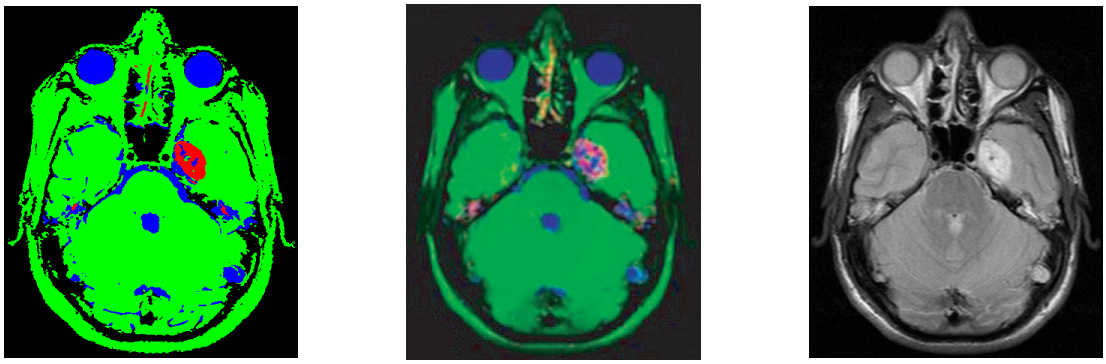
Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	2.0	1000	104618	8	99.99 %	99.82 %	99.83 %	65	13
Tejido Sano vs R.	1.9	10.0	1000	103558	7	99.99 %	99.99 %	100 %	361	19
Liquido vs R.	1.9	3.0	1000	51193	0	100 %	100 %	100 %	4	13

Tabla 5.5: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 5 del paciente 2.

	1	**2**	**3**	**4**	**% Clase**
1	0	1	0	0	99.91 %
2	7	0	6	0	99.98 %
3	0	0	0	0	100 %
4	0	0	0	0	100 %
% de Clasificación General					99.99 %

Tabla 5.6: Matriz de rendimiento de la MSV del corte 5 del paciente 2.

En la parte que sigue se muestran los resultados de las MSVs que segmentan al corte 4 de los pacientes 3, 4, 5 y 6, donde los datos de entrenamiento y validación se toman directamente de las CNHs según el patrón descrito en la metodología.



(a) (b) (c)

Figura 5.6: (a) Segmentación del corte 4 del paciente 3 con la MSV diseñada (b) Segmentación del profesor Martín, et al. (c) Imagen original.

