Nuevos modelos estadísticos para detección de patrones de hipo/perfusión-metabolismo en imágenes de tomografía funcional cerebral

Tesis Doctoral

Miriam López Pérez

Departamento de Teoría de la Señal,
Telemática y Comunicaciones
Universidad de Granada

Editor: Editorial de la Universidad de Granada
Autor: Miriam López Pérez
D.L.: GR 3473-2010
ISBN: 978-84-693-5227-4

D. Javier Ramírez Pérez de Inestrosa, Doctor por la Universidad de Granada y Profesor Titular del Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada y

D. Juan Manuel Górriz Sáez, Doctor por la Universidad de Cádiz y la Universidad de Granada y Profesor Titular del Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada,

MANIFIESTAN:

Que la presente Memoria titulada *"Nuevos modelos estadísticos para detección de patrones de hipo/perfusión-metabolismo en imágenes de tomografía funcional cerebral"*, presentada por Miriam López Pérez para optar al grado de Doctor Europeo por la Universidad de Granada, ha sido realizada bajo nuestra dirección. Con esta fecha, autorizamos la presentación de la misma.

Granada, a 25 de Mayo de 2010

Fdo: Javier Ramírez Pérez de Inestrosa Fdo: Juan Manuel Górriz Sáez

Memoria presentada por Miriam López Pérez para optar al Grado de Doctor Europeo por la Universidad de Granada.

Fdo: Miriam López Pérez

Título de Doctor con Mención Europea

Con el fin de obtener la Mención Europea en el Título de Doctor (aprobada en Junta de Gobierno de la Universidad de Granada el 5 de Febrero de 2001), se han cumplido, en lo que atañe a esta Tesis Doctoral y a su Defensa, los siguientes requisitos:

- Durante su etapa de formación, la doctoranda ha realizado una estancia de 3 meses fuera de España en una institución de enseñanza superior de otro Estado europeo cursando estudios o realizando trabajos de investigación que le han sido reconocidos por el órgano responsable del programa. En concreto la doctoranda ha realizado una estancia de investigación de 3 meses en el Institut für Biophysik de la Universidad de Regensburg, Alemania.

- Parte de la Tesis se ha redactado y presentado en una de las lenguas oficiales de la Unión Europea distinta a alguna de las lenguas oficiales en España. En concreto en Inglés.

- La Tesis ha sido informada por dos expertos pertenecientes a alguna institución de educación superior o instituto de investigación de un Estado miembro de la Unión Europea distinto de España.

- Un experto perteneciente a alguna institución de educación superior o instituto de investigación de un Estado miembro de la Unión Europea distinto de España, con el grado de doctor, y distinto de los mencionados antes, forma parte del tribunal evaluador de la Tesis.

Financiación

Las principales fuentes de financiación de esta Tesis Doctoral han sido las siguientes:

- Contrato asociado al proyecto "Detección de enfermedades neurológicas mediante clasificación y separación de señales" (PET2006-0253). Entidad financiadora: Ministerio de Educación y Ciencia. Entidades participantes: Universidad de Granada, Universidad de Cádiz, Universidad de Regensburg (Alemania) y Universidad de Aveiro (Portugal), Hospital Virgen de las Nieves de Granada y la empresa PTEC (Plataforma Tecnológica) Duración: desde 2007 hasta 2010.

- "Diagnóstico avanzado de enfermedades neurológicas mediante técnicas de reconstrucción y modelado de imágenes tomográficas cerebrales" (Proyecto de excelencia: TIC-02566). Entidad financiadora: Junta de Andalucía. Duración: desde 2008 hasta 2012.

- "SADIEN: Sistemas de ayuda al diagnóstico inteligente de enfermedades neurodegenerativas". Entidad financiadora: Universidad de Granada. Duración: desde enero 2008 hasta diciembre 2008.

- "Toolbox for Biomedical Signal and Image Processing based on Information Theory". Ref.: HD2008-0029. Entidad financiadora: Ministerio de Ciencia e Innovación. Entidades participantes: Universidad de Granada y Universidad de Regensburg. Duración: desde 2009 hasta 2010.

- "Nuevas Técnicas de Reconstrucción, Procesado, Clasificación y Fusión de Imágenes Médicas para Diagnóstico Precoz de la Enfermedad de Alzheimer" (TEC2008-02113/TEC). Entidad financiadora: Ministerio de Ciencia e Innovación. Duración: desde enero 2009 hasta diciembre 2012.

- "Plataforma abierta de procesamiento de imágenes para ayuda al diagnóstico de alteraciones neurológicoas (PAPI-ADAN)". (TIC-4530). Entidad financiadora: Junta de Andalucía. Duración: desde 2010 hasta 2014.

Agradecimientos

En primer lugar en los agradecimientos de esta Tesis quiero nombrar a Javier y Juanma, no sólo por su dedicación y profesionalidad como directores de la misma, sino también por el trato personal siempre amable que me han ofrecido durante este tiempo. Gracias a Javier por mantener siempre esa actitud positiva y motivadora que en muchos momentos a lo largo del desarrollo de este trabajo he agradecido tanto, y gracias a Juanma por su implicación y por transmitir su constancia en el trabajo y la sensación de que todo el esfuerzo merece la pena.

También me gustaría agradecer al resto de componentes del grupo SiPBA por su siempre buena disposición, y en particular a Ignacio y Diego por su ayuda en la recta final del trabajo. Gracias a todos por crear este ambiente tan amigable en el que el trabajo se convierte en una tarea gratificante.

Los resultados que se presentan en esta Tesis han sido producto de una colaboración multidisciplinar con expertos clínicos que han aportado material y conocimiento de gran utilidad para el desarrollo de este trabajo. Gracias a D. Manuel Gómez Río del Servicio de Medicina Nuclear del Hospital Virgen de las Nieves de Granada, así como a los expertos Francisco Moya y Eduardo Gil de la Clínica PET Cartuja (Sevilla). También quiero agradecer a D. Diego Pablo Ruiz Padillo su colaboración.

El agradecimiento más sincero a mis padres y hermanos, por su apoyo incondicional y por vivir cada paso de esta Tesis como suyo. Poder compartir todas mis inquietudes con vosotros ha sido para mí de gran ayuda.

No quisiera olvidarme de los compañeros de Ciencias por los buenos momentos que hemos compartido, y que de algún modo han contribuido también a que hoy esté escribiendo estas líneas.

Miriam López

Tren Regensburg-Nuremberg

9 de Mayo de 2010

Resumen

La presente Tesis Doctoral propone nuevos modelos estadísticos para la detección de patrones de hipo/perfusión-metabolismo en imágenes de tomografía funcional cerebral de tipo SPECT (Single Photon Emission Computed Tomography) y PET (Positron Emission Tomography), para la ayuda al diagnóstico precoz de enfermedades neurológicas como la enfermedad de Alzheimer. Tras la descripción de los procesos de normalización de las imágenes espacialmente y en intensidad, se presentan tres técnicas diferentes de extracción de características. En primer lugar se describe un método de exploración de imágenes que permite la localización automática de las regiones de interés para la clasificación. En segundo lugar se aplican técnicas de análisis de componentes principales y análisis discriminante lineal en combinación con el FDR (Fisher Discrimiant Ratio) como métodos de compresión de imagen y extracción de características, resolviendo el problema del pequeño tamaño muestral. Por último se llevan a cabo transformaciones basadas en la aplicación de kernels que permiten la extracción de dependencias no lineales presentes en las imágenes. Estas técnicas se combinan con métodos de clasificación supervisada basados en reglas bayesianas, máquinas de vectores de soporte y redes neuronales, alcanzando valores de precisión de hasta 95.6 % y 100 % para imágenes SPECT y PET y de 91.43 % para la reconocida base de datos ADNI, respectivamente. Estos resultados mejoran las tasas de acierto obtenidas mediante los métodos existentes hasta el momento para la detección precoz de la enfermedad de Alzheimer.

Índice general

Índice de figuras

Índice de tablas

Introducción

En 1906, Alzheimer describió el caso, considerado *único*, de una mujer de 51 años en la cual se desarrolló un proceso demencial progresivo, iniciado por una actitud delirante de celos y alteración mnémica gravísima (olvidaba el camino de regreso a su domicilio, cambiaba sin cesar los objetos de su sitio habitual pensando después que se los habían robado, etc.) alcanzando pronto un estado de perplejidad y desorientación temporo-espacial completa [Alzheimer, 1907]. Desde esta fecha hasta nuestros días, la enfermedad de Alzheimer ha adquirido un especial protagonismo a causa de la creciente frecuencia con que se presenta, especialmente en los países desarrollados con altos índices de longevidad, llegando a constituir en cierta forma una verdadera endemia demencial. Según la Fundación Alzheimer España, 8 millones de europeos están afectados por la enfermedad de Alzheimer, alcanzándose la cifra de 30 millones en todo el mundo. Teniendo en cuenta el envejecimiento de la población y el futuro incremento de personas mayores de 80 años, se prevé que el número de enfermos se duplique en 2020 y triplique en 2050. Es uno de los retos más importantes a los que deberá enfrentarse la sociedad en el transcurso del siglo XXI.

0.1. La enfermedad de Alzheimer

La Organización Mundial de la Salud (OMS) define la enfermedad de Alzheimer (EA) como una enfermedad degenerativa cerebral primaria, de etiología desconocida que presenta rasgos neuropatológicos y neuroquímicos característicos. El trastorno se inicia por lo general de manera insidiosa y lenta y evoluciona progresivamente durante un período de años. Considerada durante muchos años como una enfermedad rara, afecta al 5-7 % de las personas de más de sesenta y cinco años y al 80 % de los mayores de 85 años. Es la causa de invalidez, dependencia y mortalidad más frecuente en los mayores. Sus características clínicas y evolución están siendo cada día más conocidas, y los estudios histológicos han aclarado cuáles son las peculiaridades anatomopatológicas de este proceso. Diversos caminos de investigación están abriendo, igualmente, nuevas fronteras para su conocimiento genético, bioquímico y neuropatológico.

Sin embargo, a pesar de que la EA constituye una de las causas más frecuentes y graves de la demencia humana, existen todavía pocas posibilidades de utilizar una terapia eficaz, real, que detenga o ralentice este proceso patológico. La terapia basada en fármacos en muchas ocasiones y a muy diversos niveles, ayuda o atempera las situaciones de gravedad, agitación, depresión, ansiedad, insomnio, delirio, etc., facilitando un mejor comportamiento y adaptación del paciente.

Desde el punto de vista del diagnóstico, los esfuerzos llevados a cabo en investigación se centran en la detección precoz de la enfermedad, puesto que en esta fase la aplicación del tratamiento resulta más efectiva. Ante todo se requiere un diagnóstico clínico de probable EA utilizando criterios establecidos así como el NINCDS-ADRDA (de las siglas en inglés National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association) [McKhann et al., 1984] o el DSM-IV (cuarta edición de Manual diagnóstico y estadístico de los trastornos mentales) de la Asociación Psiquiátrica de los Estados Unidos [Americana de Psiquiatría, 1994] y subsiguiente confirmación mediante examen histológico del cerebro. Este último requerimiento supone que en la vida de los pacientes puede darse tan sólo un diagnóstico de EA probable. Además, dado que, por definición, la demencia supone "un deterioro global de la inteligencia y de las capacidades emocionales y conativas del individuo en un estado de conciencia inalterada" [Roth, 1981], es clara la dificultad para investigar la habilidad cognitiva de pacientes con un diagnóstico de probable EA. A pesar de las dificultades para explorar la capacidad cognitiva de pacientes

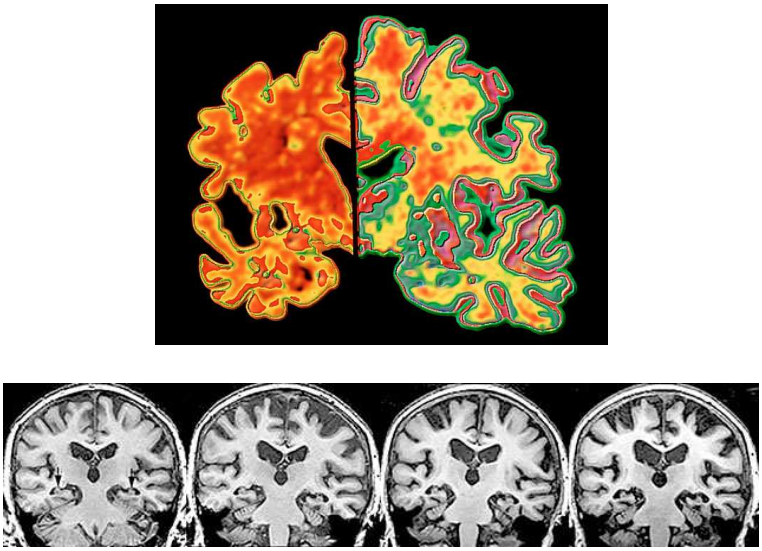


Figura 1: Arriba: Diferencias estructurales entre un cerebro afectado por la EA (izquierda) y uno sano (derecha). Abajo: cuatro imágenes por resonancia magnética que muestran, de izquierda a derecha, el aumento de la atrofia del cerebro (índice de pérdida de células) a medida que evoluciona la enfermedad. Las zonas oscuras corresponden al líquido cefalorraquídeo que ocupa el lugar de las neuronas muertas y, por tanto, es cada vez más voluminoso.

que sufren demencia con presumible EA, y los requisitos para un examen histopatológico del cerebro, ha sido posible caracterizar tres estadios en la progresión clínica de la enfermedad. En el primer estadio, la degeneración de la memoria para sucesos recientes y la afectación para nuevos aprendizajes constituyen normalmente las características más destacadas [McKhann et al., 1984; Morris and Kopelman, 1986]. En el segundo estadio, todas las modalidades de la memoria se deterioran progresivamente, hasta que, en el tercer estadio, todas las funciones intelectuales se hallan gravemente afectadas y los pacientes se mantienen en un estado casi vegetativo. Muchos de los déficits de memoria presentes en los casos iniciales de EA semejan a los existentes en los casos de amnesia del lóbulo temporal. En los estadios subsiguientes de la enfermedad, otras facetas del déficit intelectual se superponen a estos déficit de la memoria oscureciéndolos.

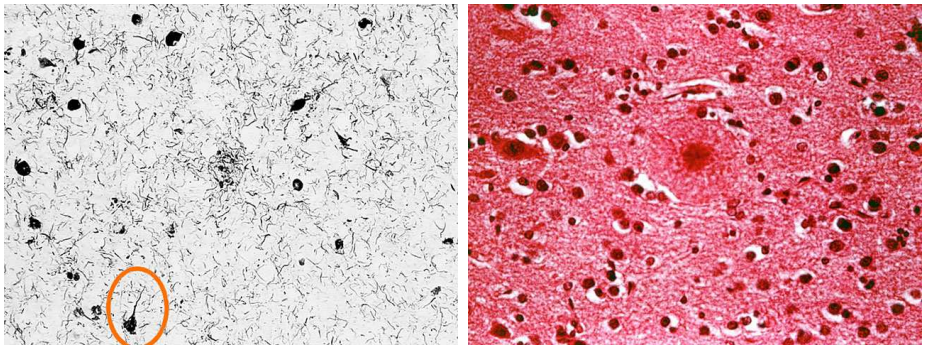


Figura 2: Las lesiones específicas de la enfermedad de Alzheimer sólo pueden observarse con microscopio. Izquierda: degeneración neurofibrilar que se caracteriza por la agregación anormal de proteínas muy específicas (proteínas Tau) en las neuronas. Derecha: placa amiloide en el centro. Está constituida por un depósito de péptido beta-amiloide, del que hasta la fecha no se conoce la función exacta.

0.1.1. Anatomía patológica

El análisis *post mortem* es el único elemento que permite efectuar un diagnóstico retrospectivo exacto de la enfermedad. Las principales alteraciones histológicas de las demencias de tipo Alzheimer son:

- Degeneración Neurofibrilar (DNF): Las lesiones de degeneración neurofibrilar son intraneurales. Se comprueba mediante microscopia fotónica, tras impregnación argéntica de los citoplasmas de las neuronas, bajo la forma de un haz de fibrillas anómalas que pueden tomar el aspecto de llamas o de bolas que deforman, a menudo, la neurona y vuelven el núcleo excéntrico. En microscopia electrónica, esta lesión corresponde a un enmarañamiento de filamentos de 10 nm de espesor, organizados por pares, según una configuración helicoidal con un paso de 80 nm.

- Degeneración Granulovacuolar (DGV). Se trata de vacuolas intracitoplasmáticas de 4-5 μm de diámetro, que contienen un gránulo central eosinófilo, observadas casi exclusivamente en las neuronas piramidales del hipocampo, donde se asocian en abundancia a las DNF.

- Placas seniles (PS). Estas estructuras redondeadas situadas exclusivamente en las zonas corticales y con predominio en la segunda y la tercera capas se componen de un centro amiloide, que se pone perfectamente de manifiesto mediante el rojo congo, rodeado de una corona

de prolongaciones nerviosas en degeneración. Con el microscopio electrónico se puede observar que el centro amiloide está constituido por un enmarañamiento de filamentos rectilíneos con una configuración beta plegada de 7 a 10 nm de diámetro, que forman la sustancia amiloide (sustancia proteica). Por regla general se asocian a una importante densidad de DNF. Los estudios demuestran que la PS no está constituida por materia inerte sino que presenta actividades metabólicas múltiples para diversos sistemas de neuromediadores.

En la Figura 2 se muestra el aspecto de las PS y del DNF. Estas lesiones elementales predominan en las zonas corticales y en las áreas asociativas del cerebro, en particular las regiones posteriores (encrucijada parietotemporooccipital y la parte interior y externa de los lóbulos temporales). Se acompañan de una despoblación neuronal con atrofia cortical. La alteración de ciertos núcleos grises centrales, como el núcleo basal de Meynert, indica, sin embargo, una participación subcortical abiotrófica del proceso, cuya importancia sigue en la actualidad en discusión. En definitiva, las características anatómicas de la enfermedad no se basan tanto en la naturaleza misma de las lesiones, cuya especificidad es insignificante (la DNF y la PS se encuentran, asimismo, en la trisomía 21, en el Parkison postencefalítico, en la demencia de los boxeadores y, sobre todo, en individuos de edad avanzada sin demencia), sino especialmente en el número y en la difusión a todas la capas corticales de estas alteraciones elementales. Además, aunque las anomalías histológicas son cualitativamente idénticas en las formas preseniles y seniles de la enfermedad, la topografía de la atrofia y la intensidad de las alteraciones diferirían notablemente en estas dos formas.

0.2. Diagnóstico de la EA: Estado del arte

En la actualidad el diagnóstico de la EA representa todavía un reto, especialmente en su etapa temprana que es cuando hay más oportunidades de tratar sus síntomas, así como de probar y desarrollar nuevos tratamientos. En la práctica clínica, el diagnóstico se basa en cuidadosos análisis clínicos, entrevistas con el paciente y sus familiares, y tests neuropsicológicos. A menudo la medicina nuclear se utiliza como apoyo al diagnóstico mediante imágenes funcionales del cerebro, aunque la enfermedad, tal y como fue apuntado en la sección anterior, solo podrá ser confirmada tras la autopsia.

El descubrimiento de nuevos fármacos efectivos para el tratamiento de los

síntomas de la enfermedad junto con otros agentes que están siendo clínicamente evaluados, abriría una nueva esperanza en el tratamiento de la EA. Existe un claro consenso en la necesidad de desarrollar técnicas de diagnóstico precoz más efectivas para una rápida intervención, para prevenir o disminuir la progresión de la enfermedad y para que se pueda hacer el máximo aprovechamiento de los servicios y tratamientos disponibles para los pacientes que la sufren.

Varias líneas convergentes de investigación sugieren que el proceso neurodegenerativo asociado con la demencia comienza varios años antes de que las características clínicas puedan ser detectadas con los instrumentos actuales. La duración precisa de este periodo pre-clínico y los detalles de los procesos moleculares que se generan son todavía desconocidos. Esta incertidumbre sobre las fases tempranas sin síntomas de la enfermedad reside en la ausencia de técnicas validadas y precisas para el diagnóstico.

El uso de imágenes médicas como herramienta de ayuda al diagnóstico de enfermedades neurológicas como la EA ha sido extensamente discutida. Históricamente, el patrón que se ha asociado a la EA se ha definido mediante la inspección de las áreas corticales, que son más fácilmente identificables en las imágenes funcionales. Basándose en la inspección visual de anormalidades en estas áreas, las imágenes son clasificadas como indicativas de EA cuando se encuentra: *i*) Bajos niveles de captación en regiones locales del córtex, en los lóbulos parietal, temporal y/o frontal, o *ii*) Un nivel global reducido de captación, relativo al córtex sensorimotor y visual, tálamo, el ganglio basal y el cerebelo [Hoffman and Phelps, 1986]. Usando estos criterios, el diagnóstico visual tiene un alto poder predictivo para detectar la presencia de enfermedades neurodegenerativas, y produce una precisión en el diagnóstico mayor que los criterios clínicos (77 %) [Silverman et al., 2003, 2001]. Sin embargo, la precisión con que actualmente se realiza el diagnóstico precoz de este tipo de enfermedades neurodegenerativas no supera el 80 % [Cummings et al., 1998; Carr et al., 1997].

Las imágenes SPECT muestran las alteraciones en el flujo sanguíneo mientras que las imágenes PET muestran los niveles de consumo de glucosa de las diferentes regiones cerebrales. Los patrones de anormalidad son similares para ambas técnicas, aunque PET tiene la ventaja de que los escáneres tridimensionales de esta modalidad tienen más sensibilidad que los escáneres SPECT de alta resolución (la resolución disponible en la actualidad está ahora en el rango 3-5 mm para PET y 7-8 mm para SPECT). Una alta resolución es deseable para estudios de imágenes cerebrales dedicados a pequeñas estructuras como el hipocampo, que es considerado como una región afectada

por la EA, y por lo tanto existe una cierta ventaja de las técnicas PET sobre las SPECT [Messa et al., 1994; Herholz et al., 2002; Silverman, 2004].

Nuevas técnicas basadas en análisis de regiones de interés o ROIs (del inglés, "Regions of Interest") diferentes al neocortex, han detectado patrones característicos de baja captación en regiones como el lóbulo temporo-parietal y cíngulo posterior o el lóbulo temporal medio [Santi et al., 2001; Leon et al., 2001; Nestor et al., 2003]. A menudo estos estudios dependen de un co-registro intra-sujeto de imágenes MRI/PET, dando lugar a una especificidad anatómica mayor que otras técnicas, pero requiere, por lo general, el alineado manual de áreas específicas y la consiguiente dependencia en el operador y en largos tiempos de manejo. El análisis de ROIs está supeditado a la necesidad de asumir *a priori* las regiones que podrían mostrar un efecto particular de interés.

Una metodología diferente de análisis de imágenes consiste en el uso del ordenador para manejar la información a nivel de voxel, conocidos como sistemas CAD (del inglés Computer Aided Diagnosis). Ejemplos comúnmente utilizados son SPM [Friston et al., 2007] o NEUROSTAT & 3D-SSP [Minoshima et al., 1995], programas que permiten hacer comparaciones estadísticas voxel a voxel, creando mapas paramétricos de efectos significativos. Estas técnicas univariadas permiten el examen global del cerebro a nivel de vóxel, y como consecuencia no están sujetas a ninguna hipótesis regional previa.

Sin embargo, la información importante para el diagnóstico tiene un carácter regional, por lo que varias técnicas multivariadas han sido desarrolladas para obtener datos a analizar con otras herramientas estadísticas, como SPM, NEUROSTAT & 3D-SSP, ANOVA o MANCOVA [Drzezga et al., 2003; Ishii et al., 2006; Scarmeas et al., 2004; Teipel et al., 2007; Salmon et al., 2009; Markiewicz et al., 2009; Nobili et al., 2008; Chen et al., 2009]. El análisis de componentes principales (PCA de sus siglas en inglés Principal Component Analysis) o el análisis de componentes independientes (Independent Component Analysis o ICA) son ejemplos de análisis multivariados que han sido usados para identificar patrones prominentes de correlación o conectividad funcional de regiones cerebrales [Friston et al., 2007; López et al., 2009c; Álvarez et al., 2009a]. Se ha argumentado que estas técnicas tienen la limitación de imponer ligaduras biológicamente inverosímiles, a la vez que su carácter lineal es limitado para una descomposición en conjuntos de medidas neuropsicológicas.

En la actualidad es necesario introducir nuevas técnicas más eficientes de procesado, modelado y clasificación de imágenes, así como mejorar el

rendimiento de las técnicas existentes. Este trabajo propone alternativas para conseguirlo basándose en la *teoría del aprendizaje estadístico por computador*. Los métodos de clasificación estadística, a pesar de haber producido notables resultados para el reconocimiento de patrones, no se han popularizado aún en este área. Esto puede deberse al hecho de que las imágenes representan grandes cantidades de datos, y la mayoría de los estudios cuentan con un número reducido de imágenes (generalmente <100) [Ishii et al., 2006; Stoeckel et al., 2001, 2004].

0.3. Objetivos

Los objetivos de este trabajo se dirigen al desarrollo de un sistema de ayuda al diagnóstico computarizado para la detección automática de la EA en su fase temprana. Para ello, es necesario implementar algoritmos en diferentes fases, que comprenden desde la normalización y preprocesado de las imágenes, hasta la decisión final del sistema que emite un voto sobre el estado particular de un paciente. Con este sistema se pretende:

- Anticipar la detección de enfermedades neurológicas como la EA en su etapa temprana. Con este fin se desarrollan algoritmos que manejen de forma eficiente imágenes neurológicas funcionales, ya que en éstas se aprecian los primeros síntomas de la EA frente a las imágenes estructurales que presentan cambios en etapas más avanzadas de la enfermedad.

- Emitir un voto objetivo sobre el diagnóstico particular de un paciente reduciendo la variabilidad o diferencia de criterio entre diferentes observadores. El aprendizaje del sistema está basado en el etiquetado previo de un conjunto de imágenes llevado a cabo por expertos clínicos. El fin del sistema por tanto es reproducir con la máxima precisión posible este etiquetado mediante técnicas de aprendizaje supervisado.

- Disminuir el tiempo de diagnóstico mediante la implementación de algoritmos eficientes que superen las limitaciones de la alta dimensionalidad de las imágenes neurológicas, y aumentar la precisión de los procesos que actualmente se llevan a cabo manualmente, como la reorientación y normalización de las imágenes.

0.4. Resumen de capítulos

El contenido de este trabajo se ha organizado en tres partes del siguiente modo:

- En la primera parte se describe el tipo de imagen empleada actualmente en el diagnóstico de enfermedades neurológicas como la EA, desde el proceso de obtención de dichas imágenes hasta el preprocesado, y se describen las bases de datos empleadas en este trabajo (Capítulo 1). Posteriormente, los criterios de diagnóstico de la EA y las técnicas computarizadas desarrolladas hasta el momento con el fin de obtener un diagnóstico asistido por computador se revisan en el Capítulo 2. En el Capítulo 3 se resumen las bases del aprendizaje estadístico supervisado, y una descripción más detallada de los métodos de clasificación empleados en este trabajo se presenta en el Capítulo 4.

- En la segunda parte del trabajo se exponen las técnicas empleadas en este trabajo para la detección automática de patrones de hipoperfusión e hipometabolismo. En el Capítulo 5 se presenta una técnica de clasificación basada en búsqueda de regiones de interés y agregado de votos como versión mejorada de la técnica VAF tomada como referencia. En los Capítulos 6 y 7 se proponen transformaciones lineales como métodos de compresión y extracción de características para la clasificación de las imágenes, y en el Capítulo 8 se generalizan estas transformaciones lineales mediante la aplicación de transformaciones basadas en kernels. Finalmente, en el Capítulo 9 se discuten las conclusiones extraídas del trabajo desarrollado.

- La tercera y última parte consiste en un resumen en inglés del trabajo presentado.

0.5. Publicaciones

Parte del trabajo realizado en esta tesis ha derivado en las siguientes publicaciones en revistas indexadas por el ISI:

Primer autor

1. **M. López**, J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, C. G. Puntonet. Automatic tool for the Alzheimer's disease diagnosis using PCA and Bayesian classification rules. (2009) IET Electronics Letters, vol 45, no. 8, pp. 389-391.

2. **M. M. López**, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, R. Chaves. SVM-based CAD system for early detection of the Alzheimer's disease using kernel PCA and LDA. (2009) Neuroscience Letters, vol 464, Issue 3, pp. 233-238.

3. **M. López**, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, R. Chaves, P. Padilla, M. Gómez-Río, the Alzheimer's Disease Neuroimaging Initiative. Principal Component Analysis-based techniques and supervised classification schemes for the early detection of the Alzheimer's Disease. (2010) Neurocomputing, in press.

Otros

1. R. Chaves, J. Ramírez, J. M. Górriz, **M. López**, D. Salas-Gonzalez, I. Álvarez and F. Segovia. SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. (2009) Neuroscience Letters, vol 461, Issue 3, pp. 293-297.

2. I. Alvarez Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, **M. López**, C. G. Puntonet, F. Segovia. Alzheimer's Diagnosis Using Eigenbrains and Support Vector Machines. (2009) IET Electronics Letters, vol 45, no. 7, 342-343.

3. D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, **M. López**, I. A. Illán, C. G. Puntonet, M. Gómez-Río. Analysis of SPECT brain images for the diagnosis of Alzheimer's disease using moments and support vector machines. (2009) Neuroscience Letters, vol 461, Issue 1, pp. 60-64.

4. J. Ramírez, J. M. Górriz, R. Chaves, D. Salas-Gonzalez, I. Álvarez, **M. López** and F. Segovia. SPECT Image Classification Using Random Forests. (2009) IET Electronics Letters, vol 45, Issue 12, pp. 604-605.

5. D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, **M. López**, I. Álvarez, F. Segovia, R. Chaves and C. G. Puntonet. Computer-aided diagnosis

of Alzheimer's disease using support vector machines and classification trees. (2010) Physics in Medicine and Biology, vol 55, no. 10, pp. 2807-2817.

6. F. Segovia, J. M. Górriz, J. Ramírez, D. Salas-González, I. Álvarez, **M. López**, R. Chaves and P. Padilla. Classification of functional brain images using a GMM-based multi-variate approach. (2010) Neuroscience Letters, vol 474, Issue 1, pp. 58-62.

7. J. Ramírez, J. M. Górriz, F. Segovia, R. Chaves, D. Salas-Gonzalez, **M. López**, I. Álvarez and P. Padilla. Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. (2010) Neuroscience Letters, vol 472, Issue 2, pp. 99-103.

8. J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, **M. López**, I. Álvarez, M. Gómez-Río, Computer Aided Diagnosis of Alzheimer Type Dementia Combining Support Vector Machines and Discriminant Set of Features. Information Sciences, in press.

9. I. Álvarez Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, **M. López**, F. Segovia, P. Padilla and C. G. Puntonet. Projecting independent components of SPECT images for computer aided diagnosis of Alzheimer's disease. (2010) Pattern Recognition Letters, in press.

Asimismo ha producido las siguientes aportaciones a congresos internacionales:

1. I. Álvarez, **M. López**, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, C. G. Puntonet and F. Segovia. Automatic Classification System for the Diagnosis of Alzheimer Disease Using Component-Based SVM Aggregations. Proceedings of the 15th International Conference on Neural Information Processing (ICONIP 2008), Lecture Notes in Computer Science, vol 5507 402-409. Auckland, New Zealand. November 2008.

2. J. Ramírez, J. M. Górriz, **M. López**, D. Salas-Gonzalez, I. Álvarez, F. Segovia and C. G. Puntonet. Early Detection of the Alzheimer Disease Combining Feature Selection and Kernel Machines. Proceedings of the 15th International Conference on Neural Information Processing (ICONIP 2008), Lecture Notes in Computer Science, vol 5507 410-417. Auckland, New Zealand. November 2008.

3. **M. López**, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia and C. G. Puntonet. Computer Aided Diagnosis of Alzheimer's Disease Using Principal Component Analysis and Bayesian Classifiers. The Sixth International Symposium on Neural Networks (ISNN 2009), Advances in Soft Computing, vol 56, pp. 213-221. Wuhan, China. May 2009.

4. I. Álvarez, J M. Górriz, J. Ramírez, D. Salas-Gonzalez, **M. López**, C. G. Puntonet and F. Segovia. Independent Component Analysis of SPECT Images to Assist the Alzheimer's Disease Diagnosis. The Sixth International Symposium on Neural Networks (ISNN 2009), Advances in Soft Computing, vol 56, pp. 411-419. Wuhan, China. May 2009.

5. D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, I. Álvarez, **M. López**, F. Segovia and C. G. Puntonet. Selecting Regions of Interest for the Diagnosis of Alzheimer Using Brain SPECT Images. The Sixth International Symposium on Neural Networks (ISNN 2009), Lecture Notes in Computer Science, vol 5553, pp. 399-406. Wuhan, China. May 2009.

6. **M. López**, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia and C. G. Puntonet. Automatic System for Alzheimer's Disease Diagnosis Using Eigenbrains and Bayesian Classification Rules. Proceedings of the 10th International Work-conference on Artificial Neural Networks (IWANN 2009), Lecture Notes in Computer Science, vol 5517, pp. 949-956. Salamanca, Spain. June 2009.

7. J. Ramírez, R. Chaves, J. M. Górriz, **M. López**, D. Salas-Gonzalez, I. Álvarez and F. Segovia. SPECT Image Classification Techniques for Computer Aided Diagnosis of the Alzheimer Disease. Proceedings of the 10th International Work-conference on Artificial Neural Networks (IWANN 2009), Lecture Notes in Computer Science, vol 5517, pp. 941-948. Salamanca, Spain. June 2009.

8. I. Álvarez, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, **M. López**, F. Segovia, C. G. Puntonet and B. Prieto. Alzheimer's Diagnosis Using Eigenbrains and Support Vector Machines. Proceedings of the 10th International Work-conference on Artificial Neural Networks (IWANN 2009), Lecture Notes in Computer Science, vol 5517, pp. 973-980. Salamanca, Spain. June 2009.

9. **M. López**, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia and M. Gómez-Río. Support Vector Machines and Neural Networks for the Alzheimer's Disease Diagnosis Using PCA. Pro-

ceedings of the 3rd. International Work-Conference on the interplay between natural and artificial computation (IWINAC 2009), Lecture Notes in Computer Science, vol 5602, pp. 142-149. Santiago de Compostela, Spain. June 22-26, 2009.

10. **M. Lopez**, J. Ramirez, J. M. Gorriz, D. Salas-Gonzalez, I. Alvarez, F. Segovia and R. Chaves. Neurological image classification for the Alzheimer's Disease diagnosis using Kernel PCA and Support Vector Machines. Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, pp. 2486-2489. Orlando, Fl. Oct. 24 2009-Nov. 1 2009.

11. **M. Lopez**, J. Ramirez, J. M. Gorriz, D. Salas-Gonzalez, I. Alvarez, F. Segovia and R. Chaves. Multivariate approaches for Alzheimer's disease diagnosis using Bayesian classifiers. Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, pp. 3190-3193. Orlando, Fl. Oct. 24 2009-Nov. 1 2009.

Finalmente, parte de este trabajo ha sido aceptado para publicación en el siguiente capítulo de libro:

1. Nombre del capítulo: Functional image classification techniques for early Alzheimer's disease detection. Autores: I. A. Illán; **M. López**; J. M. Górriz; J. Ramírez; F. Segovia; D. Salas-Gonzalez; R. Chaves; C. G. Puntonet. Título del libro: Recent Advances in Biomedical Signal Processing. Editorial: Bentham Science Publishers Ltd. Año: 2010.

Parte I

Fundamentos Teóricos

CAPÍTULO 1

Tomografía Computarizada

La capacidad de predicción de la medicina nuclear con respecto a la EA y otras demencias ha quedado patente en un un número elevado de estudios [Hoffman et al., 2000; Silverman et al., 2001; Higdon et al., 2004]. Las tomografías de emisión computarizadas o ECT (de sus siglas en inglés Emission-computed tomography) se han empleado en la investigación biomédica y en la práctica clínica durante las últimas tres décadas. Dos ejemplos de ello son las técnicas SPECT y PET, que difieren de otras técnicas de imagen médicas como las resonancias magnéticas en que producen mapas de las funciones fisiológicas, en lugar de formar imágenes de estructuras anatómicas. Estas imágenes tomográficas radiofarmacéuticas proporcionan mapas 3D *in vivo* de fármacos etiquetados mediante isotopos radioactivos, que son inyectados al paciente normalmente por vía intravenosa. Habitualmente, estas imágenes son valoradas por expertos, quienes evalúan visualmente la presencia de patrones cerebrales característicos compatibles con esta enfermedad. En este capítulo se proponen métodos de reconstrucción, procesado automáticos para la normalización de las imágenes que permita un tratamiento automático de las mismas.

1.1. Imágenes funcionales

1.1.1. SPECT

La tomografía por emisión de fotón único (SPECT) es una técnica de imagen ECT que fue inicialmente desarrollada en la década de 1960, pero su uso no se extendió a la práctica clínica hasta los 80. Se usa principalmente cuando la información estructural no es suficiente para detectar o monitorizar un problema funcional. Por tanto, SPECT es una modalidad no invasiva tridimensional que proporciona información clínica relacionada con los procesos psicológicos y bioquímicos que se producen en los pacientes. Cuando el núcleo de un radioisótopo se desintegra, se emite un fotón gamma en una dirección aleatoria y uniformemente distribuida en la esfera que rodea al núcleo. Si el fotón no sufre colisiones con electrones u otras partículas en el interior del cuerpo, su trayectoria será una línea recta o rayo. Para discriminar la dirección de incidencia usando un detector de fotones externo al paciente, se requiere una colisión física. Típicamente se coloca un colisionador antes del detector de tal modo que los fotones procedentes de todas las direcciones excepto de una quedan bloqueados por las placas. Esto garantiza que sólo los fotones procedentes de la dirección deseada golpearán el detector de fotones. SPECT es esencial para visualizar tanto agentes de flujo de sangre en regiones cerebrales (rCBF) como receptores del cerebro, así como para obtener información de la perfusión del miocardio. El SPECT cerebral se ha convertido en una herramienta de investigación y diagnóstico muy importante en la medicina nuclear.

En el proceso de adquisición de imágenes SPECT el paciente se posiciona cómodamente en una camilla con la cabeza "inmovilizada"(Figura 1.1). El detector debe posicionarse tan próximo al cerebro del paciente como sea posible, preferiblemente con un radio de rotación de 14 cm o menos desde la superficie del detector de colisiones al centro del cerebro del paciente. Al paciente se le inyecta un fármaco emisor de rayos gamma y el scan SPECT se adquiere gracias a una cámara de tres cabezales Picker Prism 3000. Se toman un total de 180 proyecciones por cada paciente con una resolución angular de 2 grados. Los trazadores 99mTc-ECD y 99mTc-HMPAO se emplean con frecuencia para estudiar la perfusión cerebral mediante SPECT.

Tras la adquisición de las imágenes, es necesario el filtrado y reconstrucción de las mismas (ver sección 1.2). Los métodos de reconstrucción necesitan la adquisición de distintas proyecciones tomadas desde diferentes vistas del paciente. Estos datos permiten generar imágenes que muestran la distribu-

Figura 1.1: Gamma Cámara Picker Prism 3000.

ción en tres dimensiones del radiofármaco, obteniéndose un mapa de activación que representa el rCBF. Esto hace que esta técnica sea aplicable al diagnóstico de enfermedades neurológicas como por ejemplo la enfermedad de Alzheimer, donde regiones específicas del cerebro sufren una disminución en el rCBF que puede ser efectivamente medida a través de técnicas de imágenes cerebrales SPECT. [Hellman et al., 1989; Holman et al., 1992; Johnson et al., 1993; Stoeckel et al., 2001, 2004; Fung and Stoeckel, 2007]. La figura 1.2 muestra cortes transversales de un SPECT con 99mTc-HMPAO en el que se observa claramente disminución en la perfusión cerebral.

1.1.2. PET

La tomografía por emisión de positrones (PET) es, al igual que SPECT, una técnica de imagen médica funcional ya que proporciona información sobre el funcionamiento o metabolismo de diferentes sistemas biológicos, siendo complementaria a las técnicas morfológicas. En la actualidad es una técnica de gran interés en el estudio del metabolismo cerebral en diversas patologías, como tumores cerebrales, epilepsias y en las demencias como la EA. El fundamento de la PET consiste en marcar con un átomo radiactivo (radioisótopo)

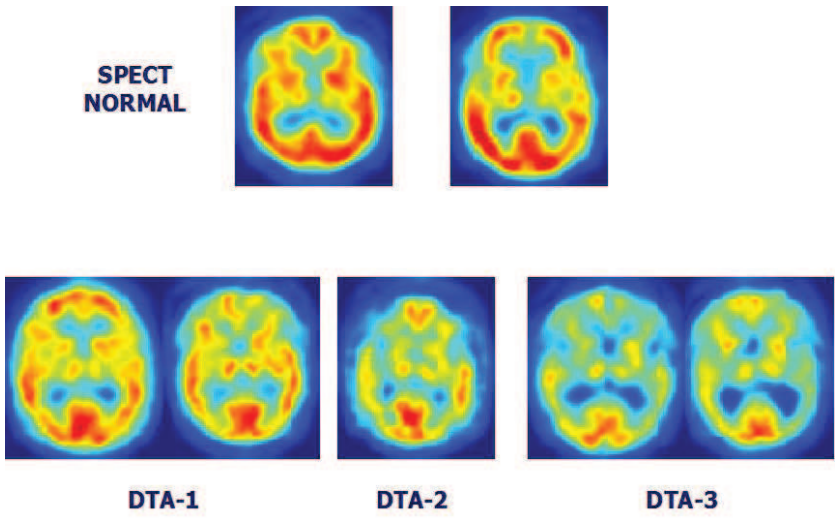


Figura 1.2: Patrones de perfusión cerebral regional con SPECT en sujetos sanos y pacientes con demencia de tipo Alzheimer.

una molécula biológica (trazador), cuyo comportamiento deseamos seguir o "trazar". De este modo, en función del trazador utilizado, es posible estudiar distintos procesos biológicos. Por ejemplo el isótopo ^{15}O se emplea en combinación con H_2O como trazador para estudiar la perfusión cerebral, o el ^{11}C con flumazenil para examinar los receptores cerebrales. Sin embargo, la PET tiene su aplicación clínica más importante en el estudio del metabolismo, en cuyo caso el radioisótopo más utilizado es el Flúor 18 (^{18}F) y el trazador, un análogo de la glucosa llamado flúor-desoxi-glucosa (FDG). Una vez introducida en el flujo sanguíneo, la FDG es almacenada por la célula, pero a diferencia de la glucosa normal, no puede ser utilizada para la producción de energía y queda atrapada dentro de ella. De este modo, la captación de FDG en las células es proporcional a su nivel metabólico. En la EA, regiones características muestran un decrecimiento en el metabolismo de glucosa.