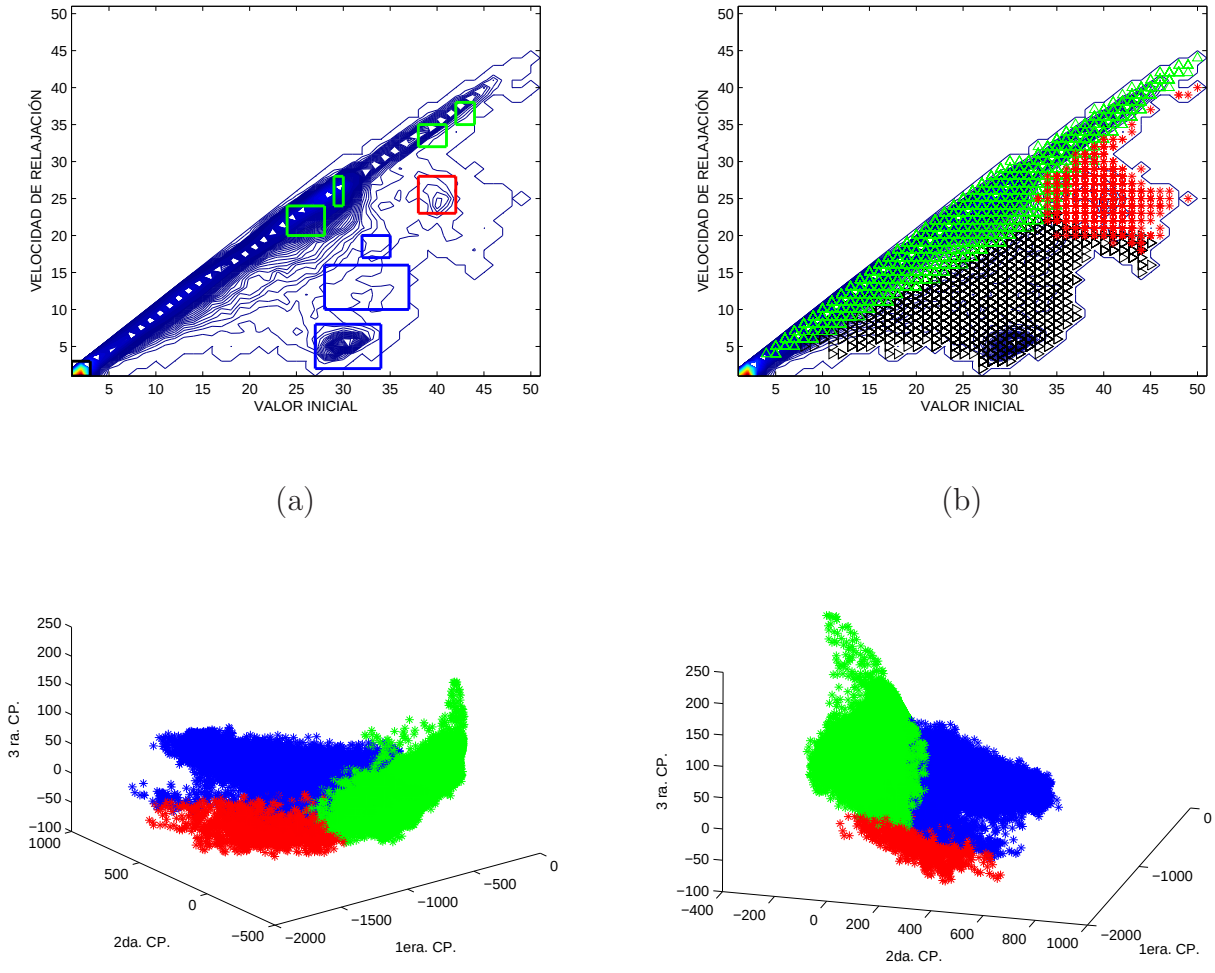


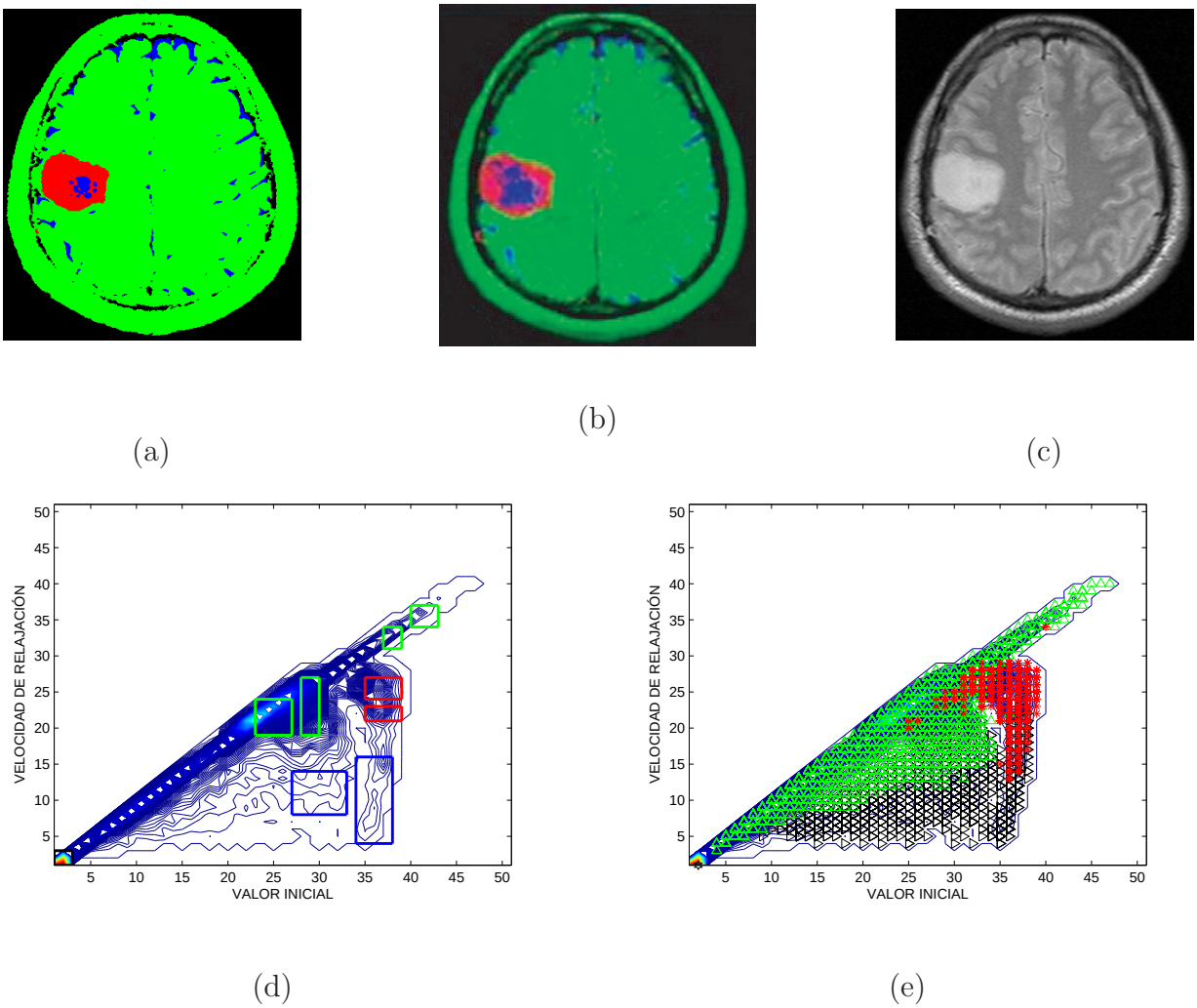
Figura 5.7: (a) CNH del corte 4 del paciente 3 (b) Mapeo de las regiones obtenidas por la MSV en las CNH (c)(d) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	3.0	1000	94567	18	99.98 %	100 %	99.98 %	787	18
Tejido Sano vs R.	1.9	5.0	1000	93648	0	100 %	100 %	100 %	258	17
Liquido vs R.	1.9	2.0	1000	61119	0	100 %	100 %	100 %	6	27

Tabla 5.7: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 3.

	1	2	3	4	% Clase
1	0	0	0	0	100 %
2	17	0	0	0	99.95 %
3	1	0	0	0	99.98 %
4	0	0	0	0	100 %
% de Clasificación General					99.98 %

Tabla 5.8: Matriz de rendimiento de la MSV del corte 4 del paciente 3.



(a) (b) (c)

(d) (e)

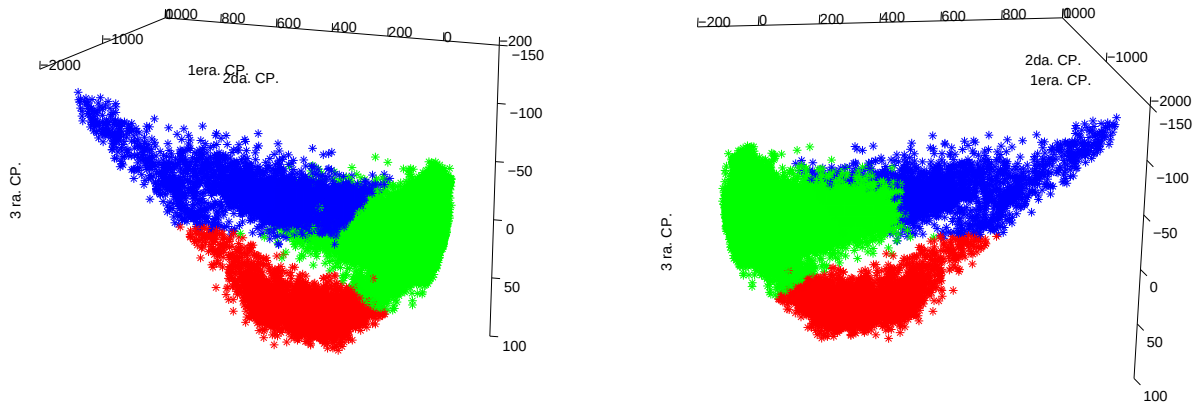
Figura 5.8: (a) Segmentación del corte 4 del paciente 4 con la MSV diseñada (b) Imagen original (c) CNH (d) Segmentación obtenida por el profesor Martín, et al. (e) Mapeo de las distintas regiones obtenidas por la MSV en las CNH .

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	10.0	1000	84047	28	99.97 %	100 %	99.97 %	686	10
Tejido Sano vs R.	1.9	20.0	1000	81983	0	100 %	100 %	100 %	452	25
Liquido vs R.	1.9	30.0	1000	33264	0	100 %	100 %	100 %	118	16

Tabla 5.9: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 4.

	1	**2**	**3**	**4**	**% Clase**
1	0	0	0	0	100 %
2	25	0	0	0	99.95 %
3	3	0	0	0	99.72 %
4	0	0	0	0	100 %
% de Clasificación General					99.97 %

Tabla 5.10: Matriz de rendimiento de la MSV del corte 4 del paciente 4.



(a) (b)

Figura 5.9: (a)(b) Vectores de T2 del corte 4 del paciente 4 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

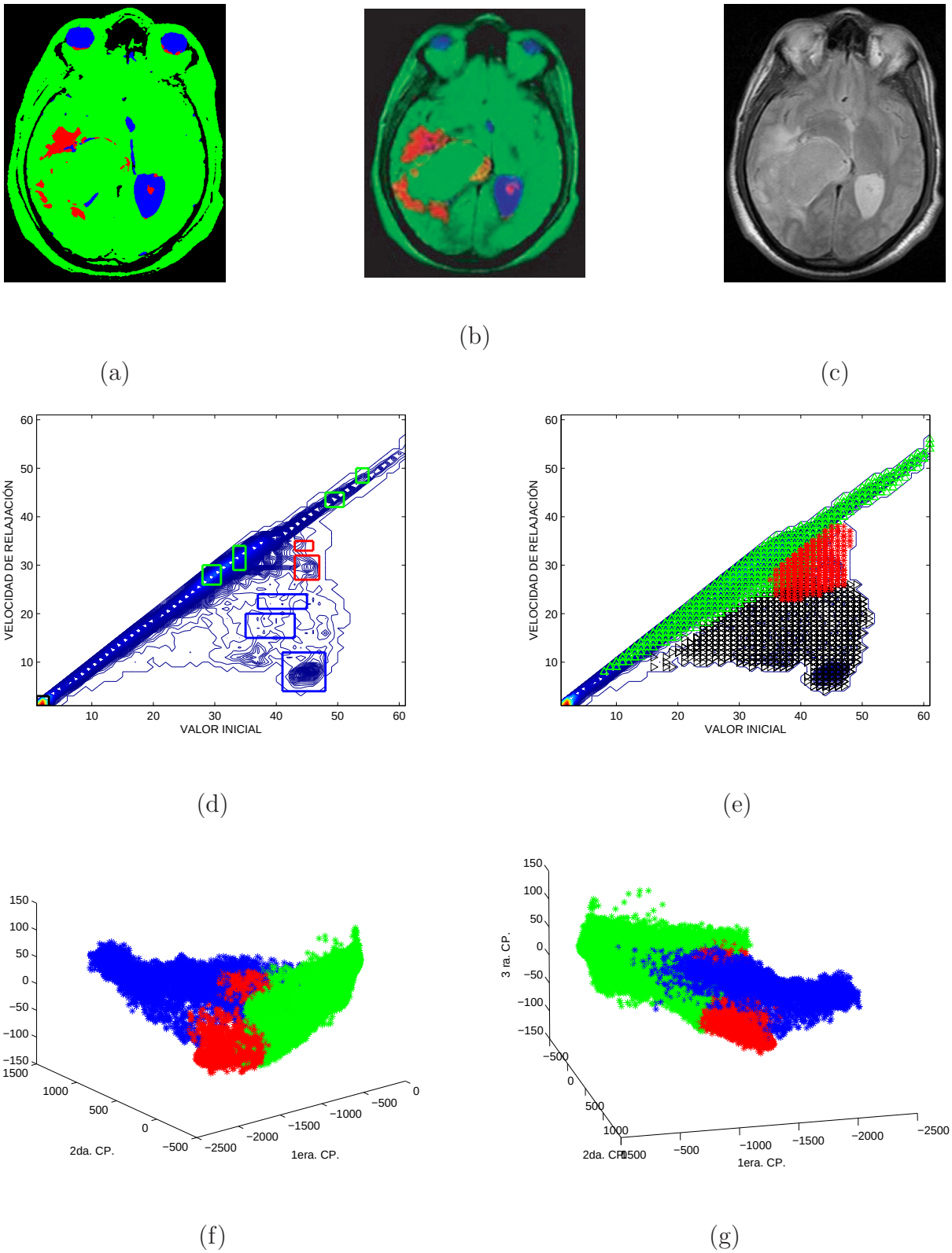


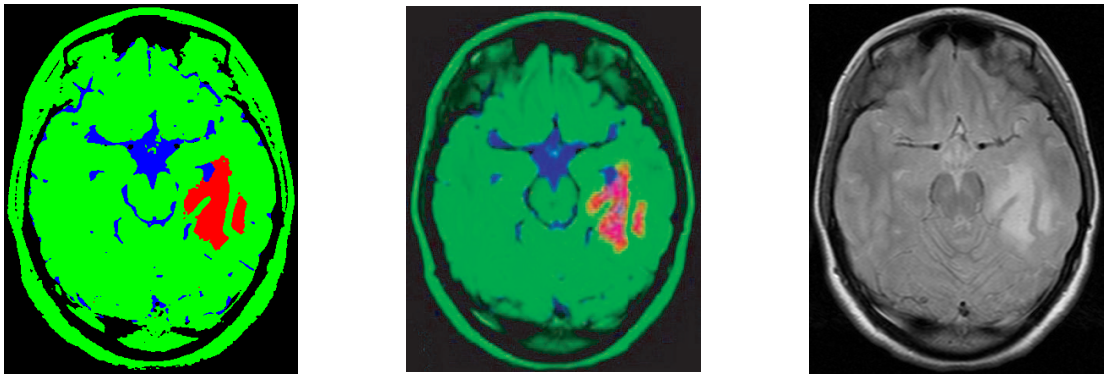
Figura 5.10: (a) Segmentación del corte 4 del paciente 5 con la MSV diseñada (b) Segmentación del profesor Martín, et al. (c) Imagen original (d) CNH (e) Mapeo de las regiones obtenidas por la MSV en las CNH (f)(g) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	2.0	800	80681	41	99.95 %	98.99 %	99.96 %	2720	13
Tejido Sano vs R.	1.9	10.0	1000	75133	0	100 %	100 %	100 %	222	26
Liquido vs R.	1.9	20.0	1000	37031	0	100 %	100 %	100 %	99	11

Tabla 5.11: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 5.

	1	**2**	**3**	**4**	**% Clase**
1	0	5	0	0	99.44 %
2	2	0	0	0	99.99 %
3	34	0	0	0	98.86 %
4	0	0	0	0	100 %
% de Clasificación General					99.99 %

Tabla 5.12: Matriz de rendimiento de la MSV del corte 4 del paciente 5.



(a) (b) (c)

Figura 5.11: (a) Segmentación del corte 4 del paciente 6 con la MSV diseñada (b) Segmentación del profesor Martín, et al [13]. (c) Imagen original.

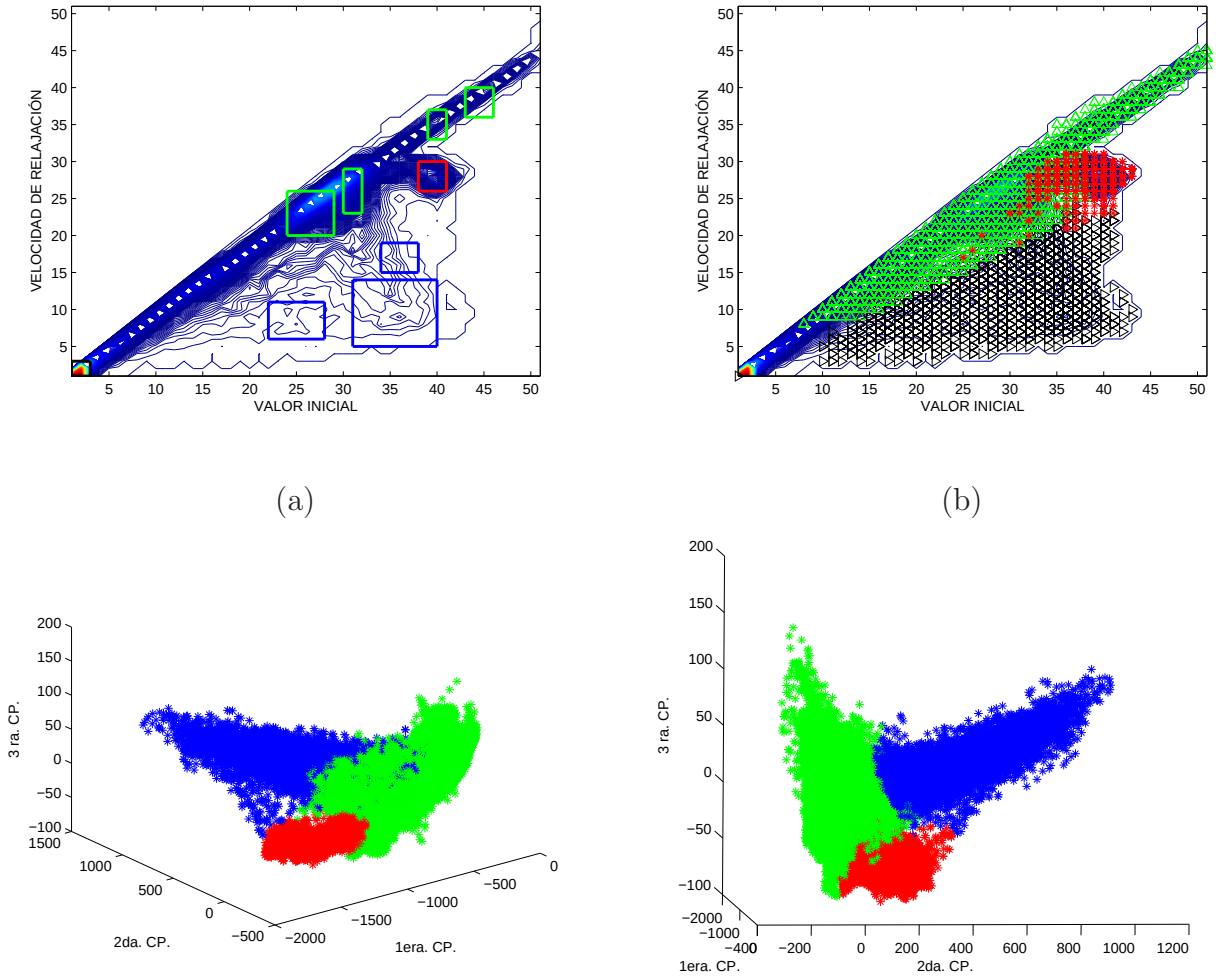


Figura 5.12: (a) CNH del corte 4 del paciente 6 (b) Mapeo de las regiones obtenidas por la MSV en las CNH (c)(d) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	1.0	1000	94536	19	99.98 %	99.85 %	99.98 %	46	43
Tejido Sano vs R.	1.9	2.5	1000	92011	0	100 %	100 %	100 %	43	25
Liquido vs R.	1.9	3.0	1000	37269	0	100 %	100 %	100 %	3	13

Tabla 5.13: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 6.