Al tiempo que la FDG es atrapada en la célula, el radioisótopo de ^{18}F, como es inestable, sufre una desintegración nuclear que da lugar a la liberación de un positrón que se aniquila rápidamente al recombinarse con un electrón cercano, emitiendo dos fotones de 512 KeV que viajan en sentidos opuestos. Para detectar la emisión de estos fotones, se utilizan las denominadas cámaras PET, que consisten en una serie de anillos de fotodetectores de centelleo que rodean la cabeza del paciente. Éstos convierten los fotones emitidos en luz visible y están, a su vez, rodeados de una serie de tubos fotomultiplicadores, los cuales, cuando dos fotones son detectados al mis-

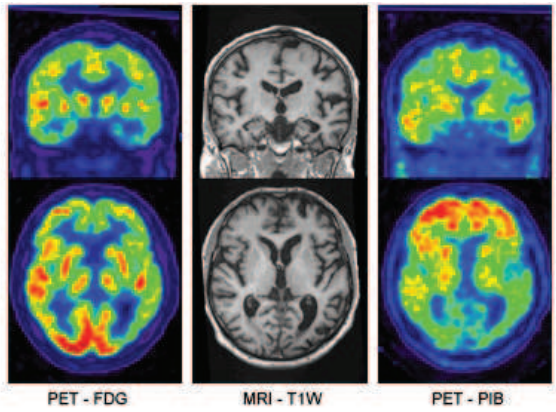


Figura 1.3: Imágenes PET (^{18}F-FDG), de resonancia magnética y PET (^{11}C-PIB) para un mismo paciente afectado por la enfermedad de Alzheimer.

mo tiempo a 180° en el anillo, son capaces de determinar la posición de aniquilación del positrón. A partir de estas coincidencias en el anillo, se es capaz de reconstruir un volumen de imágenes formadas por vóxels (volume elements) que muestran el metabolismo de la zona bajo estudio. Las dosis de radioisótopos utilizadas en este tipo de estudios son extremadamente bajas y no provocan efectos nocivos para el paciente, hecho que ha sido verificado mediante el seguimiento de cientos de miles de pacientes durante los últimos veinte años. Sin embargo, el coste de las cámaras PET es considerablemente elevado, lo que redunda en el precio de las exploraciones. Aún así, lo que más encarece a éstas es la necesidad de disponer de un ciclotrón cercano para la síntesis de los radioisótopos, ya que éstos se desintegran rápidamente dada su corta vida media.

1.1.3. Patrón de perfusión de la EA

Existe un gran número de estudios relacionados con los patrones típicos de perfusión para detectar la EA, y para distinguir entre pacientes con la EA y pacientes normales o entre pacientes con la EA y pacientes que sufren otro tipo de demencia. A pesar de la cantidad de estudios realizados, no existe una respuesta final a cuáles son las mejores regiones para diagnosticar la EA. Otro problema presente es que para la mayoría de las regiones, se ha demostrado que aún teniendo un valor alto de predicción positiva, existe un número significativo de pacientes que no muestran estos signos [Nitrini et al., 2000]. En la literatura encontramos que las regiones comúnmente afectadas

por la EA son:

1. Región temporo-parietal: Muchos estudios han mostrado que esta región es típica de la EA. Sin embargo, la mayoría de los estudios fueron llevados a cabo incluyendo pacientes en una etapa avanzada de la EA. Por tanto, esta región no aparece afectada en una fase temprana de la enfermedad. Aunque las anomalías bilaterales temporo-parietales, con o sin defectos regionales, se conocen como los patrones predominantes de la EA [Claus et al., 1994; Messa et al., 1994; Jobst et al., 1992; Talbot et al., 1998] no parecen ser ni sensibles ni específicos de la EA en una fase temprana [Gool et al., 1995; McMurdo et al., 1994].

2. Cíngulo posterior y precunei: [Kogure et al., 2000] mostraron que estas regiones están afectadas en un estudio de sujetos con afección cognitiva leve (MCI) que evolucionaron a la EA tras dos años de seguimiento. Otros estudios confirman estos descubrimientos sobre la fase temprana de la EA [Minoshima et al., 1997; Ishii et al., 1997; Ibañez et al., 1998]. Estos déficits de perfusión son probablemente más específicos y frecuentes en la fase temprana de la EA que los déficits temporo-parietales.

3. Lóbulo temporal medio: La hipo-perfusión en estas regiones fue advertida mediante un seguimiento de pacientes realizado por Kogure et al. [2000]. Estos resultados fueron sorprendentes puesto que previos estudios anatómicos y patológicos habían sugeridos que estas regiones eran las primeras en verse afectadas por la EA [Braak and Braak, 1997; Bobinski et al., 1999]. Otros estudios sugieren que estas regiones, al igual que el hipocampo, no se habían encontrado en la fase MCI debido a las dificultades para visualizar estas estructuras profundas del cerebro [Rodríguez et al., 2000].

La Figura 1.3 muestra imágenes obtenidas mediante PET ^{18}F-FDG, MRI y PET (^{11}C-PIB) para un paciente afectado por la EA. La discusión sobre las regiones de hipo-perfusión y las diferencias que existen incluso entre pacientes sanos muestra las dificultades que existen para los expertos clínicos cuando analizan imágenes de perfusión para la EA. Por tanto sería útil contar con herramientas que puedan ayudar a los expertos en esta difícil tarea. Para un estudio más detallado de las regiones típicas de afección de la EA véase Leon et al. [1983]; Foster et al. [1983, 1984]; Chase et al. [1984]; Duara et al. [1986]; McGeer et al. [1990]; Minoshima et al. [1994, 1995]; Ibañez et al. [1998]; Hoffman et al. [2000]; Kogure et al. [2000]; Alexander et al. [2002]; Mosconi et al. [2008]; Langbaum et al. [2009].

1.2. Reconstrucción de imágenes

Las imágenes de sección cruzada del cerebro pueden reconstruirse a partir de los datos de proyección [Lange and Carson, 1984; Vardi et al., 1985; Hudson and Larkin, 1994; Bruyant, 2002; Chornoboy et al., 1990]. En condiciones ideales, las proyecciones son conjuntos de medidas de los valores integrados de algunos parámetros del objeto. Si el objeto se representa por una función bidimensional $f(x, y)$ y cada integral de línea por los parámetros (θ, t), la integral de línea se define como:

$$P_\theta(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - t)dxdy, \tag{1.1}$$

donde $P_\theta(t)$ se conoce como la transformada Radon de la función $f(x, y)$.

La clave en imagen tomográfica es el *Teorema de Fourier de Cortes* el cual relaciona las medidas de los datos de proyección con la transformada de Fourier de la sección cruzada del objeto. El teorema establece lo siguiente:

Teorema. Teorema de Fourier de Cortes. La transformada de Fourier $S_\theta(w)$ de una proyección paralela $P_\theta(t)$ de una imagen $f(x, y)$ tomada con un ángulo θ y definida del siguiente modo:

$$S_\theta(w) = \int_{-\infty}^{+\infty} P_\theta(t)\exp(-j2\pi wt)dt, \tag{1.2}$$

proporciona una rebanada de la transformada de Fourier bidimensional:

$$F(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\exp(-j2\pi(ux + vy))dxdy, \tag{1.3}$$

que se encuentra a un ángulo θ del eje u, es decir,

$$S_\theta(w) = F(u = w\cos\theta, v = w\sin\theta). \tag{1.4}$$

El resultado de arriba es la esencia de la tomografía de rayo directo e indica que teniendo proyecciones de un objeto con ángulos θ_1, θ_2, ..., θ_k y tomando su transformada de Fourier, los valores de $F(u, v)$ pueden determinarse en líneas radiales. En ese caso, está claro que la función $F(u, v)$ se conoce sólo a lo largo de un número finito de líneas radiales de modo que uno debe interpolar dichos puntos radiales a los puntos de la rejilla cuadrada.

Los datos de proyección usados en este estudio se reconstruyen usando el algoritmo de retroproyección filtrada (FBP) que se deriva fácilmente del teorema de Fourier de Cortes. Una imagen de la sección cruzada $f(x, y)$ de un objeto se puede recuperar por:

$$f(x, y) = \int_0^{\pi} Q_\theta(x \cos\theta + y \sin\theta) d\theta, \tag{1.5}$$

donde

$$Q_\theta(t) = \int_{-\infty}^{+\infty} S_\theta(w)|w| \exp(j2\pi wt) dw. \tag{1.6}$$

El algoritmo FBP consiste entonces en dos pasos: la parte de filtrado, la cual puede verse como una simple ponderación de cada proyección en el dominio frecuencial, y la parte de proyección hacia atrás.

Un gran inconveniente de FBP es que se amplifica de forma indeseada el ruido de altas frecuencias impactando sobre la calidad de la imagen. Estos efectos se producen por la de multiplicación de $S_\theta(w)$ por $|w|$ en la ecuación 10.6. Para atenuar el ruido de altas frecuencias amplificado durante la reconstrucción FBP se han propuesto diversas funciones tipo ventana. De este modo, el método de reconstrucción descrito por las ecuaciones 10.5 y 10.6 se redefine normalmente aplicando una ventana en frecuencia con valores cercanos a cero cuando la frecuencia tiende a π. Entre las ventanas más comunes para la reconstrucción FBP se encuentran *i*) sinc (filtro de Shepp-Logan), *ii*) coseno, *iii*) Hamming y, *iv*) Hanning. Sin embargo, incluso cuando el ruido de reconstrucción se mantiene bajo usando la aproximación de FBP con control de ruido, se necesita filtrar el ruido capturado por el sistema de adquisición para mejorar la calidad de las imágenes reconstruidas. Además, la etapa de preprocesado de la mayoría de sistemas de preprocesado de imágenes a menudo incorpora prefiltrado, reconstrucción y postfiltrado para minimizar el ruido captado por la cámara gamma así como el ruido amplificado por la reconstrucción FBP.

1.3. Registro de imágenes

El preprocesado empleado en el paradigma basado en SPM (Statistical Parametric Mapping) [Friston et al., 2007] es el utilizado en todos los métodos de diagnóstico basados en imagen. Por ejemplo, éstos requieren que las imágenes usadas en el procedimiento sean comparables vóxel a vóxel estableciendo una correspondencia exacta entre posición anatómica y posición del

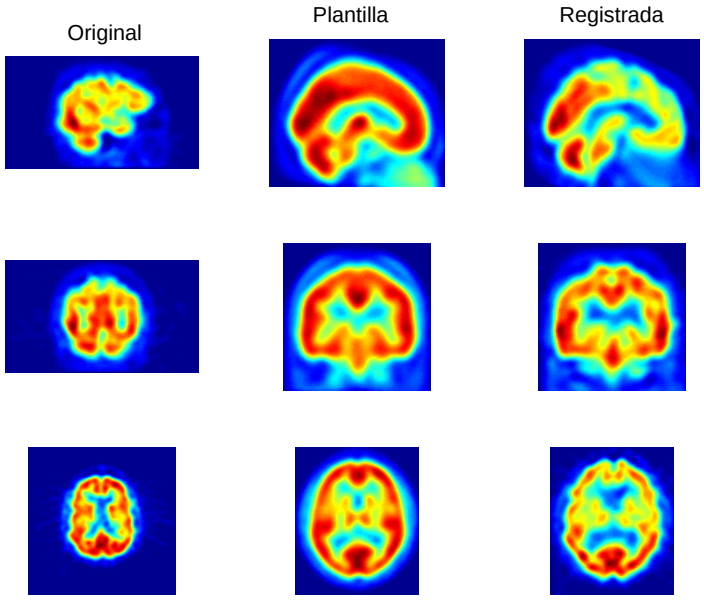


Figura 1.4: Tres imágenes SPECT. *Columna izquierda*: Imagen original. *Columna central*: Plantilla. *Columna derecha*: Imagen transformada tras el proceso de registro.

vóxel en la imagen. Este preprocesado consta de tres etapas: *realineado*, *normalización espacial* y *filtrado espacial*.

En la etapa de *realineado* se consigue corregir la diferente colocación de la cabeza en distintas imágenes de un mismo sujeto dentro del dispositivo de imagen (PET, SPECT, fMRI). Para corregirla, se aplican las traslaciones y rotaciones adecuadas que compensen esta diferencia, de modo que las imágenes coincidan en el mismo espacio común. En el caso de diagnóstico que nos concierne no se aplica tal transformación dado que solo se dispone de una imagen de cada paciente. Estas transformaciones sencillas parten del hecho de que el cerebro a realinear tiene la misma morfología en cada adquisición. Sin embargo cuando disponemos de distintos pacientes no se cumple esta condición y para solucionar este problema aplicamos la etapa de *normalización espacial*. En efecto, para realizar un análisis vóxel a vóxel, los datos de distintos sujetos deben corresponderse con un espacio anatómico estándar que permite la comparación entre sujetos y la presentación de los resultados de un modo convencional.

En esta etapa se realiza una deformación elástica [Friston et al., 2007] de las imágenes de modo que concuerden con una plantilla que sirve de patrón anatómico estandarizado. Para que la transformación espacial sea correcta, las imágenes deben ser razonablemente similares al patrón utilizado, tanto morfológicamente como en contraste. Este patrón se obtiene promediando un conjunto de controles normales de manera que se obtiene una imagen plantilla suavizada que sirve como referencia. De este modo, se ponen en correspondencia cada una de las regiones cerebrales de cada sujeto con una localización homóloga en el espacio estándar. Esta normalización, además de permitir la comparación vóxel a vóxel de las imágenes, también facilita la localización de las áreas funcionales. El concepto de sistematizar la localización cerebral de las regiones funcionales se debe originalmente a Talairach [Talairach and Tournoux, 1988], y si bien SPM presenta los resultados finales mediante este método, el sistema de coordenadas empleado para informar acerca de las localizaciones no es el mismo que el que aparece en el atlas de Talairach, lo que puede inducir a error.

El método de normalización espacial asume un modelo genérico afín con 12 parámetros [Woods, 2000] y una función de coste la cual presenta un valor extremo cuando la plantilla y la imagen se corresponden una con la otra. La función objetivo que se ha de optimizar es la diferencia cuadrática media entre ambas imágenes, la fuente y la plantilla.

$$\text{CF} = \sum_i (f(\mathbf{M}\mathbf{x}_i) - g(\mathbf{x}_i))^2, \tag{1.7}$$

donde f denota la imagen original y g la plantilla. Para cada voxel $\mathbf{x} = (x_1, x_2, x_3)$ de una imagen, la transformación afín a las coordenadas $\mathbf{y} = (y_1, y_2, y_3)$ se expresa mediante la multiplicación matricial $\mathbf{y} = \mathbf{M}\mathbf{x}$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} \tag{1.8}$$

Tras la normalización afín, la imagen resultante se registra usando un modelo de transformación no rígido más complejo. Las deformaciones se parametrizan por una combinación lineal de componentes de las más bajas frecuencias de las bases de la transformada coseno tridimensional [Ashburner and Friston, 1999].

La Figura 1.4 muestra un ejemplo de la operación del proceso de normalización en imágenes SPECT. La columna izquierda muestra una imagen

fuente arbitraria de la base de datos, la columna central muestra la plantilla usada en el registro de la imagen, y finalmente la imagen normalizada correspondiente en la columna derecha. En dicha figura, en la obtención de la imagen transformada se ha aplicado un modelo de deformación pequeña, y una regularización de la energía de curvatura del campo de desplazamiento. Se observa claramente cómo la imagen transformada se ajusta a la forma de la plantilla.

Tras la normalización espacial, una vez garantizada la correspondencia espacial de los voxels entre distintas imágenes, se obtiene una representación de cada individuo de 95×69×79 voxels tanto para imágenes SPECT como para PET. Este paso es esencial a la hora de hacer posible comparaciones entre imágenes, y nuestros resultados dependerán en gran medida de un correcto registro de las imágenes. El reto de la normalización espacial supone conseguir minimizar los efectos que producen las diferencias en la imagen debidas a características individuales de cada sujeto sin alterar las diferencias debidas a efectos de la enfermedad.

1.4. Descripción de las bases de datos

El conjunto de imágenes sobre el cual se realiza nuestro estudio pertenece a dos bases de datos diferentes: un conjunto de 91 imágenes SPECT obtenidas en el Hospital Virgen de las Nieves, (Granada) y dos conjuntos de imágenes PET: uno de ellos consiste en 219 imágenes PET obtenidas de la base de datos ADNI (Alzheimer's Disease Neuroimaging Initiative) de la Universidad de California, y el segundo se trata de una pequeña base de datos de 60 pacientes procedentes de la clínica privada "PET Cartuja" (Sevilla). A continuación se describen con más detalle las características específicas de cada una de ellas.

1.4.1. Base de datos ADNI

El proyecto ADNI fue iniciado en 2003 por el NIA (National Institute on Aging), NIBIB (National Institute of Biomedical Imaging and Bioengineering), el FDA (Food and Drug Administration), compañías farmacéuticas privadas y organizaciones sin ánimo de lucro, como un proyecto conjunto con financiación publico-privada que asciende a 60 millones de dolares. El primer objetivo de ADNI es probar si se pueden combinar las técnicas de MRI, PET y otros marcadores biológicos y evaluaciones neuropsicológicas y clínicas,

para medir la progresión de la enfermedad de Alzheimer en sus primeras etapas (MCI). El descubrimiento de marcadores sensibles y específicos de las etapas mas tempranas de la enfermedad se espera que sirva de ayuda a los investigadores y médicos para desarrollar nuevos tratamientos y probar su efectividad, así como para disminuir el tiempo y el coste de las pruebas clínicas.

El investigador principal de la iniciativa ADNI es Michael W. Weiner (VA Medical Center and University of California, San Francisco). Los pacientes de ADNI fueron seleccionados de mas de 50 lugares entre E.E.U.U y Canadá. El objetivo inicial de ADNI era conseguir que 800 adultos participaran en el proyecto con edades entre 55 y 90 años, resultando aproximadamente unos 200 sujetos normales de edad avanzada con un seguimiento previsto de 3 años, 400 personas con MCI para un seguimiento de 3 años, y 200 personas con los primeros síntomas de la enfermedad de Alzheimer para un seguimiento de 2 años (véase www.adni-info.org para obtener más detalles sobre los objetivos de este proyecto).

Protocolo de adquisición

Las exploraciones FDG PET se tomaron todas siguiendo un protocolo estandarizado. Puesto que un gran número de centros participaron en el estudio, no todos los centros disponían de los mismos recursos materiales para efectuar las exploraciones. Por ello, las exploraciones fueron adquiridas según uno de los tres protocolos siguientes:

- Dinámicas: las exploraciones de emisión dinámica consisten en 6 tomas de 5 minutos, adquiridas de 30 a 60 minutos después de la inyección intravenosa de 5.0 $\pm$0.5 mCi de ^{18}F-FDG.
- Estáticas: las estáticas consisten en una única toma de 30 minutos comenzando de 30 a 60 minutos después de la inyección. Los escáneres Siemens PET/CT no tienen la posibilidad de tomar exploraciones dinámicas.
- Cuantitativas: las cuantitativas consistían en una exploración dinámica del doble de duración que la anterior, con 33 tomas, comenzando en el momento de la inyección y continuando 60 minutos. Esta puede ser usado para calcular la tasa absoluta de metabolismo de glucosa, obtenida de la función de entrada del radioisótopo medida en las arterias carótidas.

La mayoría de las exploraciones de ADNI fueron tomados siguiendo el primer protocolo. Los pacientes permanecieron tumbados relajadamente, con sus ojos abiertos y la estimulación sensorial mínima.

Una vez adquiridas las imágenes, la iniciativa ADNI pretende también minimizar en lo posible las diferencias en los tipos de imágenes debido a la diferencia entre los escáneres usados para obtenerlas. De esta manera, el proyecto proporciona diferentes conjuntos de datos que se pueden descargar de sus archivos ordenados según el tratamiento recibido. El pre-procesamiento de las imágenes ADNI esta detallado a continuación:

1. Co-registro dinámico: las imágenes PET originales de todos los sitios se descargan para el control de calidad en la Universidad de Michigan. Estas imágenes se convierten a un formato de archivo estándar. Si la imagen contiene diferentes secuencias separadas, éstas son extraídas de los archivos de imagen a efectos de registro. En la mayoría de los casos, seis secuencias de cinco minutos se adquieren entre 30 y 60 minutos después de la inyección. Cada secuencia extraída es co-registrada hasta la primera secuencia extraída del archivo de imagen original (la secuencia adquirida en los 30-35 primeros minutos después de la inyección). Las diferentes secuencias se vuelven a co-registrar a una imagen conjunto. Estas series de imágenes tienen el mismo tamaño de la imagen (por ejemplo, $128 \times 128 \times 63$ voxels) y dimensiones (por ejemplo, $2{,}0 \times 2{,}0 \times 2{,}0mm$) y mantienen la misma orientación espacial que la imagen original PET. A esto se les denomina 'nativos'. Estos archivos se cargan en formato DICOM. Sólo los adquiridos siguiendo el protocolo 1 ó 3 tendrán una imagen procesada de este tipo. En resumen, el 'co-registro dinámico' tiene dos principales diferencias principales con la imagen 'original' PET:

- Las secuencias separadas han sido co-registradas entre sí para reducir los efectos de movimiento del paciente y
- Los archivos de la imagen están en formato DICOM.

2. Imagen co-registrada promediada: Este tipo de imagen procesada conjunto se genera como un promedio de sólo 6 secuencias de cinco minutos (o las últimas 6 secuencias de los estudios cuantitativos) de la imagen de conjunto co-registrada que se ha descrito anteriormente. Esto crea una única imagen PET de 30 minutos, todavía en el espacio 'nativo'. Como en el caso anterior, sólo las exploraciones de PET adquiridas en virtud del protocolo de 1 ó 3, tendrá una entrada de este tipo.