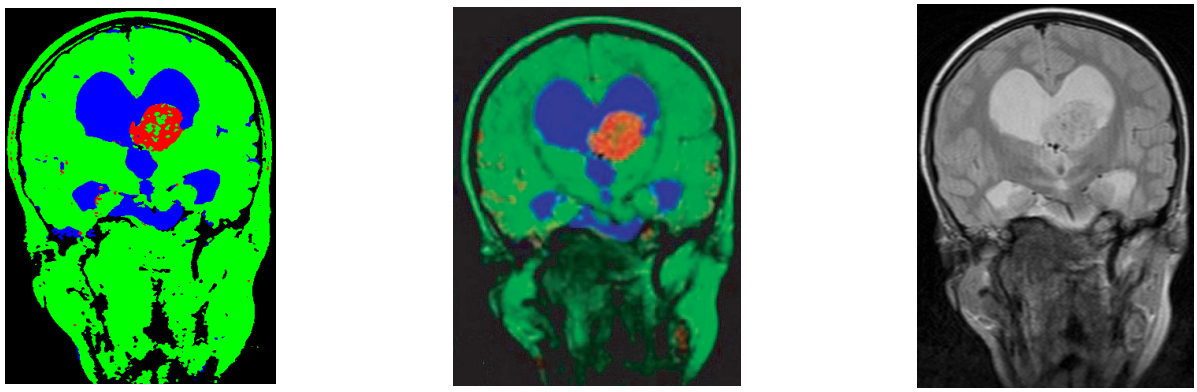
	1	2	3	4	% Clase
1	0	3	0	0	99.88 %
2	16	0	0	0	99.97 %
3	0	0	0	0	100 %
1	0	0	0	0	100 %
% de Clasificación General					99.98 %

Tabla 5.14: Matriz de rendimiento de la MSV del corte 4 del paciente 6.

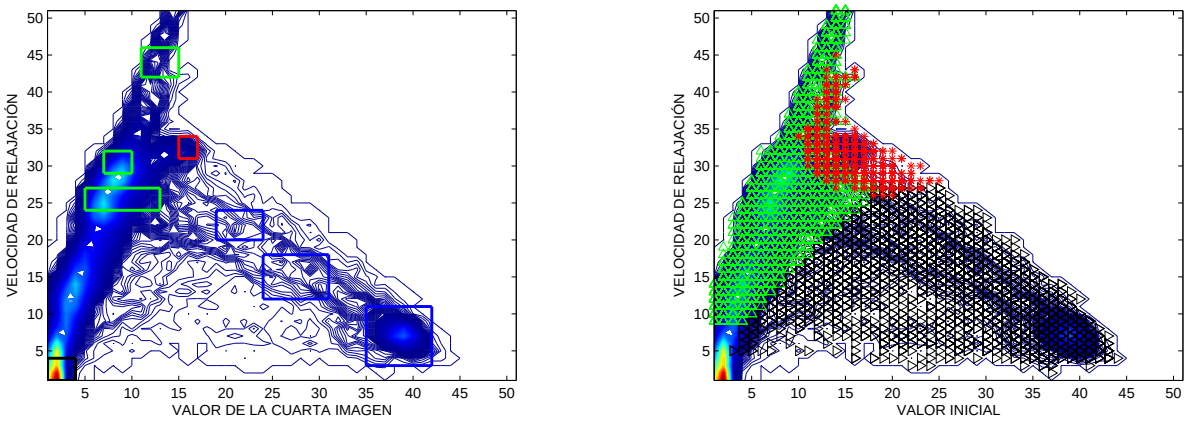
Se puede observar que las segmentaciones del corte 4 de los pacientes 3,4,5 y 6 son muy similares a las obtenidas por el profesor Martín, et al [13]. Estas segmentaciones fueron obtenidas por las MSVs con solo 1000 vectores de entrenamiento, aproximadamente 1 % de los vectores de T2 que conforman cada corte. En las figuras 5.7, 5.8, 5.10 y 5.12 se muestran en las CNHs las regiones creadas por cada MSV correspondientes a los distintos tejidos. En estas se observan zonas de transición donde en un mismo punto se identifican tejidos diferentes, caracterizadas por puntos de colores y formas distintas. Esto ocurre porque cada punto de las CNHs representa a un conjunto de vectores de T2, que son clasificados en clases diferentes.

El porcentaje de patrones bien clasificados sobre los datos de validación de cada una de estas máquinas de aprendizaje es mayor que el 99.9 %, en sus matrices de rendimiento se observa que los pocos errores que presentan corresponden a vectores identificados como tejido sano que son confundidos con tumor, salvo el paciente 5, donde el líquido es confundido con tumor.

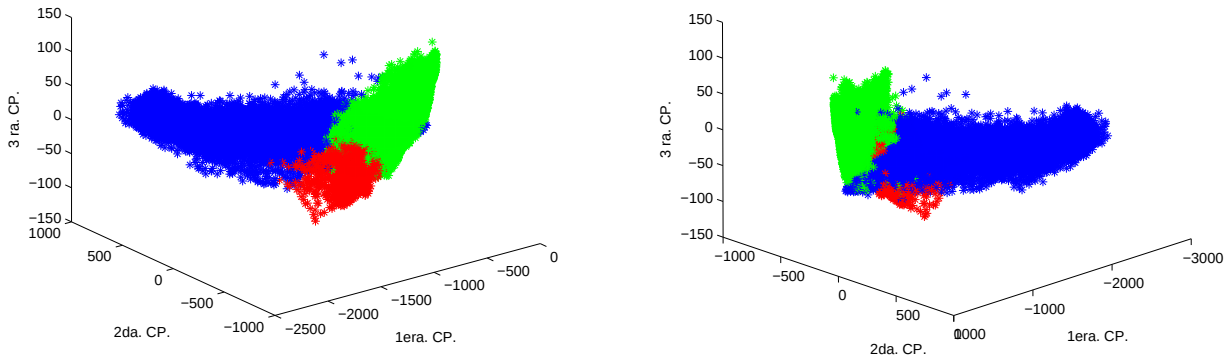
A continuación se presentan los resultados de las MSVs que segmentan el corte cuatro de los pacientes 7 y 8. Como se mencionó en el capítulo anterior, las CNHs de estos pacientes corresponden a histogramas bidimensionales del vector aleatorio (K_4, V) donde K_4 corresponde al valor de la cuarta imagen, a diferencia de los casos anteriores que consideran el valor de la primera imagen. Por esta razón el histograma cambia un poco, sin embargo mantiene el mismo comportamiento; por lo que se toman los datos de entrenamiento y validación siguiendo el patrón.