3. Imagen de tamaño de voxel estandarizado: Cada imagen a la que se le habían aplicado los anteriores pasos se reorientó a un estándar de $160 \times 160 \times 96$ voxels, teniendo cada voxel un tamaño cúbico de lado $1{,}5mm$. Esta imagen de la red está orientada de tal manera que el eje anterior-posterior del sujeto es paralela a la línea de las comisuras anterior y posterior (AC-PC). Esto se conoce como el espacio 'AC-PC' en el programa de búsqueda Loni. Este primer realineado de la imagen sirve como una imagen de referencia para todas las exploraciones de PET sobre el mismo paciente. Las secuencias individuales de cada exploración PET (el estudio de referencia, así como todos los estudios posteriores (6 meses de exploración, de 12 meses de exploración, etc) son co-registrados sobre esta línea de base de referencia. Al hacer el co-registro de la imagen original en un solo paso, sólo una interpolación de los datos de la imagen es necesaria, y, por tanto, la degradación de la resolución por interpolación se mantiene al mínimo, y es el mismo para todas las exploraciones. Un promedio de la imagen se genera a partir de las secuencias corregistradas 'AC-PC' de intensidad normalizada y, a continuación se utiliza una máscara específica para cada sujeto a fin de que la media de los voxels en la máscara sea exactamente uno. Tanto la orientación espacial (AC-PC) como la normalización de intensidad de la imagen se toman como un punto de partida para posteriores análisis. Con una imagen de normalización de la matriz, el PET de datos de diferentes modelos de escáner se pueden comparar más fácilmente.

4. Imagen de resolución uniforme: Estas imágenes son el resultado de suavizado de las mencionadas imágenes. Cada imagen se filtra conjunto con un escáner específico de función de filtro (puede ser un filtro no isotrópico) para producir imágenes de una resolución isotrópica uniforme de $8mm$ FWHM, la resolución aproximada de la resolución más baja de los escáneres utilizados en ADNI. Los conjuntos de imágenes de mayor resolución de los escáneres, obviamente, se suavizan más que las imágenes de escáneres de baja resolución. Las funciones del filtro se determinan a partir de las exploraciones de PET Hoffman fantasma que fueron adquiridos durante el proceso de certificación.

En nuestro trabajo seleccionamos un conjunto de datos FDG PET de 219 participantes de ADNI, adquiridos con escáneres Siemens, General Electric (GE) y Philips PET, del conjunto ofrecido por el Laboratorio de Neuroimagen ADNI LONI (Laboratory of NeuroImage, University of California, Los Angeles, http://www.loni.ucla.edu/ADNI/). En este primer estudio, los datos adquiridos con escáneres Siemens HRRT y BioGraph HiRez fueron excluidos debidos a las diferencias en el patrón de toma de imágenes.

Criterios de Etiquetado

Los criterios de elección que se siguieron para aceptar a participantes en el proyecto ADNI se basaron en una serie de entrevistas y test realizados individualmente. Los resultados de los candidatos debían cumplir ciertas condiciones para ser admitidos en el proyecto. A continuación se detallan los criterios de selección de pacientes para cada una de las clases de interés para el estudio:

- Pacientes normales o NC (Normal Control): La puntuación obtenida en el test MMSE debía estar entre 24-30 (ambos inclusive), un CDR de 0, no deprimido, no MCI, y sin demencia. El rango de edad de los pacientes normales debe coincidir con la de los sujetos AD y MCI.
- Pacientes MCIs: La puntuación obtenida en el test MMSE debía estar entre 24-30 (ambos inclusive), debía presentar quejas por pérdida de memoria, tener una pérdida objetiva de memoria medida en términos de su puntuación en el test de Wechsler Memory Scale Logical Memory II, un CDR de 0.5, ausencia de discapacidades, conducta normal en las actividades de la vida cotidiana y ausencia de demencia.
- Pacientes AD (Alzheimer's Disease): La puntuación obtenida en el test MMSE debía estar entre 24-30 (ambos inclusive), debía presentar un CDR de 0.5 ó 1.0 y satisfacer los criterios de NINCDS/ADRDA que definen un AD probable.

Consecuentemente, los datos FDG PET se separaron en 3 clases diferentes: sujetos de control NC, sujetos con afección cognitiva leve MCI y enfermos de Alzheimer AD. En nuestro estudio contamos con imágenes de 219 sujetos diferentes, divididos en 53 AD (rango de edad: 77.2 ± 7.2 (media $\pm$ desviación estándar)), 114 MCI (rango de edad: 75.1 ± 7.4), y 52 NC (rango de edad: 76.5 ± 4.8). De los 114 MCI, y tras un seguimiento de 2 años, se concluyó que 4 de ellos se habían convertido al grupo AD (denominados MCI converters) mientras 110 se mantuvieron estables durante ese tiempo (MCI no converters).

1.4.2. Base de datos SPECT

El rendimiento de algoritmos que se presentan en este trabajo se evalúa también sobre una base de datos que contiene 91 imágenes reales SPECT

proporcionadas por el Servicio de Medicina Nuclear del Hospital "Virgen de las Nieves" de Granada (España).

A los pacientes se les inyectó un radiofármaco emisor de rayos gamma ^{99m}Tc-ECD, adquiriendo la imagen original mediante una gamma cámara de tres cabezales Picker Prism 3000. Se tomaron 180 proyecciones con resolución angular de 2 grados. Estas imágenes de secciones del cerebro fueron reconstruidas de los datos de proyección usando el algoritmo de filtrado mediante proyección hacia atrás (FBP) en combinación con un filtro de Butterworth para eliminar ruido de alta frecuencia. Las imágenes SPECT son primeramente normalizadas usando SPM [Friston et al., 2007], de manera que podamos aseverar que los voxels de diferentes imágenes representan la misma posición anatómica subyacente en el cerebro. Este paso nos permite comparar las intensidades de los voxels de distintos sujetos [Salas-Gonzalez et al., 2008a].

Las imágenes SPECT fueron, a su vez, etiquetadas por los expertos del "Virgen de las Nieves" mediante 4 categorías: controles normales (NOR), AD posible (AD1), AD probable (AD2) y AD cierto (AD3) para distinguir entre distintos niveles de la presencia del patrón de AD.

1.4.3. PET *Cartuja*

Por último, algunos algoritmos desarrollados en este trabajo se evalúan sobre una base de datos PET de 60 muestras proporcionada por la clínica privada "PET *Cartuja*" en Sevilla. El protocolo de adquisición de estas imágenes comenzó a partir de 30 minutos desde la administración de la ^{18}FDG al paciente. En este tiempo el paciente permanece en reposo en una habitación en silencio y con iluminación tenue para que el contraste se distribuya adecuadamente por todo el organismo. Se recomienda fijar un tiempo estándar, por ejemplo, 30 minutos, para que los estudios de distintos pacientes a los de distintos controles sean comparables. El paciente se posiciona en decúbito supino. Debe dedicarse un sistema dedicado para apoyar confortablemente la cabeza, fijándola mediante cintas para evitar movimientos involuntarios. El paciente se sitúa acostado en la camilla de la cámara y se desplazará progresivamente por el centro del anillo de la Cámara PET durante un tiempo aproximado de 30 minutos en un estudio de PET cerebral.

Durante este tiempo la cámara PET recoge las señales emitidas por el contraste en todo el cuerpo. Después, un ordenador recoge las señales emitidas y las convierte en imágenes funcionales tridimensionales en los tres planos del

espacio (axial, coronal y sagital) y una imagen volumétrica del organismo en tres dimensiones. Los positrones emitidos por el radiofármaco colisionan con los electrones (con carga negativa) de los átomos que componen las moléculas tisulares. La interacción positrón-electrón origina la aparición de un par de fotones con el aniquilamiento de las masas del positrón y electrón. Estos dos fotones presentan una energía de $512KeV$ cada uno, y se desplazan en la misma dirección y en sentidos opuestos, excitando de forma simultánea 2 detectores de la cámara PET que se encuentran en coincidencia en un ángulo de 180°. Esta detección permite "por coincidencia" la reconstrucción tomográfica tridimensional del organismo que representa la distribución tisular del radiofármaco.

Los 60 pacientes PET se dividen según el etiquetado en 42 AD y 18 normal, lo que supone una prevalencia de la clase AD que ha de tenerse en cuenta en análisis posteriores.

CAPÍTULO 2

Técnicas de Diagnóstico de la Enfermedad de Alzheimer

Un diagnóstico precoz y exacto de la enfermedad de Alzheimer ayuda a los pacientes y a sus familias a planear el futuro. El diagnóstico precoz también ofrece la mejor oportunidad para tratar los síntomas de la enfermedad. Actualmente la única manera definitiva para diagnosticar la enfermedad de Alzheimer es investigar sobre la existencia de placas y enredos en el tejido cerebral. Para observar el tejido cerebral los médicos deben esperar hasta el examen patológico *postmortem*, tras practicar la autopsia. Por consiguiente, los médicos deben hacer un diagnóstico de 'posible' o 'probable' enfermedad de Alzheimer basada en la experiencia. Los exámenes del cerebro mediante escáneres pueden permitir al médico observar indirectamente el cerebro para detectar la presencia de anormalidades en él. Hasta ahora es todo el partido que se está sacando a las nuevas técnicas de imágenes cerebrales a nivel médico en los servicios de medicina nuclear, quedando abierta la posible automatización del proceso, así como la clasificación-detección de estas alteraciones desde un punto de vista matemático.

2.1. Criterios de diagnóstico

El diagnóstico de la demencia se basa fundamentalmente en la evaluación clínica y ésta requiere una exhaustiva evaluación de la función cognitiva, en concreto de la memoria, atención, percepción, lenguaje, praxias y gnosias. La evaluación neuropsicológica puede subdividirse en dos niveles de complejidad:

- Un primer nivel que consiste en el uso de test breves, estandarizados y sencillos como el Mini-Mental State Examination (MMSE), que posibilita el diagnóstico de la demencia, y
- Un segundo nivel de mayor complejidad en el que se refina la evaluación de la severidad del deterioro, al tiempo que se establecen los dominios de la función cognitiva que se hallan afectados. Existen diferentes escalas que otorgan un valor estandarizado en función del grado de deterioro funcional, como CDR (Clinical Dementia Rating) o GDS (Global Deterioration Scale). En general, estas escalas permiten clasificar la demencia según los criterios clínicos clásicos: demencia leve, moderada o severa.

Los cuestionarios o escalas han sido diseñados para cuantificar determinadas funciones cognitivas, es decir, no establecen un diagnóstico sino que cuantifican la severidad de la alteración de determinadas áreas intelectuales, siendo particularmente valiosos para discriminar entre envejecimiento normal y demencias leves. El diagnóstico siempre ha de realizarse basado en la historia clínica del paciente y de acuerdo con los criterios establecidos al respecto. En este sentido, los cuestionarios representan sólo una ayuda en el proceso de valoración de la enfermedad.

2.1.1. Examen del Estado Mental Mínimo (MMSE)

El MMSE es uno de los test de valoración de la demencia más utilizados. Es un test que tiene alta dependencia del lenguaje y consta de varios ítems relacionados con la atención. Cada ítem tiene una puntuación, llegando a un total de 30 puntos. En la práctica diaria una puntuación menor de 24 sugiere demencia, entre 23-21 una demencia leve, entre 20-11 una demencia moderada y menor de 10 de una demencia severa. Para poder efectuar el MMSE es necesario que el paciente se encuentre despierto y lúcido. En la demencia por enfermedad de Alzheimer la tasa promedio anual de cambio en

la puntuación del MMSE es de 2-5 puntos por año, por lo que el test muestra su utilidad para el seguimiento de los pacientes afectados. El MMSE tiene baja sensibilidad para el diagnóstico de deterioro cognitivo leve, la demencia frontal-subcortical y el déficit focal cognitivo. Las características esenciales que se evalúan son las siguientes:

- Capacidad de atención, concentración y memoria.
- Capacidad de abstracción (cálculo).
- Capacidad de lenguaje y percepción visoespacial.
- Orientación espacio-tiempo.
- Capacidad para seguir instrucciones básicas.

Esta prueba proporciona un instrumento para la detección de deterioro cognitivo que se puede realizar en poco tiempo. Ésta es una cualidad especialmente importante en la evaluación de la demencia, ya que el paciente en ocasiones es propenso a perder la atención con facilidad y por tanto dejar de colaborar en la elaboración del test.

2.1.2. Escala de Deterioro Global (GDS)

La GDS [Sheikh and Yesavage, 1986] establece siete estadios posibles:

- 1 = normal
- 2 = deterioro muy leve
- 3 = deterioro leve
- 4 = deterioro moderado
- 5 = deterioro moderadamente severo
- 6 = deterioro severo
- 7 = deterioro muy severo

La escala define cada clase en términos operacionales y asume que el deterioro es homogéneo. Sin embargo, dado que la secuencia de aparición de los síntomas es a menudo variable, se ha argumentado que la inclusión de un paciente en un estadio de acuerdo a un criterio rígido podría conducir a error; a pesar de ello se trata de una de las escalas más completas, simples y útiles para la estimación de la severidad de la demencia.

2.1.3. Clasificación Clínica de la Demencia (CDR)

La evaluación de otros tipos de demencia aparte del Alzheimer se realiza normalmente a través de la CDR [Morris, 1993] que es más general. Su escala establece cinco estadios posibles:

- 0 = normal
- 0,5 = cuestionable
- 1 = demencia leve
- 2 = demencia moderada
- 3 = demencia severa

La estimación se basa en el rendimiento del sujeto en seis modalidades de tipo cognitivo y funcional. Estas modalidades son: memoria, orientación, razonamiento, actividades sociolaborales, actividades recreativas y cuidado personal.

2.1.4. Escala de Evaluación para la Enfermedad de Alzheimer (ADAS)

En 1984 con la aparición del ADAS (Alzheimer's Disease Assessment Scale) o Escala de Evaluación para la Enfermedad de Alzheimer (EEEA) fue posible contar con un instrumento fiable y breve diseñado especialmente para la enfermedad de Alzheimer y capaz de medir puntualmente los síntomas característicos de la misma así como su progresión a estadios más avanzados.

El ADAS es un test que consta de 21 ítems organizados a su vez en dos subescalas: cognitiva (ADAS-Cog) y conductual (ADAS-Noncog), de 11 y 10 ítems respectivamente. En la práctica se ha hecho muy popular el uso

del ADAS-Cog, mientras que el ADAS-Noncog ha sido reemplazado por escalas conductuales más generales como el Cuestionario de Inventario Neuropsiquiátrico o NPI-Q (NeuroPsychiatric Inventory Questionnaire). Éste es un instrumento de aplicación relativamente breve que mide una serie de síntomas habituales en función de la frecuencia de aparición y la intensidad con que aparecen.

2.2. Diagnóstico asistido por computador

Hasta la fecha se han propuesto numerosos sistemas de ayuda al diagnóstico de enfermedades neurológicas, con el objetivo de analizar imágenes de tipo SPECT u otros tipos de imágenes funcionales de forma automática [Ramírez et al., 2009; Ramírez et al., 2009; Górriz et al., 2009; López et al., 2009b; Segovia et al., 2009b]. Las técnicas existentes en la literatura pueden dividirse entre aquellas que hacen uso de métodos estadísticos univariados o multivariados. Otra posible clasificación distingue aquellas que obtienen resultados de manera no supervisada, frente a aquellas que hacen uso de algún tipo de aprendizaje estadístico supervisado, aplicando cierto entrenamiento previo. Estudiaremos los casos más relevantes de estas técnicas.

La aproximación univariada más relevante está basada en SPM (del inglés Statistical Parametric Mapping) y sus numerosas variantes [Friston et al., 2007]. SPM consiste en hacer un test estadístico univariado a nivel de voxel (unidad de volumen mínimo de la imagen), por ejemplo un test t-student de dos muestras, que compara los valores del vóxel de la imagen bajo estudio con un grupo de pacientes normales que representan la muestra 'control'. A continuación, los voxels relevantes de este test son inferidos usando la teoría de campos aleatorios [Adler, 1981]. Su marco de actuación fue pensado para el análisis de estudios de imágenes SPECT y PET, pero actualmente se aplica principalmente para el análisis de la imagen de resonancia magnética funcional o fMRI (functional Magnetic Resonance Imaging).

Sin embargo, SPM no está estrictamente diseñada para resolver el problema del diagnóstico automático usando exclusivamente un paciente de estudio sino para la comparación de conjuntos de imágenes a las cuales se les asigna una etiqueta implícitamente (como es el caso del diagnóstico que nos ocupa). De hecho, su aplicación en este contexto proporciona resultados de clasificación pobres (semejantes a nuestra metodología de referencia [Stoeckel et al., 2001]) dado que una de las poblaciones consiste en un único individuo (estimación sesgada de la media de la población), y la otra consiste en un

conjunto de individuos normales (el test t no incluye ninguna información sobre el patrón de imagen típico de la enfermedad bajo estudio) [Stoeckel et al., 2001]. Más aún, este método sufre los inconvenientes de las aproximaciones locales y univariadas.

Por otro lado, se han propuesto otras aproximaciones multivariadas como ManCova, que consideran como una observación todos los voxels de un solo "scan" con el objetivo de hacer inferencias sobre los efectos de activación distribuidos. La importancia de ellas radica en que los efectos debidos a activaciones, los efectos indefinibles ("confounding effects") y los efectos de error son evaluados estadísticamente en términos de efectos a nivel de vóxel y también a nivel de las interacciones entre voxels [Frackowiak et al., 2003]. No obstante, con estas técnicas uno no puede hacer inferencias estadísticas sobre cambios específicos regionales, y, aun más importante, requieren un número de observaciones (scans) que sea mayor que el número de componentes de la observación multivariada (voxels). Obviamente esta no es la situación en la que nos encontramos cuando trabajamos con estudios de imágenes funcionales (SPECT, PET, fMRI).

El trabajo de esta investigación se centra en el contexto de las aproximaciones *supervisadas* y multivariadas donde se plantea un nuevo método cuantitativo para evaluación de imágenes funcionales. En este campo, la clasificación se realiza habitualmente mediante la definición de vectores de características que representan los rasgos más relevantes de las diferentes imágenes (por ejemplo las de SPECT) y mediante el entrenamiento de un clasificador dado un conjunto de muestras con etiquetas conocidas [Álvarez et al., 2009a; Górriz et al., 2008; Segovia et al., 2010; López et al., 2009a], etc. Tras el proceso de entrenamiento, el clasificador (que incluye la capacidad de generalización del sistema) se aplica a nuevos casos de test para distinguir entre controles sanos y enfermos. En este régimen de entrenamiento y test se asume como hipótesis de partida que las etiquetas de entrenamiento y test son válidas por lo que una precisión elevada en la clasificación es equivalente a un diagnóstico efectivo del paciente bajo estudio.

El conjunto de clasificadores usados por los sistemas CAD están basados en distintas funciones analíticas (por ejemplo en su complejidad) que se ajustan mediante datos de entrenamiento en base a distintos procedimientos. De entre esos procedimientos destacamos, por su robustez, el uso de máquinas de vectores soporte (SVMs) [Vapnik, 1998]. Las SVMs han atraído recientemente la atención de la comunidad científica en el campo de reconocimiento de patrones dado el alto número de avances a nivel teórico y computacional derivados de la Teoría del Aprendizaje Estadístico (TAE) [Vapnik, 1998] de-

sarrollada por Vladimir Vapnik en los laboratorios de AT&T. Estas técnicas han sido usadas de manera satisfactoria en un gran número de aplicaciones entre las que destacamos la detección de actividad de voz (VAD) [Ramírez et al., 2006b], recuperación de imágenes basadas en contenido [Tao et al., 2006], clasificación de texturas [Kim et al., 2002] o el diagnóstico basado en imagen [Fung and Stoeckel, 2007; Álvarez et al., 2009a; Salas-Gonzalez et al., 2008b, 2009d].

Los sistemas CAD desarrollados hasta el momento con el fin de detectar patrones de hipoperfusión e hipometabolismo combinan la extracción de información a partir de las imágenes con diversos métodos de clasificación [Salas-Gonzalez et al., 2010; Ramírez et al., 2009, 2010]. Las características extraídas pueden estar directamente relacionadas con el nivel de intensidad que presenta cada vóxel, obteniendolas por ejemplo mediante el cálculo de estadísticos [Ramírez et al., 2008; Chaves et al., 2009] o la agrupación de voxels con características similares [Segovia et al., 2010]. Además, las imágenes pueden ser analizadas en otros espacios transformados que presentan ventajas para la clasificación con respecto al espacio imagen original [Álvarez et al., 2010; Ramírez et al., 2010]. En este trabajo se profundiza en la extracción de características en espacios de proyección tanto lineales como no lineales, obtenidos estos últimos mediante la aplicación de técnicas kernel [López et al., 2009c,d], con las que se evalúa no sólo uno, sino tres métodos de clasificación diferentes con el fin de determinar la mejor combinación de técnicas de extracción y clasificación para el problema que nos ocupa. La ventaja de todas estas aproximaciones para el diagnóstico clínico basada en la TAE es que no es necesario ningún tipo de conocimiento a priori acerca de la enfermedad bajo estudio y que el método que es automático, es aplicable a cualquier otro tipo de patología neurológica o técnica de imagen cerebral.

2.2.1. Statistical Parametric Mapping (SPM)

En el ámbito de la investigación médica, el análisis de neuroimágenes funcionales (PET, SPECT y fMRI) mediante técnicas de cuantificación estadística permite el estudio de diversos procesos cerebrales, patológicos o cognitivos. Una herramienta de creciente uso para este fin es el conocido programa SPM, gracias a su amplia disponibilidad y el gran abanico de estudios estadísticos que permite realizar. Sin embargo, el desconocimiento de los fundamentos teóricos en los que se basa puede conducir fácilmente a resultados imprecisos e incluso a conclusiones erróneas. Esta sección presenta brevemente dichos principios teóricos y discute los principales puntos críticos

en la utilización del método sin detallar los fundamentos matemáticos en los que está basada la citada herramienta.

La finalidad de SPM es la realización de mapas de estadísticos paramétricos para la búsqueda de efectos de interés presentes en imágenes funcionales PET, SPECT o fMRI. Desde su primera aparición en 1991, la comunidad de investigadores de neuroimagen funcional ha aceptado y utilizado ampliamente las actualizaciones de 1996 y 1999, gracias a que proporcionan una gran flexibilidad en el diseño de los experimentos que pueden analizarse [Friston et al., 2007]. SPM se utiliza actualmente en departamentos de psiquiatría, psicología, neurología, radiología, medicina nuclear, farmacología, ciencias cognitivas y del comportamiento, bioestadística y física biomédica de todo el mundo para la investigación de enfermedades mentales, cuantificación de efectos farmacológicos, estudios cognitivos, realización de análisis longitudinales, estudios intersujeto, e incluso morfométricos. En un estudio estadístico mediante SPM, los puntos clave son la elección del método de normalización en intensidad, la normalización espacial, el sistema de coordenadas empleado y la interpretación de la significación estadística de los resultados.

Mediante SPM es posible realizar numerosos tests estadísticos, como regresiones, tests t de Student, tests F y análisis de varianza (Anova) incluyendo covariables y permitiendo el modelado de interacciones entre ellas [Friston et al., 2007]. Todos estos tipos de análisis pueden ser englobados en un modelo general (el modelo lineal general o GLM), que es el utilizado por SPM para efectuar los cálculos matemáticos. La formulación del GLM se basa en dos conceptos: la matriz de diseño y los contrastes.

Los estudios estadísticos que pueden efectuarse mediante SPM pueden ser divididos fundamentalmente en dos tipos: estudios paramétricos o factoriales y estudios categóricos o sustractivos. Los primeros estudian la relación entre las imágenes y un parámetro, como puede ser la edad, una escala de síntomas o el resultado de un test cognitivo. Los estudios categóricos se utilizan para poner de manifiesto diferencias entre grupos, definidas por variables categóricas. Mediante el uso del GLM somos capaces de introducir de manera matemática el estudio estadístico en cuestión, que en el caso del diagnóstico de un paciente a partir de una sola imagen de test, es la definición de la matriz de diseño. Dentro de los estudios categóricos es necesaria una columna para determinar la pertenencia a cada uno de los grupos (categoría normal frente a Alzheimer). Por ejemplo, si se desea comparar un grupo de pacientes con otro de control, las filas correspondientes a los pacientes en la matriz de diseño, tendrán un uno en la columna de pacientes y un cero en la de controles, y viceversa.

Test t

El estadístico básico que se usa para comparar las diferencias entre imágenes es el estadístico t. En esta sección se considera el estudio de una única localización de voxel en diferentes imágenes. El estadístico se usa para comprobar si dos muestras independientes representan dos poblaciones normales con diferentes valores medios. Éste es exactamente el problema cuando se compara si el valor del voxel de la imagen bajo estudio es significantemente menor que los valores de los voxels de las imágenes normales, es decir, si existe hipo-perfusión significante. Aquí, una población se representa por una sola muestra, el valor del voxel de la imagen bajo estudio, y la segunda población se representa por los valores de los voxels de todas las imágenes normales. Puesto que sólo se tiene una muestra en la primera población, se asumirá que la varianza de ambas poblaciones es la misma y ésta se estimará sobre la segunda población.

El estadístico t se calcula del siguiente modo:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{(n_1-1)\bar{s}_1^2+(n_2-1)\bar{s}_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{2.1}$$

donde $\bar{x}_1$ y $\bar{x}_2$ son las medias muestrales estimadas, $\bar{s}_1^2$ y $\bar{s}_2^2$ las varianzas estimadas y n_1 y n_2 son los tamaños muestrales de la primera y segunda población respectivamente. Se puede ver que si n_1 es 1 sólo se usa la estimación de la varianza de la segunda población.

Se puede calcular el estadístico t para todos los voxels de la imagen. Esta nueva imagen, que contiene un valor de t para cada voxel se denomina *imagen* t o *mapa* t. Un nombre general para mapas que contienen un estadístico en cada posición es un SPM. No se puede realizar inferencia sobre estos mapas, pero pueden ser usados para derivar otros estadísticos útiles.

El modelo general lineal

Sea $\mathbf{Y}$ una matriz de dimensión $N \times n$, donde N es el número de imágenes y n el número de voxels en cada imagen. Por tanto, cada fila de $\mathbf{Y}$ representa una única imagen. Todos los valores de los voxels de la imagen tridimensional se colocan uno tras otro para formar una fila de $\mathbf{Y}$. Sea $\mathbf{X}$ una matriz de dimensiones $n \times p$, denominada la matriz de diseño. Ésta contiene una fila

por cada imagen y una columna p por cada variable explicativa (factor o efecto) del modelo. El GLM explica la variación en $\mathbf{Y}$ en términos de una combinación lineal de las variables explicativas, más un término de error:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

donde la matriz β de dimensión $p \times n$ contiene los parámetros desconocidos correspondientes a cada variable explanatoria p. ϵ es la matriz de términos de error de $N \times n$. El número de grados de libertad de este modelo es $\gamma = N - rank(\mathbf{X})$. Si $\mathbf{X}^T\mathbf{X}$ es invertible, entonces β se puede estimar como:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.2}$$

Los residuos se definen como $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{MY}$ donde $\mathbf{M} = \mathbf{I} - (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$.

Para continuar se asume que cada elemento de las filas de ϵ son independientes e idénticamente distribuidos. Así, ϵ se distribuye como una normal $N(0, \mathbf{I}_N \otimes \mathbf{\Sigma})$ donde $\mathbf{I}_N$ es la matriz identidad de $N \times N$. El test t puede implementarse como una combinación lineal de los parámetros estimados. El vector que define la combinación se denomina vector contraste c (de dimensión $1 \times p$), dando lugar al contraste $c\hat{\beta}$. La varianza de este contraste se estima como:

$$\hat{\sigma}^2 = diag\left(\mathbf{R}^T\mathbf{R}(c(\mathbf{X}^T\mathbf{X})^{-1}c^T)\right)/\gamma \tag{2.3}$$

Una imagen t donde cada voxel contiene la distribución t con γ grados de libertad puede definirse dividiendo el contraste por la varianza estimada del contraste

$$t = c\hat{\beta}/\hat{\sigma} \tag{2.4}$$

Una vez establecido el modelo (véase la Figura 2.1), SPM ya puede estimar de forma automática la contribución de cada efecto de forma separada. Esto permite diferenciar entre efectos de "interés" (como el efecto de grupo) y efectos "correctores" (por ejemplo, efecto de la edad en estudios paramétricos), así como diferencias entre las medias de dos factores. Mediante el GLM, esto se realiza mediante la definición de un "contraste" que se define como un vector. La longitud de este vector es igual al número de efectos incluidos

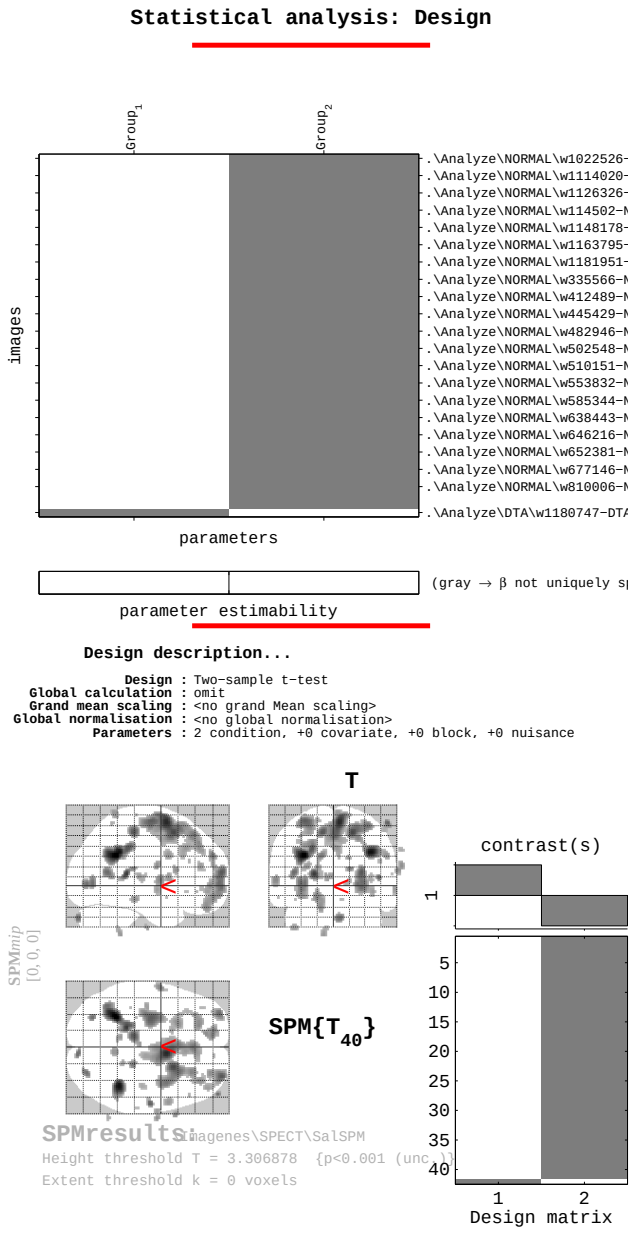


Figura 2.1: Resultados de SPM para clasificación con el modelo descrito en el texto.

en la matriz de diseño, de modo que cada efecto se pondera por su elemento correspondiente. Si el efecto es corrector, entonces se pondera con un cero en el vector contraste. En caso de que el efecto sea paramétrico, el contraste determina si la correlación buscada es positiva, mediante un “1”, o bien negativa mediante un “-1”, en la posición correspondiente a ese efecto

en el vector contraste. En caso de efectos categóricos los contrastes deben cumplir una condición importante: la suma de todos los pesos en el contraste en las columnas de efectos categóricos debe ser igual a cero. En el ejemplo, la pertenencia al grupo de pacientes contribuye con peso negativo (menor metabolismo que controles) y el grupo de controles contribuye con peso positivo (mayor metabolismo que pacientes). De este modo, el vector de contraste sería [1 -1], ya que la suma de los efectos 1 y 2 debe ser cero. En el caso contrario, para comprobar qué regiones presentan un metabolismo mayor en pacientes, el contraste sería [-1 1].

Finalmente, SPM realiza el test estadístico (un test t o un test F), descrito mediante la matriz de diseño y el contraste en todos los voxels de la imagen de forma independiente. El resultado es una imagen cuyo valor en cada vóxel es el resultado del test estadístico.

2.2.2. Voxels-As-Features (VAF)

Esta es la aproximación básica que a lo largo de este trabajo se usa como referencia. En ella, las intensidades de los voxels I_n de la imagen funcional, por ejemplo SPECT, son directamente usados para construir los vectores de características $v = (I_1, \ldots, I_N)$, véase por ejemplo Stoeckel et al. [2001, 2004]. Esta aproximación ha demostrado en las anteriores referencias ser igual de exacta que el diagnóstico visual usando como apoyo SPM, sin embargo presenta los mismos problemas que las aproximaciones multivariadas como Mancova: incluso después de bajar la resolución de la imagen submuestreando, y tras aplicar una *máscara* cerebral seguimos teniendo un problema en la clasificación de dimensión $N \sim 10000$ en los vectores de características. Por lo tanto, el tamaño de estos vectores es muy superior al número de muestras (50-100 es un número realista), lo cual conduce al llamado problema de tamaño muestral pequeño (en inglés *small sample size problem*) [Duin, 2000], que explicaremos en detalle en la sección 3.3.

CAPÍTULO 3

Aprendizaje Estadístico Supervisado en Neurociencias

El uso de algoritmos de aprendizaje estadístico en el entrenamiento de clasificadores para la decodificación de estímulos, estados mentales, comportamientos y otras variables de interés a partir de imágenes neurológicas ha tenido un interés creciente en los últimos años. Cada vez existen más estudios que muestran que el aprendizaje estadístico de clasificadores puede usarse para extraer información nueva de datos de neuroimagen [Norman et al., 2006; Haynes and Rees, 2006]. Las aplicaciones en las cuales el conjunto de entrenamiento comprende vectores de entrada con su correspondiente vector objetivo (o etiqueta) se conocen como problemas de aprendizaje *supervisado*. Una vez entrenados, los clasificadores pueden usarse para determinar si las características usadas contienen información de clase. Si el clasificador captura la relación existente entre las características de las clases, entonces será capaz de predecir la clase a la que pertenecen muestras nunca vistas anteriormente.

3.1. Aprendizaje supervisado

Una rama importante de los métodos de ayuda al diagnóstico médico por ordenador se basa en las técnicas conocidas como *métodos de clasificación*, en el contexto de la Teoría de Aprendizaje Estadístico. Estas técnicas consisten en la extracción de cierta información a partir de un conjunto de datos (aprendizaje automático), que es empleada posteriormente para determinar diagnósticos de nuevos pacientes. En el sentido más amplio, cualquier método que incorpora información de un conjunto de muestras para el diseño de una función clasificadora emplea aprendizaje. Con aprendizaje nos referimos, en este contexto, a un procedimiento algorítmico para reducir el error de la clasificación en el conjunto de entrenamiento. Dado que casi todos los problemas de reconocimiento de patrones son bastante complejos se suele invertir la mayor parte del tiempo en la fase de aprendizaje.

Una de las formas más populares de aprendizaje automático se denomina con el sufijo de *supervisado*. El aprendizaje supervisado implica el uso de datos de entrenamiento sobre los que se conoce la clase a la que pertenecen. En el caso de diagnóstico, las clases vendrán determinadas mediante una etiqueta binaria (± 1), que determinará si el paciente padece o no la enfermedad. El proceso de etiquetado dependerá de la base de datos en concreto, aunque en general se asumirá que el error en el etiquetado es despreciable. Así, las imágenes cerebrales contendrán la información de diferentes procesos que tienen lugar en el cerebro en forma de distribuciones de intensidad en 3 dimensiones y sus etiquetas correspondientes. La distribución de intensidad I corresponderá a diferentes aspectos de la actividad funcional del cerebro, ya sea una imagen SPECT o PET (como por ejemplo la tasa de consumo de glucosa), y estará discretizada a un número n finito de elementos de volumen llamados *voxels*. Podremos agrupar el conjunto de datos originales a partir de los valores que toma la intensidad en los diferentes puntos del cerebro, para formar vectores:

$$\mathbf{x} = (I_1, I_2, ..., I_N) \tag{3.1}$$

que nos permitirán construir los objetos del aprendizaje supervisado: los vectores de características.

> **Definición 1**: El conjunto de datos experimentales o muestras en el proceso de clasificación estará formado por un conjunto de *vectores de características* $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, ..., N$, siendo n

la dimensión del *espacio de características* $\mathcal{H}$. Este conjunto de vectores de características se dividirá en 2 subconjuntos: datos de entrenamiento y datos de test:

- Diremos que un vector de características $\mathbf{x}_i$ es un *vector de entrenamiento* si pertenece al subconjunto de entrenamiento $\mathcal{X} \subset \mathbb{R}^n$. Expresado matemáticamente, $\mathbf{x}_i \in \mathcal{X}$.
- Diremos que un vector de características $\bar{\mathbf{x}}_j$ es un *vector de test* si pertenece al subconjunto de test $\mathcal{Y} \subset \mathbb{R}^n$. Expresado matemáticamente, $\bar{\mathbf{x}}_j \in \mathcal{Y}$.
- La unión de estos dos subconjuntos formará el espacio de características $\mathcal{H}$, es decir, $\mathcal{H} = \mathcal{X} \cup \mathcal{Y}$. En nuestro caso concreto, asumiremos que $\mathcal{H}$ es sencillamente $\mathbb{R}^n$.

Los vectores de características se obtendrán del conjunto de datos originales (3.1), tomándose toda o parte de la información contenida en ellos. A menudo se obtendrán de ellos aquellas características más representativas para efectuar la clasificación, lo que supondrá una disminución del tamaño de los datos originales, cuestión que abordaremos en profundidad en la sección 3.3. Una vez obtenidos los vectores de características, será posible definir el clasificador:

Definición 2: Se define un clasificador como una función $f = f(\mathbf{x}_i, \omega)$, dependiente de unos parámetros ω y construida a partir de los datos de entrenamiento n-dimensionales $\mathbf{x}_i \in \mathcal{X}$ y sus correspondientes etiquetas y_i, tal que clasifica un nuevo vector de test $\bar{\mathbf{x}}_j \in \mathcal{Y}$, asignándole un valor $z_j \in \{\pm 1\}$ correspondiente a cada clase:

$$\begin{array}{rccl} f: & \mathbb{R}^n & \longrightarrow & \{\pm 1\} \\ & \bar{\mathbf{x}}_j \in \mathcal{Y} & \longmapsto & f(\bar{\mathbf{x}}_j, \omega) = z_j \end{array} \tag{3.2}$$

El objetivo del proceso de aprendizaje supervisado es construir una función o clasificador de manera que f categorice correctamente nuevos patrones $\mathbf{x}_j$. Así, el proceso de clasificación se puede dividir en dos etapas: el entrenamiento y el test. En la etapa de entrenamiento se define el clasificador de acuerdo con un conjunto de vectores de características de entrenamiento $\mathbf{x}_i \in \mathcal{X}$, cuyas etiquetas son conocidas. La forma exacta del clasificador dependerá del modelo propuesto para la tarea de clasificación, y los parámetros

desconocidos del modelo serán estimados empleando los patrones de entrenamiento $\mathbf{x}_i \in \mathcal{X}$, $i = 1, ..., l$. El algoritmo de aprendizaje automático consistirá en estimar los parámetros ω, de manera que el error de clasificación sobre el conjunto de entrenamiento sea mínimo. Una vez definido, el clasificador se emplea para establecer categorías sobre muestras desconocidas pertenecientes al conjunto de test.

La elección de los conjuntos de entrenamiento y test es un problema ampliamente estudiado en la literatura. Una condición necesaria es que los conjuntos sean independientes, es decir, que no existan elementos del conjunto de test que hayan sido utilizados en el entrenamiento del clasificador. Expresado matemáticamente requiere que:

$$\mathcal{X} \bigcap \mathcal{Y} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l\} \bigcap \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_m\} = \oslash \tag{3.3}$$

Además, es importante que estos conjuntos sean muestras representativas de la población que se quiere estudiar y clasificar, ya que esto permitiría construir el clasificador adecuadamente y dar una buena representación de su comportamiento.

El método que permite evaluar los resultados de cualquier análisis estadístico se conoce como *validación*, y se estudiará en la sección 3.6. Uno de los aspectos interesantes a evaluar de un clasificador es cómo los resultados pueden ser generalizados a un conjunto de datos independientes. La *generalización* es un concepto central en el diseño de clasificadores y hace referencia a la capacidad de un clasificador construido a partir de un conjunto muestral de entrenamiento determinado, para describir la estructura subyacente de la población total, y no de la muestra concreta. Así, un clasificador con buena capacidad de generalización será capaz de operar correctamente con nuevos datos independientes. Otros aspectos a valorar sobre el rendimiento de un clasificador se estudiarán en la sección 3.5.

Es posible que el conjunto de vectores de características quede representado de una manera más simple en un espacio diferente al original. En estos casos, la eficiencia del clasificador mejorará cuando se realice la transformación oportuna, a través de las funciones conocidas como *kernels*. El criterio a seguir lo marcará el *Teorema de Mercer*, estudiado en la sección 8.1, imponiendo las condiciones necesarias para que esta transformación pueda tener lugar.

Los métodos de clasificación, por tanto, son técnicas complejas en las que intervienen varios factores: definición del clasificador, extracción de ca-

racterísticas, aprendizaje o entrenamiento, y test. Uno de los problemas que será necesario a la hora de construir un clasificador eficaz de imágenes médicas para la ayuda al diagnóstico, es el problema del pequeño tamaño muestral, que se produce cuando el número de pacientes que intervienen en los estudios es muy pequeño en comparación con la gran cantidad de información contenida en las imágenes cerebrales. En la sección 3.3 estudiaremos cómo abordar este problema.

3.2. Procesado

3.2.1. Eliminación de *outliers*

Un *outlier* se define como un punto que se sitúa lejos de la media de la variable aleatoria correspondiente. Esta distancia se mide con respecto a un umbral determinado, habitualmente un número de veces las desviación típica. Para una variable aleatoria distribuida normalmente una distancia de dos veces la desviación típica cubre un 95 % de los puntos, y una distancia de tres veces la desviación típica cubre un 99 % de los puntos. Puntos con valores muy diferentes del valor medio producen grandes errores durante la fase de entrenamiento y pueden tener efectos desastrosos. Estos efectos pueden ser incluso peores que cuando los *outliers* proceden de medidas ruidosas. Si el número de *outliers* es muy pequeño normalmente se descartan. Sin embargo, si ese no es el caso y son resultados de una distribución con largas colas, entonces el diseñador debe adoptar funciones de coste que no sean muy sensibles a la presencia de *outliers* ya que los errores grandes dominan la función de coste debido a los términos cuadráticos [Huber, 1981].