(a) (b) (c)



(d) (e)



(f) (g)

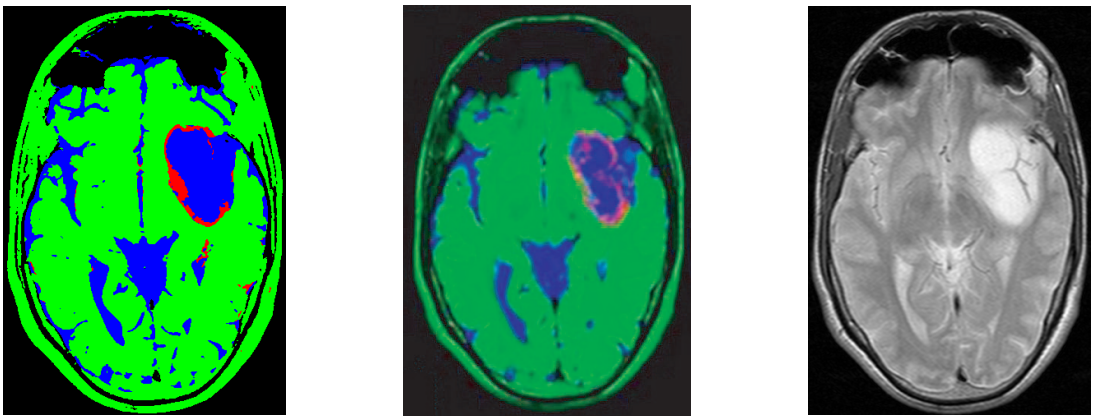
Figura 5.13: a) Segmentación del corte 4 del paciente 7 con la MSV diseñada (b) Segmentación del profesor Martín, et al. (c) Imagen original (d) CNH (e) Mapeo de las regiones obtenidas por la MSV en las CNH (f)(g) Vectores de T2 del corte 4 del paciente 7 en el espacio de componentes principales identificados con un color según la salida generada por la MSV

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	4.0	800	47671	7	99.99 %	100 %	99.99 %	9	17
Tejido Sano vs R.	1.9	10.0	1000	76607	0	100 %	100 %	100 %	1609	31
Liquido vs R.	1.9	15.0	1000	39815	0	100 %	100 %	100 %	464	18

Tabla 5.15: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 7.

	1	**2**	**3**	**4**	**% Clase**
1	0	0	0	0	100 %
2	7	0	0	0	99.98 %
3	0	0	0	0	100 %
4	0	0	0	0	100 %
% de Clasificación General					99.99 %

Tabla 5.16: Matriz de rendimiento de la MSV del corte 4 del paciente 7.



(a) (b) (c)

Figura 5.14: (a) Segmentación del corte 4 del paciente 8 con la MSV diseñada (b) Segmentación del profesor Martín, et al. (c) Imagen original.

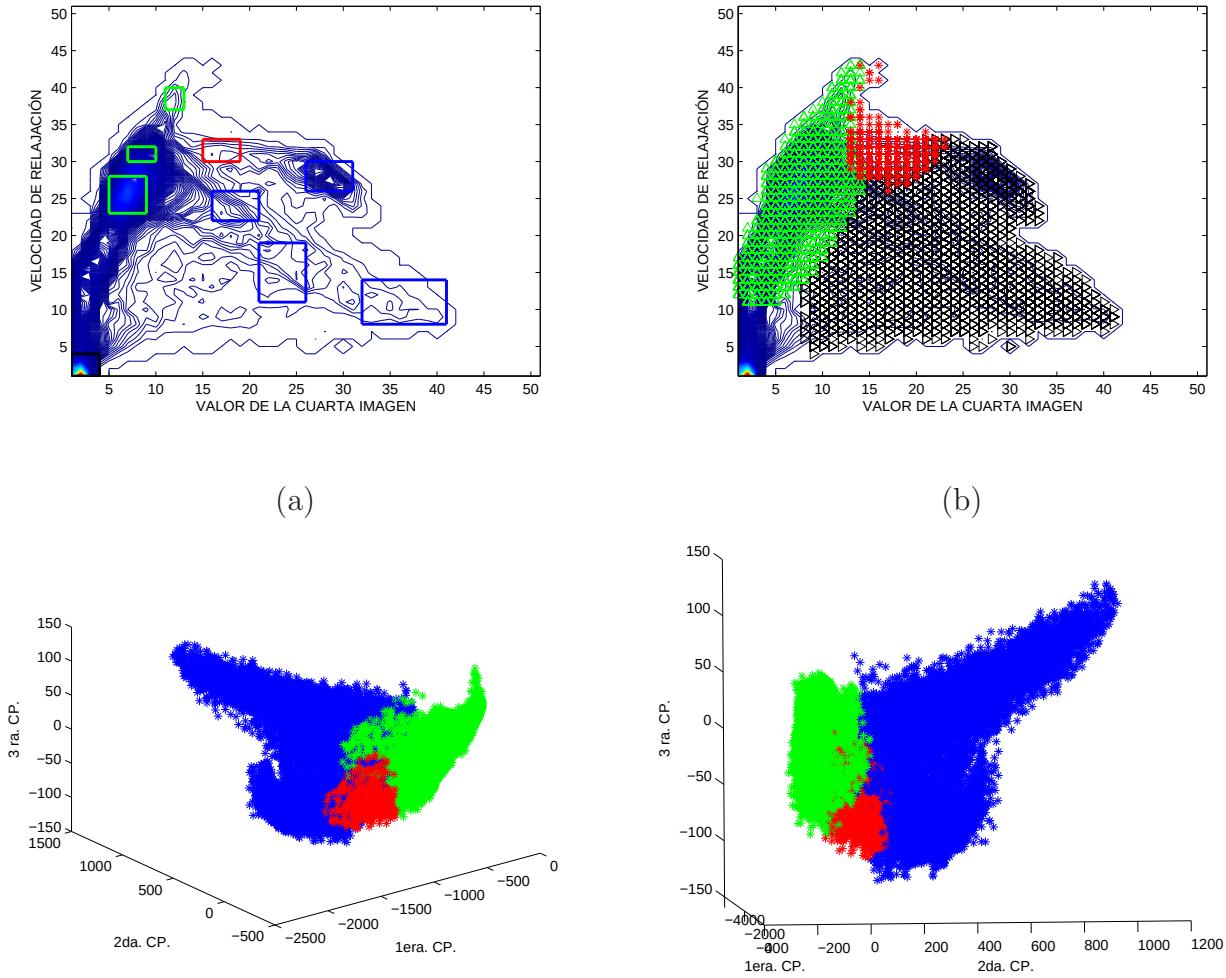


Figura 5.15: (a) CNH (b) Mapeo de las regiones obtenidas por la MSV en las CNH (c)(d) Vectores de T2 en el espacio de componentes principales identificados con un color según la salida generada por la MSV.

Entrenamiento	λ	σ	Pat. Ent.	Pat. Val.	Error Val.	% Val.	1^{era} Clase	2^{da} Clase	Iteraciones	Vectores Soporte
Tumor vs R.	1.9	3.0	1000	103583	3	99.99 %	99.65 %	99.99 %	740	18
Tejido Sano vs R.	1.9	1.5	1000	102627	0	100 %	100 %	100 %	76	13
Liquido vs R.	1.9	4.0	1000	50316	0	100 %	100 %	100 %	38	15

Tabla 5.17: Parámetros de aprendizaje óptimos de la MSV que segmenta al corte 4 del paciente 8.

	1	2	3	4	% Clase
1	0	0	1	0	99.87 %
2	1	0	0	0	99.99 %
3	1	0	0	0	99.98 %
4	0	0	0	0	100 %
% de Clasificación General					99.99 %

Tabla 5.18: Matriz de rendimiento de la MSV del corte 4 del paciente 8.

En las tres figuras anteriores se observan los resultados de los dos últimos pacientes, donde se nota el mismo patrón estudiado. El hecho de considerar las CNHs correspondientes al vector (K_4, V) permite visualizar mejor la zona entre el tejido sano y el tumor. Las segmentaciones de nuevo son muy parecidas a las obtenidas por el profesor Miguel Martín, et al. [13], con un porcentaje de efectividad de las máquinas muy alto. Además los vectores de relajación de cada tejido (identificados por las MSVs) en el espacio de componentes principales se ubican en regiones distintas (5.13 y 5.15).