3.2.2. Normalización de los datos

En muchas situaciones prácticas el diseñador de un clasificador debe enfrentarse a características cuyos valores se encuentran en rangos dinámicos diferentes. Así, características con valores altos pueden tener más influencia en la función de coste que características con valores pequeños, aunque esto no refleja necesariamente su significancia respectiva en el diseño del clasificador. El problema se soluciona normalizando las características de modo que se encuentren en rangos de valores similares. Una técnica sencilla es la normalización mediante las estimaciones de la media y la varianza. Para N datos

disponibles del k-ésimo vector n-dimensional de características, se tiene:

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \quad k = 1, 2, ..., n$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k^2) \tag{3.4}$$

$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

Las características normalizadas tendrán entonces media cero y varianza unidad. Esto es obviamente un método lineal. Otras técnicas lineales limitan los valores de las características a rangos entre $[0, 1]$ o $[-1, 1]$ mediante un escalado apropiado. Además de las técnicas lineales, existen métodos no lineales que se pueden emplear en casos en los que ni siquiera están distribuidos en torno a una media. En tales casos, transformaciones basadas en funciones no lineales (ver sección 8.1) pueden usarse para mapear los datos a intervalos específicos. Un candidato popular es el siguiente escalado:

$$y = \frac{x_{ik} - \bar{x}_k}{r\sigma_k}, \quad \hat{x}_{ik} = \frac{1}{1 + exp(-y)}, \tag{3.5}$$

Se trata de una función de escalado que limita los datos al rango $[0, 1]$. Usando una aproximación de series de expansión, es fácil ver que para pequeños valores de y se trata de una función aproximadamente lineal para x_{ik}. El rango de valores de x_{ik} que corresponde a la sección lineal depende de la desviación típica y del factor r, definido por el usuario. Valores lejanos de la media se escalan exponencialmente.

3.3. Extracción de características

Desde un punto de vista cuantitativo, se pretende seleccionar las características que conduzcan a una distancia entre clases grande y a una pequeña varianza intra-clase en el espacio del vector de características. Esto significa que las características deben estar distantes para clases diferentes y tomar valores cercanos para la misma clase. Para lograr este fin existen varios escenarios. Uno de ellos es evaluar las características individualmente y descartar aquéllas con pequeña capacidad discriminatoria [Chaves et al., 2009; Ramírez

et al., 2010; Salas-Gonzalez et al., 2009a,b]. Otra alternativa mejor es examinarlas en combinaciones [Ramírez et al., 2008]. A veces, la aplicación de una transformación lineal o no lineal al vector de características puede conducir a un nuevo vector con mejores propiedades discriminatorias [Álvarez et al., 2009b,a, 2010; López et al., 2009c,d].

La teoría de aprendizaje estadístico establece la habilidad de un clasificador para agrupar de manera efectiva N puntos de un espacio de alta dimensionalidad en dos clases diferentes [Cover, 1965]. Consideremos N puntos en un espacio de características n-dimensional. Se asume que los puntos están *bien distribuidos*, de manera que no existe ningún subconjunto de $N-1$ puntos que se sitúen en un hiperplano de dimensión menor que $n-1$. El número $O(l,N)$ de agrupaciones que pueden ser formadas por hiperplanos $(n-1)$-dimensionales para separar los N puntos en dos clases viene dado por [Cover, 1965]:

$$O(N,n) = 2\sum_{i=0}^{n} \binom{N-1}{i} \tag{3.6}$$

donde

$$\binom{N-1}{i} = \frac{(N-1)!}{(N-1-i)!i!} \tag{3.7}$$

Por lo tanto, la probabilidad de agrupar N puntos en un espacio de características n-dimensional en 2 clases linealmente separables será:

$$P_N^n = \frac{O(N,n)}{2^N} = \begin{cases} \frac{1}{2^{N-1}} \sum_{i=0}^{n} \binom{N-1}{i} & N > n+1 \\ 1 & N \leq n+1 \end{cases} \tag{3.8}$$

La Figura 3.1 muestra la probabilidad P_N^n como una función de n/N. Para espacios de características de baja dimensión ($n/N < 0{,}3$), la probabilidad de separación de clases P_N^n es prácticamente cero, esto es, los clasificadores lineales tienen un bajo rendimiento para discernir entre dos clases. Sin embargo, cuando se aumenta la dimensión del espacio de características, la probabilidad de que el conjunto de N puntos sea separable se aproxima a la unidad. Así, como intuitivamente parece claro, añadir información al vector de características mejora la separabilidad de clases para el caso de clasificador lineal.

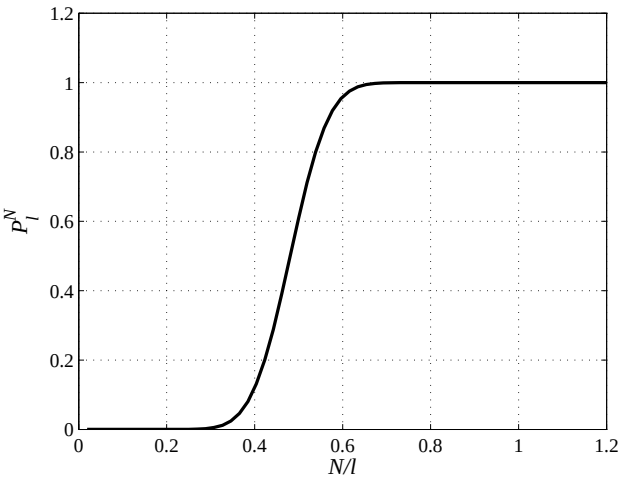


Figura 3.1: Probabilidad de agrupar N puntos en un espacio de características n-dimensional en 2 clases linealmente separables.

Por otro lado, el hecho de que un cociente n/N pequeño produce bajos valores para probabilidad de separabilidad puede resolverse para un número fijo N de puntos a clasificar mediante un mapeo a un espacio de dimensión mayor. Esto puede conseguirse a través del uso de *kernels*, que involucran el uso de un producto interno no lineal (ver sección 8.1), convirtiendo el espacio de características en otro de dimensión mayor donde P_N^n aumente, y un clasificador lineal pueda operar satisfactoriamente.

A la vista de la Figura 3.1 y de la ecuación 3.8 es natural pensar que un aumento de la dimensión del espacio de características será siempre beneficioso para discriminar entre dos clases. Paradójicamente, ocurre exactamente lo contrario, dando lugar al problema conocido como la maldición de la dimensionalidad (en inglés, *curse of dimensionality*) o el fenómeno del máximo (en inglés, *peaking phenomenon*). Este problema se describe como una reducción de la eficacia de un clasificador al añadir nuevas características a los vectores de entrenamiento cuando el número de éstos es relativamente pequeño en comparación con número de características. El problema radica en que para definir un clasificador en un espacio de características de alta dimensionalidad es necesario estimar un numero de parámetros comparable a la dimensión del espacio. Por ejemplo, en el caso de un clasificador lineal, será necesario estimar $n + 1$ parámetros en un espacio de características n-dimensional. Por lo tanto, aunque el clasificador separe los datos de entrenamiento satisfactoriamente, la fiabilidad en la estimación de los pará-

metros del clasificador será baja, ya que se estimarán muchos parámetros con un número muy reducido de vectores de entrenamiento. El clasificador construido con esta limitación tendrá, por consiguiente, una baja capacidad de generalización.

El problema de la maldición de la dimensionalidad justifica el uso de técnicas de reducción de la dimensionalidad del espacio de características, cuando el numero de características usadas para diseñar el clasificador es mucho mayor que el número de vectores de entrenamiento disponibles. Aunque éste es el principal motivo para hacerlo, existen otras motivaciones adicionales para reducir la dimensión del espacio de características hasta un mínimo razonable:

- La reducción del coste computacional de los algoritmos de entrenamiento y test.
- Eliminación de la correlación entre características.
- Selección de las características mas relevantes para la clasificación.

Así, aunque el problema de la maldición no exista, es natural suponer que el uso de un subconjunto de características con mejor capacidad de discriminación entre clases optimizará tanto el coste computacional del algoritmo de clasificación como el rendimiento del mismo.

El primer enfoque que examinaremos para la obtención del subconjunto de características para construir los vectores de entrenamiento y test introduce un proceso independiente con este fin, que ocurre antes de la categorización del vector de características. Por esta razón los podríamos denominar métodos de prefiltrado [John et al., 1994], ya que mediante ellos se descartan los atributos irrelevantes para la clasificación antes de que ésta tenga lugar. Este paso de preprocesamiento de los datos usa aspectos generales del conjunto de entrenamiento para seleccionar o extraer unas características y excluir otras. De esta manera, los métodos de prefiltrado no dependen del algoritmo de clasificación y podrán ser combinados con cualquiera de estos algoritmos, sin más que usar el subconjunto de características obtenido mediante filtrado para clasificación.

Remuestreo de la imagen

El método de filtro más sencillo que usaremos para reducir la dimensión del espacio de características en el caso de imágenes médicas será el de reducir

el tamaño de las imágenes mediante *subsampling* o *remuestreo*. Si la compresión de las imágenes no es elevada, éste es un método eficaz para disminuir la dimensión del espacio de características, obteniéndose un subconjunto de características que recoge prácticamente la misma información que los datos originales.

Descarte de voxels con baja intensidad

Es posible seleccionar un subconjunto de datos para clasificación mediante la construcción de una máscara binaria que extraiga de las imágenes aquellos voxels cuya intensidad sobrepase un límite prefijado. Una razón evidente para usar este método es que, fijando el umbral de selección adecuadamente, permitirá seleccionar aquellos voxels de la imagen tomográfica que representen regiones del cerebro, descartando todas aquellas regiones que quedan fuera del cerebro o cuya intensidad es muy baja y no contiene información útil para la clasificación. La definición exacta parte del cálculo de la imagen media del conjunto de datos. Considerando que tenemos un conjunto de imágenes cerebrales $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, ..., \mathbf{\Gamma}_N$, se define la imagen media como:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{\Gamma}_i \tag{3.9}$$

donde $\{\mathbf{\Gamma}_i \in \mathbb{R}^n\}$ es la imagen i muestral. Se define la máscara $\mathbf{E}$ binaria a partir de los valores ϵ_j:

$$\epsilon_j = \begin{cases} 1 & \text{si} \quad (\boldsymbol{\mu}_j > t) \\ 0 & \text{si} \quad (\boldsymbol{\mu}_j \leq t) \end{cases} \quad j = 1, ..., n \tag{3.10}$$

donde t es un umbral de intensidad fijado a priori. Con estos valores se puede construir la matriz de máscara $\mathbf{E}$ como:

$$\mathbf{E} = \text{diag}\,(\epsilon_1, \epsilon_2, ..., \epsilon_n) \tag{3.11}$$

De esta manera, la aplicación de la máscara $\mathbf{E}$ sobre un vector imagen $\mathbf{\Gamma}_i$:

$$\hat{\mathbf{\Gamma}}_i^T = \mathbf{\Gamma}_i^T \mathbf{E} \tag{3.12}$$

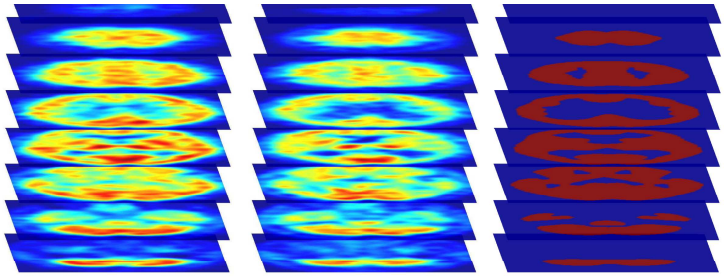


Figura 3.2: Secciones transversales de: *Columna izquierda*: Un paciente normal. *Columna central*: Un paciente con EA. *Columna derecha*: Máscara.

producirá un nuevo vector columna $\hat{\mathbf{\Gamma}}_i$. Este nuevo vector contendrá únicamente la información de aquellos voxels cuya intensidad promedio en el conjunto de muestras supere un valor fijado por t. Si t es pequeño se obtendrá una máscara que seleccione el interior del cerebro, y aumentando paulatinamente el valor de t, se irán descartando aquellas regiones en las que la intensidad promedio no sea muy alta. Este segundo caso queda reflejado en la Figura 3.2, y puede introducir mejoras para nuestros intereses, ya que las regiones cuya intensidad promedio sea baja, tanto en las imágenes de pacientes afectados por el Alzheimer como para los pacientes control, son regiones con menor capacidad de discriminación de la enfermedad de Alzheimer.

3.4. Medidas de separabilidad de clases

Entre los elementos de los vectores de características existe una inevitable correlación que influye en la capacidad de clasificación. En esta sección se proponen métodos para medir la efectividad discriminatoria de los vectores de características. Esta información se puede usar de dos modos. Por un lado, nos permite combinar características apropiadamente y terminar con "el mejor" vector para una dimensión dada n. Por otro lado, los datos se pueden transformar desde la base de un criterio óptimo que dé lugar a características de alto poder de clasificación [Theodoridis and Koutroumbas, 2003]. Algunas de estas medidas de separabilidad se presentan a continuación.

3.4.1. Divergencia

Dadas dos clases ω_1 y ω_2 y un vector de características $\mathbf{x}$, la regla de Bayes (ver Sección 4.2) selecciona ω_1 si:

$$P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \tag{3.13}$$

El factor $\frac{P(\omega_1|\mathbf{x})}{P(\omega_2|\mathbf{x})}$ puede llevar información útil acerca de la capacidad discriminatoria asociada con un cierto vector de características $\mathbf{x}$, con respecto a las dos clases ω_1 y ω_2. Alternativamente, dados los valores $P(\omega_1)$, $P(\omega_2)$, la misma información reside en la razón $\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \equiv D_{12}(\mathbf{x})$ y puede usarse como medida de la información discriminante existente de la clase ω_1 con respecto a ω_2. Claramente, para clases completamente superpuestas se obtiene $D_{12}(\mathbf{x}) = 0$. Puesto que $\mathbf{x}$ toma diferentes valores, es lógico considerar el valor medio sobre la clase ω_1

$$D_{12} = \int_{-\infty}^{+\infty} p(\mathbf{x}|\omega_1) \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} d\mathbf{x} \tag{3.14}$$

Del mismo modo, se define para la clase ω_2

$$D_{21} = \int_{-\infty}^{+\infty} p(\mathbf{x}|\omega_2) \ln \frac{p(\mathbf{x}|\omega_2)}{p(\mathbf{x}|\omega_1)} d\mathbf{x} \tag{3.15}$$

La suma

$$d_{12} = D_{12} + D_{21} \tag{3.16}$$

se conoce como *divergencia* y puede usarse como medida de separabilidad para las clases ω_1 y ω_2, con respecto al vector dado $\mathbf{x}$. La divergencia tiene las siguientes sencillas propiedades:

$$\begin{array}{l} d_{ij} \geq 0 \\ d_{ij} = 0 \quad \text{si } i = j \\ d_{ij} = d_{ji} \end{array} \tag{3.17}$$

Si las componentes del vector de características son estadísticamente inde-

pendientes, entonces se puede demostrar que:

$$d_{ij}(x_1, x_2, ..., x_n) = \sum_{r=1}^{n} d_{ij}(x_r) \tag{3.18}$$

Asumiendo ahora que las funciones de densidad de probabilidad son Gaussianas $\mathcal{N}(\mu_i, \Sigma_i)$ y $\mathcal{N}(\mu_j, \Sigma_j)$ respectivamente, el cálculo de la divergencia se simplifica y no es difícil demostrar que

$$d_{ij} = \frac{1}{2}\text{traza}\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i - 2l\} + \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{3.19}$$

Para el caso unidimensional, se convierte en

$$d_{ij} = \frac{1}{2}\left(\frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2} - 2\right) + \frac{1}{2}(\mu_i - \mu_j)^2\left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right) \tag{3.20}$$

Como ya se ha señalado, esta medida de separabilidad no depende únicamente de la diferencias entre los valores medios, sino que también es dependiente de las varianzas. Además, d_{ij} puede ser alto incluso con valores medios iguales, siempre y cuando las varianzas difieran de forma significante. Así, la separación de clases es posible incluso cuando las medias son iguales.

Indagando un poco más en la ecuación 3.19, si las matrices de covarianza de las dos distribuciones Gaussianas son iguales, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$, entonces la divergencia se simplifica aún más:

$$d_{ij} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{3.21}$$

que no es más que la distancia de Mahalanobis entre los correspondientes vectores medios. Además esto tiene otra implicación interesante: en este caso se tiene una relación directa entre la divergencia d_{ij} y el error de Bayes, es decir, el mínimo error que se puede alcanzar cuando se adopta un específico vector de características. Ésta es la propiedad más deseable para cualquier medida de separabilidad. Desafortunadamente, dicha relación de la divergencia con el error de Bayes no es posible para otras distribuciones más generales. Además, en Swain and King [1973]; Richards [1995] se señala que la dependencia específica de la divergencia de los vectores medios puede llevar a confusión, en el sentido de que pequeñas variaciones en dicha diferencia

pueden dar lugar a cambios grandes en la divergencia, los cuales sin embargo no se vean reflejados en el error de clasificación.

3.4.2. Matrices de dispersión

Una desventaja de la medida de separabilidad propuesta en la sección anterior es que no es fácilmente computable a menos que se asuman distribuciones Gaussianas. Aquí se propone un criterio más sencillo formado a partir de la información relacionada con el modo en el que las muestras de los vectores de características están dispersas en el espacio n-dimensional. Con ese fin se definen las siguientes matrices:

Matriz de dispersión intra-clase

$$\mathbf{S}_w = \sum_{i=1}^{c} P_i \mathbf{S}_i \tag{3.22}$$

donde c es el número de clases y $\mathbf{S}_i$ es la matriz de covarianza para la clase i

$$\mathbf{S}_i = E[(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T] \tag{3.23}$$

y P_i es la probabilidad *a priori* de la clase i. Es decir, $P_i \approx N_i/N$, donde N_i es el número de muestras de la clase ω_i del total de muestras, N. Obviamente, la traza de $\mathbf{S}_w$ es una medida de la media de la varianza de las características sobre todas las clases.

Matriz de dispersión inter-clase

$$\mathbf{S}_b = \sum_{i=1}^{c} P_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^T \tag{3.24}$$

donde μ_0 es el vector media global

$$\boldsymbol{\mu}_0 = \sum_{i}^{c} P_i \boldsymbol{\mu}_i \tag{3.25}$$

La traza de $\mathbf{S}_b$ es una medida de la distancia promediada sobre todas las clases entre la media de cada clase individual y el respectivo valor medio global.

Matriz de dispersión mixta

$$\mathbf{S}_m = E[(\mathbf{x} - \boldsymbol{\mu}_0)(\mathbf{x} - \boldsymbol{\mu}_0)^T] \tag{3.26}$$

es decir, $\mathbf{S}_m$ es la matriz de covarianza del vector de características con respecto a la media global. No es difícil demostrar que

$$\mathbf{S}_m = \mathbf{S}_w + \mathbf{S}_b \tag{3.27}$$

Su traza es la suma de las varianzas de las características en torno a la respectiva media global. De estas definiciones es fácil ver que el criterio

$$J_1 = \frac{\text{traza}\{\mathbf{S}_m\}}{\text{traza}\{\mathbf{S}_w\}} \tag{3.28}$$

toma valores mayores cuando las muestras en el espacio n-dimensional están agrupadas en torno a sus medias, dentro de cada clase, y las agrupaciones de diferentes clases están bien separadas. A veces $\mathbf{S}_b$ se usa en lugar de $\mathbf{S}_m$. Otro criterio alternativo surge si se usan los determinantes en lugar de las trazas. Esto se justifica para matrices de dispersión simétricas y definidas positivas de modo que sus autovalores sean positivos. La traza es igual a la suma de los autovalores, mientras que el determinante es el producto de los autovalores. Por tanto, valores altos de J_1 también se corresponde con valores altos de

$$J_2 = \frac{|\mathbf{S}_m|}{|\mathbf{S}_w|} = |\mathbf{S}_w^{-1}\mathbf{S}_m| \tag{3.29}$$

Una variante de J_2 comúnmente usada es

$$J_3 = \text{traza}\{\mathbf{S}_w^{-1}\mathbf{S}_m\} \tag{3.30}$$

Los criterios J_2 y J_3 tienen la ventaja de ser invariantes bajo transformaciones lineales. En Fukunaga [1990] se definen otros criterios mediante varias combinaciones de $\mathbf{S}_w$, $\mathbf{S}_b$ y $\mathbf{S}_m$ con formulaciones de "traza" o "determinante".

Sin embargo, si se usa el determinante, hay que tener especial cuidado con $\mathbf{S}_b$, puesto que $|\mathbf{S}_b| = 0$ para $c < n$. Esto es debido a que $\mathbf{S}_b$ es la suma de c matrices de dimensión $n \times n$ de rango 1 cada una.

Estos criterios toman una forma especial en el caso unidimensional de dos clases. En este caso, es fácil ver que para clases equiprobables $|\mathbf{S}_w|$ es proporcional a $\sigma_1^2 + \sigma_2^2$ y $|\mathbf{S}_b|$ proporcional a $(\mu_1 - \mu_2)^2$. Combinando $\mathbf{S}_b$ y $\mathbf{S}_w$ resulta el así llamado criterio discriminante de Fisher (FDR de sus siglas en inglés Fisher Discriminant Ratio):

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{3.31}$$

FDR a veces se usa para cuantificar las capacidades de separabilidad de características individuales.

3.5. Parámetros de valoración del rendimiento de un clasificador

El objetivo último de un clasificador consiste en asignar correctamente una etiqueta a un patrón de test. En el caso de clasificación binaria, en la que los datos están divididos entre etiquetas positivas o negativas, existirán dos posibles errores que el clasificador puede cometer: clasificar como positivo un patrón que en realidad era negativo o viceversa. Estas posibilidades vienen reflejadas en la Tabla 3.1, donde se compara el resultado del test con la etiqueta original, dando lugar a positivo verdadero o TP (True Positive) cuando los dos coinciden en valor positivo, falso positivo o FP (False Positive) cuando el test da positivo siendo la etiqueta original negativa, falso negativo o FN (False Negative) cuando el test da negativo siendo la etiqueta original positiva y negativo verdadero o TN (True Negative) cuando los dos coinciden en valor negativo.