5.2. Conclusiones

Esta tesis permitió desarrollar procesos de segmentación de imágenes de resonancia magnética cerebrales potenciadas en T2, mediante la utilización de máquinas de soporte vectorial. En el preprocesamiento de los datos se utiliza ACP para eliminar errores aleatorios y sistemáticos presentes en las imágenes y se reescriben los datos con las tres primeras componentes principales que reúnen mas del 99 % de la información. Si persisten algunas curvas de relajación que no son decrecientes, se aplica una regresión exponencial.

Se caracterizan las curvas de T2 por los vectores aleatorios (K, V) y (K_4, V), donde K y K_4 representan respectivamente el valor de la primera y la cuarta imagen de un corte y V la velocidad de relajación definida por la diferencia de la primera y la séptima imagen. Las curvas de nivel de los histogramas conjuntos de las observaciones de diferentes cortes de distintos pacientes presentan un patrón similar de comportamiento y se pueden establecer regiones de $\mathbb{R}^2$

que caracterizan a los tejidos.

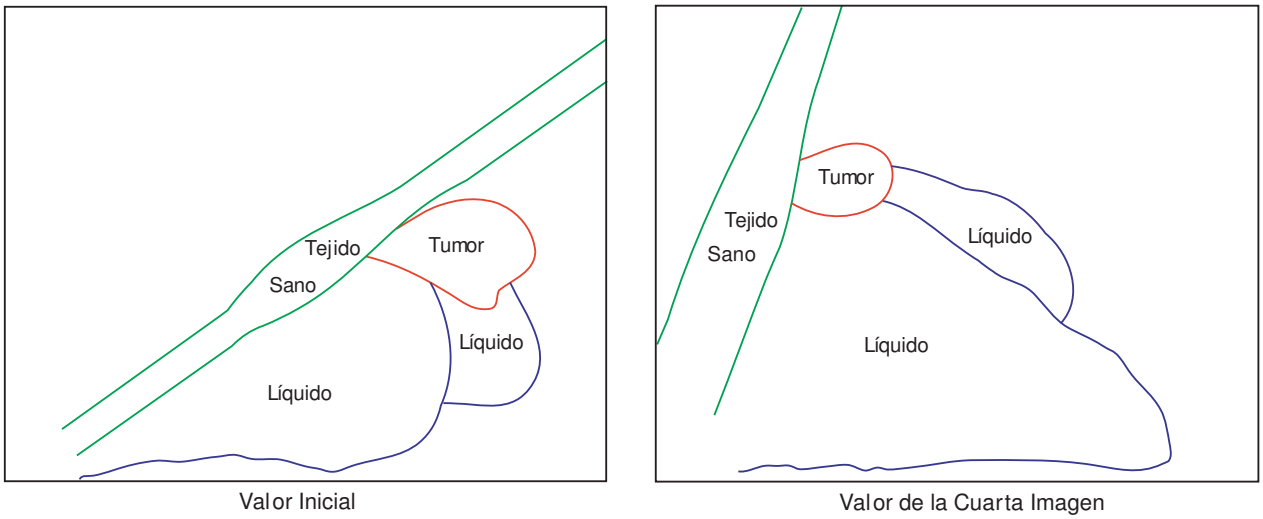


Figura 5.16: Regiones de los distintos tejidos en las CNHs

La presencia de una joroba en la zona del tejido enfermo de las CNHs indica la presencia del tumor pero no la ubicación(ver figura 5.16). Aunque en la imagen sea difícil detectar el tumor a simple vista, en las curvas de nivel se pueden presentar puntos que indiquen la existencia de un tumor.

Este patrón permite entrenar las MSVs sin la ayuda de un experto. A los vectores de T2 se les aplica un ACP para disminuir su dimensión de 8 a 3 ya que las 3 primeras componentes contienen mas del 99 % de la información. Se diseñan las máquinas para delimitar el tumor en las imágenes junto con el tejido sano y el líquido con los datos de entrenamiento escogidos en las regiones que indica el patrón establecido.

Los porcentajes de vectores bien clasificados dentro de la datos de validación de todas las MSVs construidas superan el 99.9 %; y las segmentaciones obtenidas son muy similares a las realizadas por el Dr. Miguel Martín, et al, descritas en un trabajo de su autoría [13], donde utiliza espectroscopía protónica en vivo y relaxometría. Además, el mapeo de las regiones de los distintos tejidos en las CNHs mediante las MSVs establecen claramente las zonas de cada tejido, donde se visualizan zonas de transición que comparten tejidos diferentes.

5.3. Recomendaciones

Entre las recomendaciones que surgen de este trabajo se pueden mencionar las siguientes:

- Las zonas de transición entre tejidos generada por las MSVs en las CNHs, caracterizadas por figuras solapadas de distintos colores, se podrían definir mas detalladamente tomando mas pequeño el tamaño de las celdas en la construcción del histograma de las variables (K, V) y (K_4, V), lo que permite representar menos datos por cada celda del histograma. Esto no se realizó debido a que implica imágenes mas pesadas y mas tiempo en su elaboración.

- Los resultados óptimos obtenidos de los dos últimos pacientes, donde en sus histogramas conjuntos se utiliza el valor de la cuarta imagen de los cortes sugieren construir todos los histogramas de esa manera y desarrollar la metodología presente en este trabajo para segmentar las imágenes de resonancia magnética y comparar los resultados con los obtenidos al utilizar la primera imagen.

-Con una gran cantidad de imágenes se podría dar una distribución conjunta mas precisa del vector (K_4, V), que daría la posibilidad en cada imagen de asignar probabilidad al tejido enfermo y afinar los límites entre tejidos en la figura 5.16.

Anexo

Zonas de Transición

En los resultados se muestran las regiones creadas por cada MSV en las CNHs correspondientes a los distintos tejidos. Una celda de un histograma puede contener varios pixeles que a su vez pueden ser clasificados en tejidos diferentes. Estas celdas representadas en las CNHs como puntos forman las zonas de transición entre tejidos; las cuales se caracterizan por puntos de colores y formas distintas, solapados entre si. Por sugerencia del profesor Miguel Martín, se identificó con un color cada zona de transición en las CNHs, con la finalidad de visualizarlas claramente. Además, las zonas de transición entre tejido sano y tumor; y de líquido y tumor se muestran en las segmentaciones de las imágenes estudiadas. A continuación se presentan los gráficos mencionados para cada paciente.