La capacidad de un clasificador para detectar los positivos verdaderos se mide a través de la sensibilidad, que se define como:

$$\text{Sensibilidad} = \frac{TP}{TP + FN} \tag{3.32}$$

de manera que una sensibilidad del 100 % corresponderá a un clasificador

		Etiqueta		
		Positiva	Negativa	
Test	Positivo	TP	FP	→ *Valor predictivo positivo*
	Negativo	FN	TN	→ *Valor predictivo negativo*
		↓	↓	
		Sensibilidad	*Especificidad*	

Tabla 3.1: Posibles resultados del test en función de la etiqueta.

que es capaz de detectar todos los pacientes etiquetados como positivos como tales. Por lo tanto, si un clasificador con alta sensibilidad da un resultado negativo, éste será muy fiable, lo que puede ser usado para descartar la enfermedad. La sensibilidad esta relacionada con el error de tipo I en inferencia estadística, que consiste en rechazar la hipótesis nula cuando en realidad es cierta.

Por otro lado, la capacidad para detectar negativos verdaderos vendrá dada por la especificidad, definida como:

$$\text{Especificidad} = \frac{TN}{TN + FP} \tag{3.33}$$

permitiendo que un clasificador con alta especificidad sea muy fiable a la hora de confirmar la enfermedad, ya que raramente producirá un resultado positivo que en realidad sea falso. La especificidad esta relacionada con el error de tipo II donde se acepta la hipótesis nula cuando en realidad es falsa.

Sin embargo, valores altos tanto de la sensibilidad como de la especificidad no tienen porque corresponder a un clasificador preciso. Se define la precisión como:

$$\text{Precisión} = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.34}$$

Puede ocurrir que un clasificador tenga valores cercanos al 100 % de sensibilidad y cercanos al 0 % de especificidad. Este clasificador no tendrá capacidad de discernir entre las clases, ya que será un clasificador que tome cualquier patrón como positivo. Esto es equivalente a una clasificación al azar, ya que su precisión rondará el 50 % para una muestra sin preponderancia de ninguna de las dos clases. El clasificador deseable será aquel que tenga

valores altos de sensibilidad, especificidad y precisión simultáneamente, y no solo de alguno de ellos por separado.

Otros parámetros que pueden resultar interesantes son los valores predictivos. Éstos hacen referencia a la validez de un resultado de clasificación positivo/negativo (valor predictivo positivo/negativo). Se podrá confiar más en un resultado positivo de un clasificador con un vpp alto que uno con un vpp menor. Sin embargo, los valores predictivos dependen de la preponderancia de las clases, denominada prevalencia. Éste es un término de epidemiología que determina la proporción de individuos de una población que, en este caso, padece la enfermedad. Si el conjunto de test no tiene igual número de positivos que de negativos, habrán de usarse las fracciones de probabilidad positiva o negativa (fpp/fpn):

$$\text{fpp} = \frac{\text{sensibilidad}}{1 - \text{especificidad}} \tag{3.35}$$

y

$$\text{fpn} = \frac{1 - \text{sensibilidad}}{\text{especificidad}} \tag{3.36}$$

que no dependen de la prevalencia.

3.5.1. Curva ROC

A menudo será interesante valorar cómo se modifica el rendimiento de un clasificador al modificar algún parámetro, ya sea del clasificador o en la definición de algún paso anterior. Para este análisis será útil la representación en el espacio ROC (receiver operating characteristic (ROC) [Fawcett, 2006]), que no es más que una representación bidimensional de la tasa de positivos verdaderos (sensibilidad) frente a la tasa de falsos positivos (1-especificidad).

En este espacio (véase la Figura 3.3), la mejor predicción corresponderá con un punto cercano a la esquina superior izquierda (D), representando sensibilidad del 100 % (ningún negativo falso) y especificidad 100 % (ningún positivo falso), que producirá tambien una precisión del 100 %. El punto D se llama tambien clasificación perfecta. Una clasificación completamente aleatoria daría un punto C en la línea diagonal (llamada la línea de no-discriminación) de la parte inferior izquierda a la parte superior derecha.

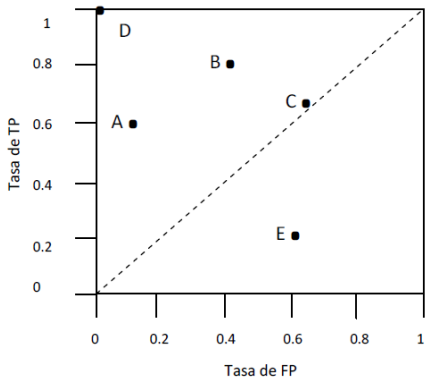


Figura 3.3: Representación en el espacio ROC

Por debajo de esta línea estarían resultados peores que la estimación al azar E. Los puntos A y B representan clasificaciones adecuadas, la primera con mayor sensibilidad que espeficicidad y la segunda con una especificidad mayor que la sensibilidad.

3.6. Métodos de validación cruzada

La Validación cruzada, a veces denominada estimación por rotación [Kohavi and John, 1995; Devijver and Kittler, 1982], es una técnica muy popular para evaluar cómo el resultado de un análisis estadístico tiene mayor o menor capacidad de generalización sobre un conjunto independiente de datos. Se usa preferentemente en aplicaciones donde el objetivo último es la predicción, y se emplea para estimar cómo de preciso es un modelo predictivo en la práctica. Una ronda de validación cruzada consiste en dividir el conjunto muestral en subconjuntos complementarios, y realizar el análisis en un subconjunto (llamado el conjunto de entrenamiento), y validar este análisis en otro subconjunto (llamado subconjunto de validación o de test). Para reducir la variabilidad en la evaluación global de la generalización, se llevan a cabo múltiples rondas de validación cruzada usando distintas particiones, y los resultados de la validación son promediados sobre todas las rondas realizadas.

En los métodos de validación cruzada, el conjunto de test no supone

un test 'real', ya que la etiqueta de los elementos del conjunto de test es conocida. De esta manera, se puede comparar el resultado del test con la etiqueta original, y determinar si se trata de un TP, FP, TN, o FN. Una vez el proceso de itera sobre cada partición, se puede calcular cualquiera de las cantidades definidas en la sección anterior.

La teoría de la Validación Cruzada fue originalmente desarrollada por Seymour Geisser y es de radical importancia para vigilar la posible presencia del error estadístico de tipo III en cualquier proceso de decisión. Este tipo de error consiste en rechazar la hipótesis nula de manera correcta por razones erróneas, que ocurre por ejemplo cuando el tamaño muestral es limitado. En las siguientes secciones se detallan algunos de estos métodos de validación.

3.6.1. Validación por sub-muestreo aleatorio repetido

Este método divide aleatoriamente el conjunto de datos en dos conjuntos de entrenamiento y validación. Para cada división, el clasificador es entrenado con el conjunto de entrenamiento y validado sobre los datos restantes. Los resultados de cada división son promediados. La ventaja de este método es que la proporción de la división entrenamiento/validación no depende del número de iteraciones. Sin embargo, presenta una desventaja y es que algunas muestras puede nunca ser seleccionadas en el subconjunto de validación, mientras que otras pueden ser seleccionadas más de una vez. Es decir, los conjuntos de validación pueden solaparse. Este método también exhibe la variación Monte Carlo, esto es, los resultados variarán si el análisis se repite con diferentes conjuntos aleatorios. Una variante de esta aproximación genera muestras aleatorias de tal forma que el valor de respuesta medio es igual en los subconjuntos de entrenamiento y test. Esto es particularmente útil cuando el conjunto muestral contiene una representación no balanceada en las respuestas de las muestras.

3.6.2. Validación cruzada K-pliegues

En la validación K-pliegues, el conjunto muestral original se divide en K subconjuntos [Breiman et al., 1984]. De los K subconjuntos, uno de ellos se destina a validación para testar el modelo y los $K-1$ restantes se usan como conjunto de entrenamiento. Después la validación cruzada se repite K veces (los pliegues), con cada uno de los K subconjuntos usados una vez como datos de validación. Los K resultados de cada pliegue son promediados (o

combinados) para producir una única estimación.

La ventaja de este método sobre el anterior es que todas las observaciones se usan tanto para entrenamiento como para test y cada observación se usa para validación sólo una vez. Este método se suele usar cuando el número de elementos de la muestra es muy grande, o cuando los algoritmos de clasificación son computacionalmente costosos, de manera que se tiene control sobre el número de veces que se itera la validación a través del número K. En este caso también puede considerarse que en cada pliegue se contenga la misma proporción de etiquetas o respuestas.

3.6.3. Validación Leave-One-Out

La validación Leave-One-Out consiste en utilizar una sola muestra como observación de test y las restantes observaciones como entrenamiento [Raudys and Jain, 1991]. Este proceso se repite N veces, siendo N el número de muestras disponibles, de manera que todas las muestras son usadas una vez como observaciones para la validación, por lo que se considera un caso particular del anterior método de validación. La ventaja del método de dejar uno fuera es que el conjunto de entrenamiento es lo más grande que la muestra permite, aumentando la estadística en la estimación de los parámetros del clasificador. Por contra, este método puede ser computacionalmente costoso dado el gran número de veces que se repite el proceso de validación, si el número de elementos de la muestra es grande.

En el caso de diagnóstico médico, éste será a menudo el método de validación elegido, ya que el número de pacientes de una base de datos habitual es suficientemente reducido para que el uso de la validación dejar uno fuera no suponga un gran coste computacional.

CAPÍTULO 4

Métodos de Clasificación

Uno de los problemas más habituales en ciencias y matemáticas es el problema de *clasificación* de objetos en un dominio particular, es decir, separar dichos objetos en clases más pequeñas y proporcionar los criterios para determinar si un objeto de dicho dominio pertenece a una de esas clases o no. Dependiendo de la aplicación, estos objetos pueden ser imágenes, formas de onda o cualquier otro tipo de medida que requiera ser clasificada. En muchos casos estos objetos se denominan *patrones*, refiriéndose entonces como *reconocimiento de patrones* a la tarea de clasificarlos. El campo de reconocimiento de patrones se ocupa por tanto del descubrimiento automático de regularidades en los datos a través del uso de algoritmos computarizados y hace uso de esas regularidades para clasificar los datos en categorías. En este capítulo se presentan las bases teóricas de los métodos de clasificación empleados en la tarea de categorización de imágenes neurológicas.

4.1. Linealidad de un clasificador

El objetivo de la clasificación es tomar un vector de entrada $\mathbf{x}$ y asignarle una de las c clases discretas ω_k donde $k = 1, ..., c$. El escenario de clasificación más común considera que las clases son disjuntas, es decir, a cada entrada se le asigna una única clase. El espacio de entrada por tanto se divide en regiones de decisión cuyas fronteras se denominan *fronteras de decisión* o *superficies de decisión*. Los modelos lineales de decisión para un vector de entrada $\mathbf{x}$ están definidos por hiperplanos de dimensión $(n-1)$, siendo n la dimensión del espacio de entrada. Cuando los conjuntos de datos pueden ser separados exactamente por superficies lineales se dicen que son *linealmente separables*.

La principal diferencia entre clasificadores lineales y no lineales es que para los clasificadores lineales, su función de decisión $g(\mathbf{x})$ se basa en una combinación lineal de los vectores de entrada

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{4.1}$$

donde $\mathbf{w} = [w_1 w_2 ... w_n]$ se conoce como el vector pesos constante, y w_0 como el umbral. Para clasificadores no lineales, las funciones de decisión no dependen únicamente de combinaciones lineales de los vectores de características, y normalmente estas funciones son más difíciles de optimizar aunque también más precisas en la clasificación.

4.2. Criterio de Bayes

En un test de dos hipótesis, la regla de decisión óptima que minimiza la probabilidad de error es el conocido clasificador de Bayes. Los clasificadores basados en las reglas bayesianas se han usado en diferentes problemas de reconocimiento de caras [Shakhnarovich and Moghaddam, 2004; Wang and Tang, 2003] y clasificación de imagen neurológica [López et al., 2009b]. El clasificador de Bayes evalúa la función de probabilidad *a posteriori* [Fukunaga, 1990]. Sea $\{\omega_1, \omega_2, ..., \omega_c\}$ las clases y $\mathbf{x}$ el vector de características de entrada. La función de probabilidad *a posteriori* de ω_i dada $\mathbf{x}$ se define como

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}, \quad i = 1, 2, ..., c. \tag{4.2}$$

donde $p(\mathbf{x}|\omega_i)$ es la función de densidad de probabilidad condicional de $\mathbf{x}$ dada ω_i, $P(\omega_i)$ es la probabilidad *a priori* y $P(\mathbf{x})$ la densidad conjunta. La regla de decisión *maximum a posteriori* (MAP) para el clasificador Bayesiano se define como

$$p(\mathbf{x}|\omega_i)P(\omega_i) = \max_j\{p(\mathbf{x}|\omega_j)P(\omega_j)\}, \quad \mathbf{x} \in \omega_i \tag{4.3}$$

es decir, al vector de test $\mathbf{x}$ se le asigna la clase ω_i cuya probabilidad a posterior dado $\mathbf{x}$ sea la mayor de entre todas las clases. Para el problema de clasificación de imágenes neurológicas que nos concierne, la probabilidad *a priori* $P(\omega_i)$ se establece inicialmente al valor 0,5, es decir, la probabilidad inicial del test de pertenecer a una de las dos clases AD (Alzheimer) y NORMAL es igualmente probable. En cuanto a la función de densidad de probabilidad condicional, habitualmente no existen muestras suficientes para estimarla para cada clase (densidad intra-clase). Se suele optar por el compromiso de asumir una forma de densidad particular y convertir por tanto la estimación general de la densidad en un problema paramétrico. Las densidades intra-clase suelen ser modeladas como distribuciones normales:

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_i|^{1/2}} \times exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{4.4}$$

donde μ_i y $\mathbf{\Sigma}_i$ son la media y la matriz de covarianza de las muestras de la clase ω_i, respectivamente.

La Figura 4.1 muestra datos pertenecientes a dos clases y la frontera de decisión definida por la regla de Bayes. Las distribuciones de densidad de las muestras se asumen como Gaussianas. Las elipses alrededor de las agrupaciones muestran las líneas de igual densidad de probabilidad de la Gaussiana. Considerando las probabilidades *a priori* iguales para los datos de la clase ω_1 (representados en la figura con puntos azules) y la clase ω_2 (representados por cruces rojas), la línea verde divide el plano 2D en dos regiones, de modo que una muestra situada encima o debajo de dicha línea es más probable que pertenezca a la clase ω_1 o a la clase ω_2, respectivamente.

En el caso de dos clases, supongamos que las covarianzas de las clases son idénticas. Entonces reescribiendo la ecuación 4.2 se obtiene que

$$P(\omega|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + \omega_0) \tag{4.5}$$

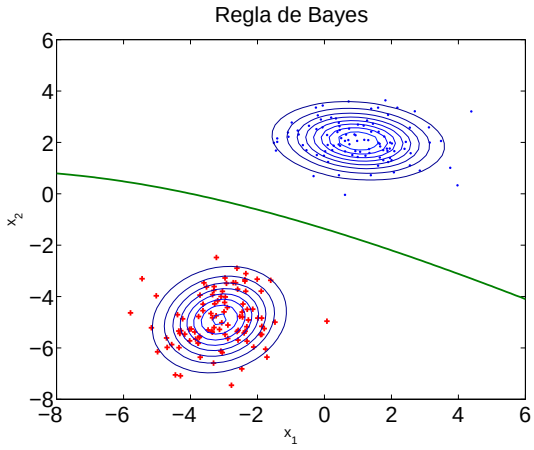


Figura 4.1: Línea de decisión diseñada por un clasificador bayesiano para dos clases cuyas funciones de distribución de densidad se asumen como Gaussianas.

siendo

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{4.6}$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{P(\omega_1)}{P(\omega_2)} \tag{4.7}$$

y $\sigma(a)$ es la función *sigmoide logística* y se define como sigue

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{4.8}$$

con

$$a = \ln\frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x}|\omega_2)P(\omega_2)} \tag{4.9}$$

Se puede ver que los términos cuadráticos en $\mathbf{x}$ de los exponentes de las densidades Gaussianas se han cancelado (debido a la asunción de matrices de covarianza comunes) dando lugar a una función lineal de $\mathbf{x}$ en el argumento de la sigmoide logística. Las probabilidades *a priori* $P(\omega_i)$ afectan sólo al parámetro ω_0 de modo que cambios en estas probabilidades a priori tienen el efecto de desplazar paralelamente la frontera de decisión.

Si no se asume una matriz de covarianza compartida, y se permite a cada densidad condicional de clase $p(\mathbf{x}|\omega_i)$ que tenga su propia matriz $\mathbf{\Sigma}_i$, entonces las cancelaciones de antes ya no se dan y se obtienen funciones cuadráticas de $\mathbf{x}$, dando lugar al *discriminante cuadrático*.

4.3. Máquinas de Vectores de Soporte (SVM)

Desde su introducción a finales de los años 70 [Vapnik, 1982], las máquinas de vectores de soporte (SVMs) marcaron el comienzo de una nueva era en el paradigma del aprendizaje a partir de ejemplos [Burges, 1998; Joachims, 1998]. SVM ha atraído recientemente la atención de la comunidad dedicada al reconocimiento de patrones debido a la cantidad de méritos derivados de la Teoría del Aprendizaje Estadístico [Vapnik, 1995, 1998] desarrollada por Vladimir Vapnik en AT&T. Estas técnicas se han usado en una gran cantidad de aplicaciones incluyendo la detección de actividad de voz (VAD) [Enqing et al., 2002a,b; Qi et al., 2004; Ramírez et al., 2006b,a; Yélamos et al., 2006], recuperación de imágenes basadas en contenido [Tao et al., 2006], clasificación de texturas [Kim et al., 2002], predicción de series temporales [Górriz et al., 2004] y diagnosis de imágenes médicas [Fung and Stoeckel, 2007; Kalatzis et al., 2003; Salas-Gonzalez et al., 2009d, 2010; López et al., 2009b].

4.3.1. SVM lineal

Una de las mayores ventajas de los clasificadores lineales es su simplicidad y atractivo computacional. En este apartado supondremos que todos los vectores de características de las clases disponibles pueden clasificarse correctamente usando un clasificador lineal. Más adelante nos centraremos en problemas más genéricos donde los clasificadores lineales no pueden clasificar correctamente todos los vectores, y trataremos de buscar modos de diseñar un clasificador óptimo lineal adoptando un criterio de optimización apropiado.

Clases linealmente separables

Sean $\mathbf{x}_i$, i=1,2,...,N, los vectores de características del conjunto de entrenamiento, X. Estos pertenecen a una de las dos clases ω_1 o ω_2 , las cuales se asumen linealmente separables. El objetivo, una vez más, es diseñar el hiperplano de la ecuación 4.1, $g(\mathbf{x}) = 0$, que clasifica correctamente todos

los vectores de entrenamiento. Dicho hiperplano no es único y el proceso de selección se centra en maximizar la generalización del clasificador. De entre los posibles criterios, se selecciona el de la búsqueda del hiperplano que maximiza el margen entre clases, conocido como el hiperplano de margen máximo.

Todo hiperplano está caracterizado por su dirección (determinada por $\mathbf{w}$) y su posición exacta en el espacio (determinada por w_0). Puesto que no se pretende dar preferencia a ninguna de las dos clases, es razonable elegir para todas las direcciones el hiperplano que dista lo mismo respectivamente de los puntos más cercanos en ω_1 y ω_2. La distancia desde un punto $\mathbf{x}$ al hiperplano viene dada por:

$$z = \frac{|g(\mathbf{x})|}{\sqrt{w_1^2 + w_2^2}} \tag{4.10}$$

escalamos $\mathbf{w}$ y w_0 de modo que el valor de $g(\mathbf{x})$ en los puntos más cercanos sea $+1$ para el punto más cercano en ω_1 y -1 para el punto más cercano en ω_2. Esto es equivalente a

- Tener un margen de $\frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$
- Con las condiciones

$$\begin{array}{ll} \mathbf{w}^T\mathbf{x} + w_0 \geq 1, & \forall \mathbf{x} \in \omega_1, \\ \mathbf{w}^T\mathbf{x} + w_0 \leq 1, & \forall \mathbf{x} \in \omega_2, \end{array} \tag{4.11}$$

Para cada $\mathbf{x}_i$ denotamos el correspondiente indicador de clase y_i ($+1$ para ω_1 y -1 para ω_2). El objetivo por tanto es calcular los parámetros w_i y w_0 del hiperplano de manera que se minimice la siguiente expresión

$$\mathbf{J}(\mathbf{w}) \equiv \frac{1}{2} \|\mathbf{w}\|^2 \tag{4.12}$$

sujeto a

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 \quad \text{para} \quad i = 1, 2, ..., N \tag{4.13}$$

Obviamente, minimizando la norma el margen se hace mínimo. Esto es una tarea de optimización (cuadrática) no lineal sujeto a un conjunto de res-

tricciones de inecuaciones lineales. Las condiciones de Karush-Kuhn-Tucker (KKT) establecen que ha de cumplirse lo siguiente:

$$\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0} \tag{4.14}$$

$$\frac{\partial}{\partial w_0}\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) = 0 \tag{4.15}$$

$$\lambda_i \geq 0 \quad \text{para} \quad i = 1, 2, ..., N \tag{4.16}$$

$$\lambda_i \left[y_i(\mathbf{w}^T \mathbf{x}_i + w_0) - 1 \right] = 0 \quad \text{para} \quad i = 1, 2, ..., N \tag{4.17}$$

donde λ es el vector de multiplicadores de Lagrange, λ_i, y $\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda})$ es la función Lagrangiana definida como

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i \left[y_i(\mathbf{w}^T \mathbf{x}_i + w_0) - 1 \right] \tag{4.18}$$

Combinando 4.18 con 4.14 y 4.15 resulta

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \tag{4.19}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{4.20}$$

Los multiplicadores de Lagrange pueden ser cero o positivos. Por tanto, el vector de parámetros $\mathbf{w}$ de la solución óptima es una combinación lineal de $N_s \leq N$ vectores características asociados a $\lambda_i \neq 0$, es decir,

$$\mathbf{w} = \sum_{i=1}^{N_s} \lambda_i y_i \mathbf{x}_i \tag{4.21}$$

Éstos se conocen como los *vectores de soporte* y el hiperplano clasificador óptimo *máquina de vectores de soporte* (SVM). Al igual que para el conjunto

de restricciones en 4.17 para $\lambda_i \neq 0$, los vectores de soporte caen en uno de los dos hiperplanos

$$\mathbf{w}^T\mathbf{x} + w_0 = \pm 1 \tag{4.22}$$

es decir, son los vectores de entrenamiento que están más cerca del clasificador lineal, y constituyen los elementos críticos del conjunto de entrenamiento.