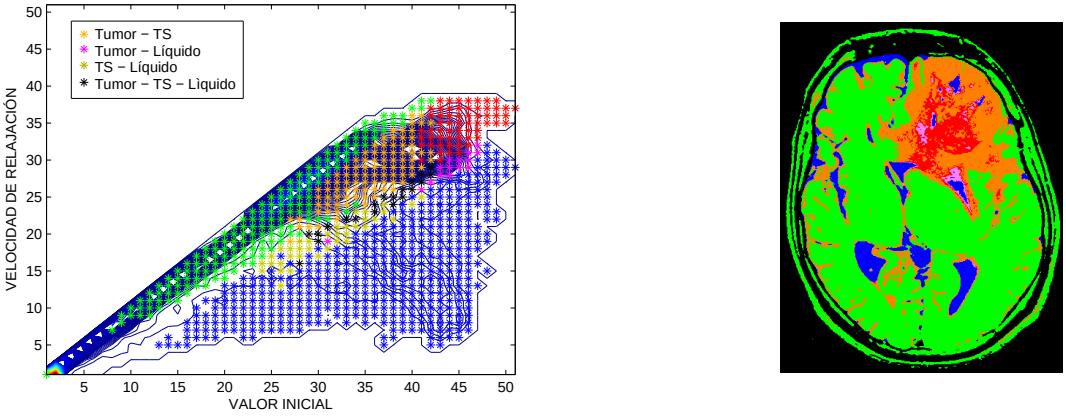
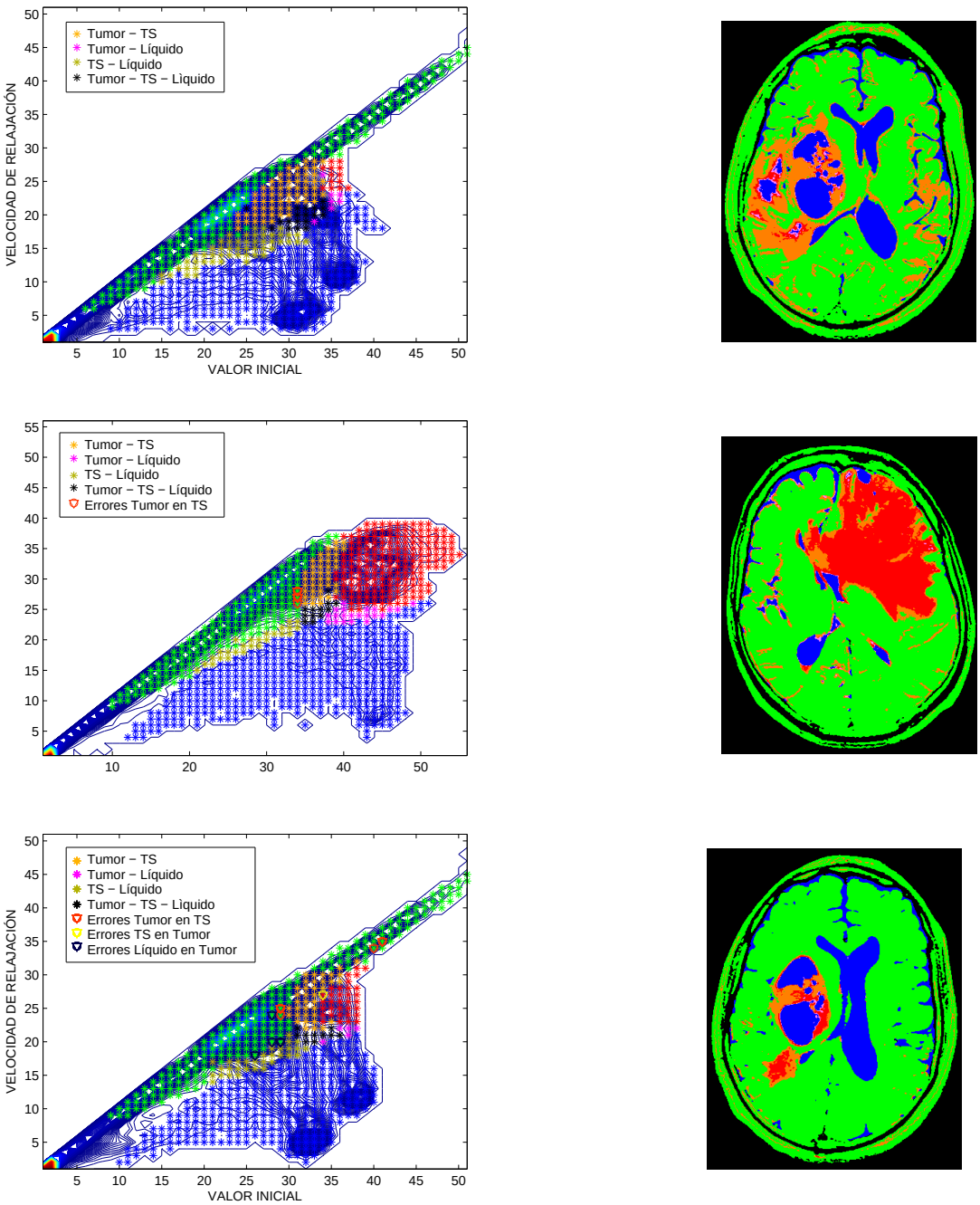
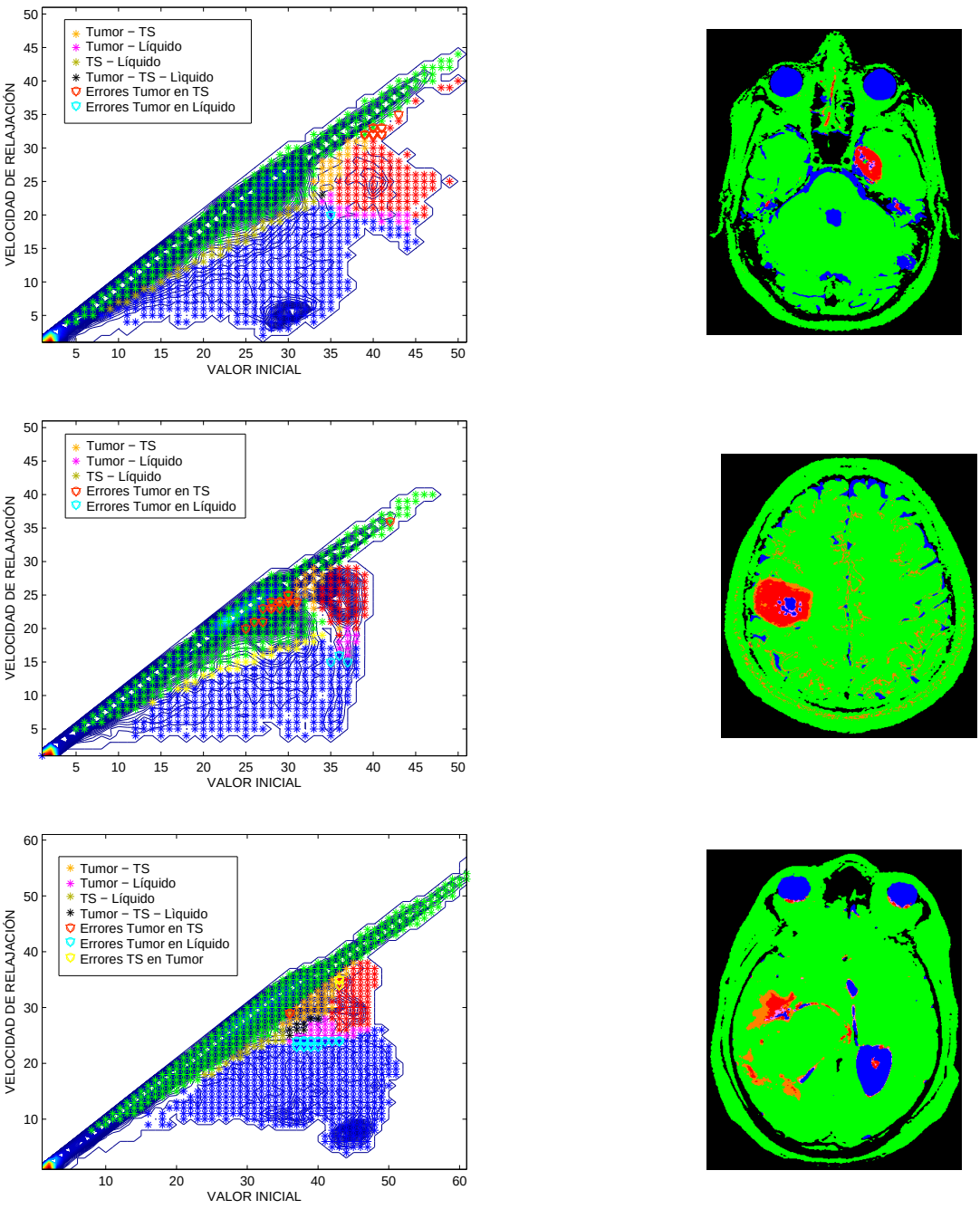


Figura 5.17: (a) Mapeo de las regiones obtenidas por la MSV en las CNH del corte 4 del paciente 1 (b) Segmentación del corte 4 del paciente 1, y del corte 5 del paciente 1 con regiones de transición entre tejidos .



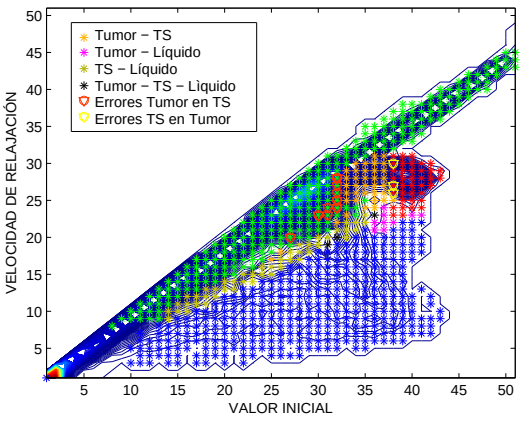
(a) (b)

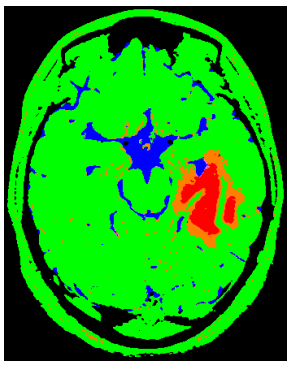
Figura 5.18: (a) Mapeo de las regiones obtenidas por las MSVs en las CNHs del corte 4 del paciente 2, y del corte 5 de los pacientes 1 y 2 (b) Segmentación del corte 4 del paciente 2, y del corte 5 de los pacientes 1 y 2 con regiones de transición entre tejidos .

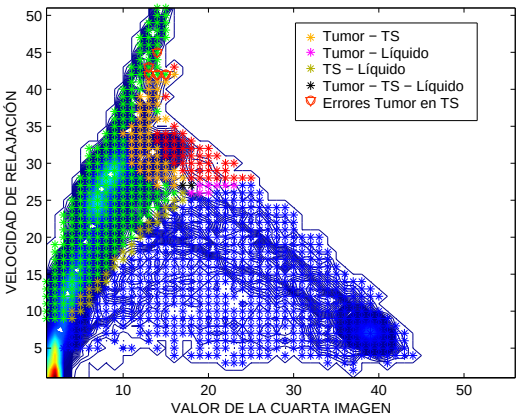


(a) (b)

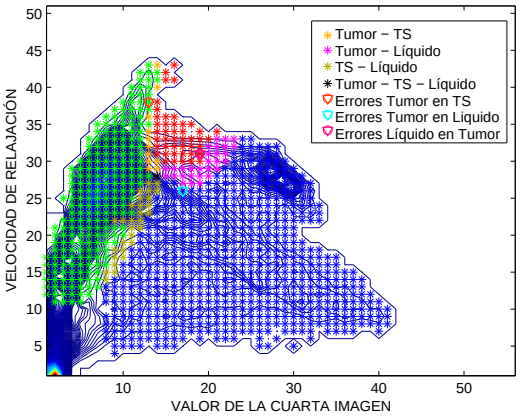
Figura 5.19: (a) Mapeo de las regiones obtenidas por las MSVs en las CNHs del corte 4 de los pacientes 3, 4 y 5 (b) Segmentaciones del corte 4 de los pacientes 3, 4 y 5 con regiones de transición entre tejidos.

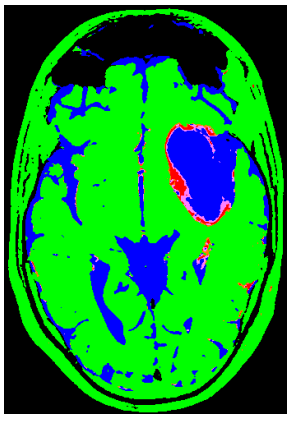






(a) (b)





(a) (b)

Figura 5.20: (a) Mapeo de las regiones obtenidas por las MSVs en las CNHs del corte 4 de los pacientes 6, 7 y 8 (b) Segmentaciones del corte 4 de los pacientes 6, 7 y 8 con regiones de transición entre tejidos.

Bibliografía

[1] Ayestaran, Rafael. Santamaría, Ignacio. Igualación Mediante Clasificadores Lineales y No Lineales de Máximo Margen.

[2] Barbola, R. Análisis Lineal y Teoría de Matrices, 1997.

[3] Clavijo, Jairo. Análisis de Componentes Principales. Universidad de Tolima, Ibague, Colombia.

[4] Coto, Ernesto. Métodos de Segmentación de Imágenes Médicas. Lecturas en ciencias de la computación. Laboratorio de computación gráfica. Escuela de computación. Facultad de ciencias. U.C.V. Marzo 2003.

[5] Cristianini, N. Shawe-Taylor, J. An Introduction to Support Vector Machines. Cambrige University Press. 2000.

[6] Drozdowicz, B. Bernasconi, Guillermo. et al. Segmentación Semiautomática de Imágenes de Resonancia Magnética Basada en Redes Neuronales Artificiales. Ciencia docencia y tecnología. Nro. 30. Año XVI. Marzo 2005. p.p. 117-155.

[7] Frieb, Thilo-Thomas. Cristianini, Nello. Campbell, Colin. The Kernel-Adatron Algorithm: Fast and Simple Learning Procedure for Support Vector Machines. Published in Proceeding of the Fifteenth International Conference on Machine Learning. July 24-27.1998.

[8] Garcia, Cristina. Moreno, José Ali. Kernel Based Method for Segmentation and Modeling of Magnetic Resonances Images.

[9] Hernández, José. Ramírez, José. Ferri César. Introducción a la minería de datos. Pearson Prentice Hall. 2004. Cap 14.

[10] Hofman, Kenneth. Algebra Lineal. Prentice Hall. Mexico. 1984.

[11] Le Saux, Bertrand. Amato, Giuseppe. Image Classifiers for Scene Analysis.

[12] Manjón, José Vicente. Segmentación Robusta de Imágenes de Resonancia Magnética Cerebral. Tesis Doctoral. Universidad Politécnica de Valencia. España. 2006. Capitulo I.

[13] Martin, Miguel. Mayobre, Finita. Bautista, Igor y Villata Raúl. Brain Tumor Evaluation and Segmentation by in Vivo Proton Spectroscopy and Relaxometry. Magma (New York, N.Y.), published in Germany. 2005-Dec. Vol 18. issue 6. p.p 316-330.

[14] Moreno, José Ali. Maquinas Lineales de Aprendizaje II. Seminario 2 del Postgrado de computación emergente de la facultad de ciencias e ingeniería UCV.

[15] Moreno, José Ali. Maquinas de Aprendizaje Lineales en Clasificación. Seminario del laboratorio de computación gráfica y geometría aplicada. Febrero. 2004. p.p 1-32.

[16] Reddick, W. Mulhern, R. y col. A Hybrid Neural Network Analysis of Subtle Brain Volume Differences in Children Surviving Brain Tumors. Magn Reson Imagin. May 1998. p.p. 413-421.

[17] Ruiz, Alejandra. Estimación de la Matriz de Trafico por el Método de las Componentes Principales. Tesis de pregrado presentada en la escuela de matemáticas de la U.C.V.

[18] Simons, Stephen. From Hahn-Banach to Monotonicity. Springer. Second Edition. 2008.

[19] Apostol, Tom. Calculus. Jon Wiley & Sons,Inc. 1967.

[20] Tang, H. Wu, Ex. et al. MRI Brain Image Segmentation by Multi-Resolution Edge Detection and Region Selection. Comput Med Imaging Graph. Nov-Dec 2004. p.p. 349-357.

[21] Vapnik, Vladimir. Statistical Learning Theory. A Wiley Interscience Publication, John Wiley SONS, INC. New York. 1998.

[22] Villardon, José Luis. Análisis de Componentes Principales.

[23] Zhang, J. Luang, M. y Chong, V. Tumor Segmentation from Magnetic Resonance Imaging by Learning via One-Class Support Machine. http://lear.inrialpes.fr/people/zhang/IWAIT04.pdf.