Aunque $\mathbf{w}$ se da explícitamente, w_0 se puede obtener implícitamente por una de las condiciones 4.17. En la práctica, w_0 se calcula como un valor medio obtenido usando todas las condiciones de este tipo. Por otro lado, las propiedades de la función de coste 4.12 garantizan que la matriz Hessiana correspondiente es definida positiva. Además, las restricciones consisten en funciones lineales. Estas dos condiciones garantizan que cualquier mínimo local es también global y único. El hiperplano clasificador de una máquina de vectores de soporte es único.

Habiendo establecido todas estas propiedades interesantes del hiperplano óptimo de una máquina de vectores de soporte, el siguiente paso es el cálculo de los parámetros involucrados. Desde un punto de vista computacional esto no siempre es una tarea fácil, y existen numerosos algoritmos para ello. Se trata de un problema de la familia de programación convexa. Estos problemas se resuelven considerando la denominada *dualidad Lagrangiana*, y el problema puede formularse equivalentemente como sigue

Maximizar

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\lambda}) \tag{4.23}$$

sujeto a

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \tag{4.24}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{4.25}$$

$$\boldsymbol{\lambda} \geq \mathbf{0} \tag{4.26}$$

Las dos restricciones de igualdad son el resultado de igualar a cero el gradiente de la Lagrangiana, con respecto a $\mathbf{w}$ y w_0. Los vectores de características de entrenamiento aparecen en el problema mediante las restricciones de igualdad y no mediante las inecuaciones, lo cual hace más fácil de manejar. Sustituyendo 4.24 y 4.25 en 4.23 y haciendo algunas operaciones se llega a la tarea de optimización equivalente

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right) \tag{4.27}$$

sujeto a

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{4.28}$$

$$\boldsymbol{\lambda} \geq \mathbf{0} \tag{4.29}$$

Una vez que los multiplicadores de Lagrange han sido calculados, maximizando 4.46, el hiperplano óptimo se calcula vía 4.24 y w_0 como antes.

Además de ser una manera más cómoda, existe otra razón por la cual se opta por la formulación de 4.28. Los vectores de entrenamiento aparecen en parejas, en la forma de productos escalares. Esto es lo más interesante, porque la función coste no depende explícitamente de la dimensionalidad del espacio de entrada. Esta propiedad permite generalizaciones eficientes para el caso de clases linealmente no separables.

Clases linealmente no separables

En el caso de que las clases no sean linealmente separables, lo dicho anteriormente deja de ser válido. Recordando que el margen se define como la distancia entre el par de hiperplanos paralelo descritos por

$$\mathbf{w}^T \mathbf{x} + w_0 = \pm 1 \tag{4.30}$$

Los vectores de características de entrenamiento ahora pertenecen a una de las siguientes tres categorías:

- Vectores que caen fuera de la banda y que son correctamente clasificados. Estos vectores cumplen con las restricciones en 4.12.

- Vectores que caen dentro de la banda y que son correctamente clasificados. Estos satisfacen la inecuación

$$0 \leq y_i(\mathbf{w}^T\mathbf{x} + w_0) < 1 \tag{4.31}$$

- Vectores que son clasificados erróneamente. Éstos cumplen la inecuación

$$y_i(\mathbf{w}^T\mathbf{x} + w_0) < 0 \tag{4.32}$$

Estos tres casos se pueden tratar como un solo tipo de restricciones introduciendo un nuevo conjunto de variables

$$y_i(\mathbf{w}^T\mathbf{x} + w_0) \geq 1 - \xi_i \tag{4.33}$$

La primera categoría de datos corresponde con $\xi_i = 0$, la segunda con $0 < \xi_i \leq 1$ y la tercera con $\xi_i > 1$. Las variables ξ_i se conocen como *variables débiles*. La tarea de optimización se vuelve más complicada aunque se basa en los mismos principios que antes. El objetivo ahora es hacer el margen tan grande como sea posible pero al mismo tiempo mantener la cantidad de puntos con $\xi_i \geq 0$ tan pequeña como sea posible. En términos matemáticos, esto es equivalente a minimizar la función de coste

$$J(\mathbf{w}, w_0, \xi) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\mathbf{I}(\xi_i), \tag{4.34}$$

donde $\boldsymbol{\xi}$ es el vector de parámetros ξ_i y

$$\mathbf{I}(\xi_i) = \begin{cases} 1, & \xi_i > 0 \\ 0, & \xi_i = 0 \end{cases} \tag{4.35}$$

El parámetro C es una constante positiva que controla la influencia relativa de los dos términos competitivos. Sin embargo, la optimización de arriba es difícil puesto que incluye una función discontinua $\mathbf{I}(\cdot)$. Como es común en

casos así, se elige optimizar una función de coste estrechamente relacionada, y el objetivo se convierte en minimizar

$$J(\mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i, \tag{4.36}$$

sujeto a

$$\begin{array}{cc} y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1 - \xi_i, & i = 1, 2, ..., N \\ \xi_i > 0, & i = 1, 2, ..., N \end{array} \tag{4.37}$$

El problema es de nuevo un problema de programación convexa, y la Lagrangiana correspondiente viene dada por

$$\mathcal{L}(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i[y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 + \xi_i] \tag{4.38}$$

Realizando pasos similares a los del caso de clases separables, llegamos al siguiente problema de optimización equivalente

$$\max_{\lambda}(\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j) \tag{4.39}$$

sujeto a

$$0 \leq \lambda_i \leq C \quad \text{para} \quad i = 1, 2, ..., N \tag{4.40}$$

$$\sum_{i=1}^{N}\lambda_i y_i = 0 \tag{4.41}$$

La única diferencia con el caso previamente considerado de clases linealmente separables está en la primera de las dos restricciones, donde es necesario establecer limitar superiormente a los multiplicadores de Lagrange por C. El caso linealmente separable corresponde con $C \to \infty$. Las variables débiles ξ_i, y sus multiplicadores de Lagrange asociados, μ_i, no intervienen

en el problema explícitamente. Su presencia está reflejada indirectamente mediante C.

En todo este estudio se ha considerado sólo el caso de clasificación con dos clases. En el caso de M-clases, se puede extender fácilmente mirando el problema como M problemas de dos clases. Para cada una de las clases, tratamos de diseñar una función discriminante óptima, $g_i(\mathbf{x})$, i=1,2,...,M, de modo que $g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i$ si $\mathbf{x} \in \omega_i$. Adoptando la metodología de SVM podemos diseñar las funciones discriminantes de modo que $g_i(\mathbf{x}) = 0$ sea el hiperplano óptimo para separar la clase ω_i de todas las demás, dado por supuesto que esto es posible. Así, la función lineal resultante dará $g_i(\mathbf{x}) > 0$ para $\mathbf{x} \in \omega_i$ y $g_i(\mathbf{x}) < 0$ en caso contrario. La clasificación se consigue de acuerdo a la siguiente regla

$$\text{Asignar } \mathbf{x} \text{ a } \omega_i \text{ si } i = \arg_k \text{máx} \{g_k(\mathbf{x})\}$$

Esta técnica, sin embargo, puede conducir a regiones indeterminadas, donde más de un $g_i(\mathbf{x})$ es positivo. Otra aproximación es extender la formulación matemática de SVM de dos clases al problema de M clases.

4.3.2. SVM no lineal

En el apartado anterior se discutieron las máquinas de vectores de soporte (SVM) como una metodología óptima de diseño de un clasificador lineal. Asumimos ahora que existe un mapeo

$$\mathbf{x} \in \mathbb{R}^n \rightarrow \mathbf{y} \in \mathbb{R}^k$$

desde el espacio de entrada a un espacio k-dimensional, donde las clases se pueden separar satisfactoriamente por un hiperplano lineal. Recordamos que los vectores de características participan por pares mediante la operación del producto interno. Una vez que se calcula el hiperplano óptimo $(\mathbf{w}, w_0)$, la clasificación se realiza según si el signo de

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \sum_{i=0}^{N_s} \lambda_i y_i \mathbf{x}^T\mathbf{x} + w_0 \tag{4.42}$$

es positivo o negativo, donde N_s es el número de vectores de soporte. Así, una vez más, sólo el producto interno entra en escena. Si el diseño se va a llevar a cabo en el espacio k-dimensional, la única diferencia es que los vectores involucrados estarán en los mapeos k-dimensionales del vector de características original. Una simple ojeada a esto nos llevaría a la conclusión de que ahora la complejidad es mucho mayor, puesto que habitualmente k es mucho más alto que la dimensión n del espacio de entrada, para poder hacer las clases linealmente separables.

El producto interno de los vectores en el nuevo espacio dimensional se puede expresar como función del producto interno de los correspondientes vectores en el espacio de características original.

Ejemplos típicos de kernels usados en aplicaciones de reconocimiento de patrones son

- Polinómicos:
$$K(\mathbf{x}, \mathbf{y}) = [\gamma(\mathbf{x} \cdot \mathbf{y}) + c]^d. \tag{4.43}$$
- Funciones de base radial (RBF):
$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2). \tag{4.44}$$
- Tangente hiperbólica:
$$K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma(\mathbf{x} \cdot \mathbf{y}) + c). \tag{4.45}$$
Para valores apropiados de γ y c de modo que las condiciones de Mercer se cumplan (ver Capítulo 8).

Una vez que el kernel adecuado se adopta, que implícitamente define un mapeo a un espacio de dimensión mayor, la tarea de clasificación se convierte en
$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{4.46}$$

sujeto a

$$0 \leq \lambda_i \leq C, \quad i = 1, 2, ..., N \tag{4.47}$$

$$\sum_i \lambda_i y_i = 0 \tag{4.48}$$

y el clasificador lineal resultante es

$$\text{Asignar } \mathbf{x} \text{ a } \omega_1(\omega_2) \text{ si } g(\mathbf{x}) = \sum_{i=1}^{N_s} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + w_0 > (<)0$$

Una característica notable de las máquinas de vectores de soporte es que la complejidad computacional es independiente de la dimensionalidad del espacio kernel, donde las características de entrada son mapeadas. Así, uno diseña en un espacio de dimensión alta sin tener que adoptar modelos explícitos usando una gran cantidad de parámetros. A su vez, una limitación importante de las máquinas de vectores de soporte es la alta carga computacional que se requiere, tanto durante el entrenamiento como durante el test. Para problemas con una cantidad relativamente pequeña de datos de entrenamiento se puede usar cualquier algoritmo de optimización de propósito general. Sin embargo, para una gran cantidad de puntos de entrenamiento (del orden de unos miles), se requiere un tratamiento especial. Entrenar con SVM normalmente se realiza por tandas. Para grandes problemas esto significa una demanda de necesidades de memoria del ordenador. Para solventar este problema se han ideado ciertos procedimientos. Su filosofía se basa en la descomposición, de una manera o de otra, del problema de optimización en una secuencia de otros más pequeños. Otra limitación importante de las máquinas de vectores de soporte es que, hasta ahora, no hay un método práctico para seleccionar la mejor función de kernel. Esto es todavía un problema sin solución.

4.4. Redes neuronales

El término de red neuronal (NN de sus siglas en inglés Neural Network) [McCulloch and Pitts, 1943] tiene su origen en los intentos por encontrar una representación matemática del procesamiento de información en sistemas biológicos, como el cerebro [McCulloch and Pitts, 1943]. Las NNs pueden agruparse en dos categorías en función de la arquitectura o patrón de conexión.

- Redes neuronales feed-forward, en las cuales no existen bucles, y

- Redes recurrentes (o retroalimentadas), en las cuales existen bucles debido a conexiones hacia atrás.

Diferentes conectividades dan lugar a diferentes comportamientos de red. En términos generales, las redes feed-forward son estáticas, es decir, producen sólo un conjunto de salidas en lugar de una secuencia de valores para una entrada dada. Las redes feed-forward son sin memoria en el sentido de que su respuesta a una entrada es independiente del estado anterior de la red. Las redes recurrentes, por otro lado, son sistemas dinámicos. Cuando se presenta un patrón de entrada nuevo, se calculan las salida de las neuronas. Debido a las conexiones de retroalimentación, las entradas de las neuronas se modifican, lo cual conduce a un nuevo estado de la red.

Las redes feed-forward a menudo tienen una o más capas ocultas de neuronas sigmoides seguidas por una capa de salida de neuronas lineales. Múltiples capas de neuronas con funciones de transferencia no lineales permiten a la red aprender relaciones lineales y no lineales entre vectores de entrada y salida. El proceso de aprendizaje en el contexto de NNs puede entenderse como el problema de actualizar la arquitectura de la red y los pesos de modo que la red pueda realizar de forma eficiente una tarea específica. La habilidad de las NNs para aprender automáticamente a partir de ejemplos las hace un clasificador atractivo para este tipo de aplicaciones. El desarrollo de algoritmos de aprendizaje de retro-propagación para determinar los pesos de un perceptrón multicapa ha hecho que estas redes sean las más populares entre investigadores de NNs.

Los modelos lineales de clasificación se basan en combinaciones lineales de las bases fijas y no lineales $\phi_j(\mathbf{x})$ y toman la forma

$$y(\mathbf{x}, \mathbf{w}) = f\left(\sum_{j=1}^{M} w_j \phi_j(\mathbf{x})\right) \tag{4.49}$$

donde $f(\cdot)$ es una función de activación no lineal. El objetivo es extender este modelo haciendo que las funciones base $\phi(\mathbf{x})$ dependan de parámetros y permitir entonces que esos parámetros se ajusten, junto con los coeficientes w_j durante el entrenamiento. Existen muchas posibilidades para construir funciones base paramétricas no lineales. Las NNs usan funciones base que siguen la misma forma que la ecuación 4.49, de modo que cada función base es en sí una función no lineal de una combinación lineal de las entradas, donde los coeficientes de la combinación lineal son parámetros adaptativos.

Esto conduce a un modelo básico de NN, que puede describirse como una serie de transformaciones funcionales. Primero se construyen M combinaciones lineales de las variables de entrada $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ de la forma

$$a_j = \sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{4.50}$$

donde $j = 1, ..., M$ y el superíndice (1) indica que los parámetros correspondientes se encuentran en la primera capa de la red. Nos referimos a los parámetros $w_{ij}^{(1)}$ como los *pesos* y a los parámetros $w_{i0}^{(1)}$ como los *sesgos*. Las cantidades a_j se conocen como *activaciones*. Cada una de ellas se transforma usando una función de activación $h(\cdot)$ diferenciable y no lineal que lleva a

$$z_j = h(a_j) \tag{4.51}$$

Estas cantidades corresponden con las salidas de las funciones base de la ecuación 4.49, que en el contexto de NN se denominan *unidades ocultas*. Las funciones no lineales $h(\cdot)$ se eligen generalmente como funciones sigmoidales como la sigmoide logística o la función 'tanh'. Siguiendo 4.49, estos valores se combinan de nuevo de forma lineal para dar las *unidades de activación de salida*

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{4.52}$$

donde $k = 1, ..., K$ y siendo K el número de salidas. Esta transformación corresponde con la segunda capa de la red y de nuevo los coeficientes $w_{k0}^{(2)}$ son los parámetros sesgo. Finalmente, las unidades de activación de salida se transforman usando una función de activación apropiada para dar el conjunto de salidas de la red y_k. La elección de la función de activación viene determinada por la naturaleza de los datos.

Combinando estas diferentes etapas se llega a función de la NN que, para una función de salida sigmoidal toma la forma

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma \left(\sum_{j=1}^{M} w_{kj}^{(2)} \left(\sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} \right) \tag{4.53}$$

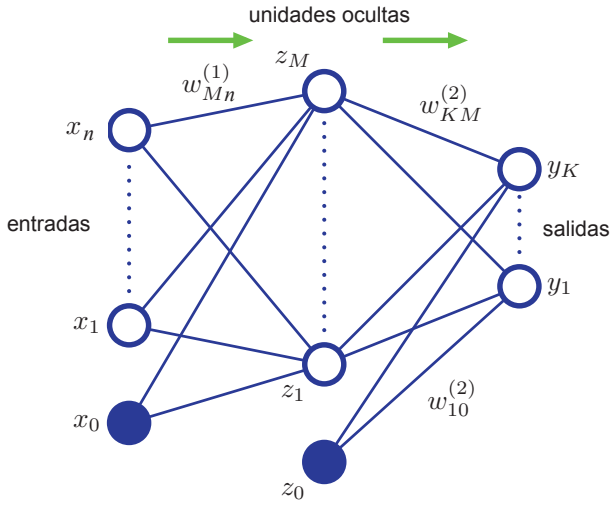


Figura 4.2: Red neuronal correspondiente a la ecuación 4.53. Las variables de entrada, las ocultas y las de salida se representan por nodos, y los parámetros de los pesos se representan mediante enlaces entre nodos. Los parámetros sesgo se representan por enlaces procedentes de entradas adicionales y variables ocultas x_0 y z_0. Las flechas indican la dirección del flujo de información a través de la red durante la propagación hacia adelante.

donde la función σ se definió en la ecuación 4.8. El conjunto de todos los pesos y sesgos se agrupan en un vector $\mathbf{w}$. Así, el modelo de red neuronal es simplemente una función no lineal que trasnforma un conjunto de variables de entrada $\{x_i\}$ a un conjunto de variables de salida $\{y_i\}$ controlado por un vector $\mathbf{w}$ de parámetros ajustables. Esta función se representa en el diagrama de red de la Figura 4.2. El proceso de evaluación 4.53 puede interpretarse como una propagación hacia adelante de la información a través de la red.

Para los experimentos presentados en este trabajo, se usa una red neuronal feed-forward con la siguiente configuración:

- Una capa oculta con un número creciente de neuronas y una capa de salida lineal.
- Función de transferencia sigmoide hiperbólica: $f(n) = 2/(1 + exp(-2 * n)) - 1$, para las capas de entrada.
- Función lineal de transferencia: $f(n) = n$, para la capa de salida.

- Peso y sesgo actualizados de acuerdo al algoritmo de optimización Levenberg-Marquardt.
- Función de aprendizaje que usa gradiente descendiente con peso y sesgo.

Parte II

Desarrollos Experimentales

CAPÍTULO 5

Regiones de Interés

En el proceso de detección de la EA a través de la exploración visual de las imágenes, los expertos focalizan la atención en las regiones cerebrales donde la enfermedad produce un decremento de la actividad funcional. Dicho de otra manera, no todo el volumen cerebral es relevante a la hora de diagnosticar la EA, y por tanto aquellas regiones que no aportan información para el diagnóstico pueden descartarse del estudio. En este capítulo se persigue la localización automática de las regiones de interés o ROIs (del inglés, Regions of Interest) mediante una exploración exhaustiva del volumen completo, que proporciona conjuntos de voxels agrupados que forman las ROIs. De esta manera quedan localizados los voxels de interés que serán empleados en la tarea de clasificación, pudiéndose descartar el resto. La relevancia de cada región viene determinada por la precisión que alcanza dicha región en la clasificación de pacientes normales y AD mediante el entrenamiento y test de clasificadores SVM empleando como características los niveles de gris de los voxels que la contienen.

5.1. Técnicas basadas en ROIs

Las técnicas basadas en regiones de interés (ROIs), también conocidas como técnicas basadas en componentes, han sido muy utilizadas en los últimos años en varias tareas de reconocimiento y detección de patrones, y en concreto en detección y reconocimiento de caras [Heisele et al., 2001b; Schneiderman and Kanade, 2000; Heisele et al., 2001a; Brunelli and Poggio, 1993; Nefian and Hayes, 1999; Wiskott, 1995]. En las aproximaciones globales, se emplea un único vector de características representando la imagen completa (resultante tras los pasos previos de preprocesado) como entrada para el clasificador. Para aplicaciones en detección de patrones de hipoperfusión, la primera técnica desarrollada fue Voxels-As-Features (VAF) [Stoeckel et al., 2001], que será tomada como referencia a lo largo de este trabajo. Las técnicas globales funcionan bien para detectar planos frontales, pero no son robustas ante traslaciones y rotaciones. Para evitar este problema se puede añadir una etapa de alineado o registro espacial (ver Sección 1.3) antes de la etapa de clasificación. Sin embargo las aproximaciones globales presentan dos inconvenientes. Por un lado sufren el denominado problema del pequeño tamaño muestral (en inglés *small sample size*), es decir, el número de características empleadas en la fase de entrenamiento del clasificador es mucho menor que el número de ejemplos para el aprendizaje, por lo que el aprendizaje no se efectúa de manera óptima. Por otro lado, las aproximaciones globales no explotan la propiedad de vecindad de lo voxels que componen las ROIs al considerar el volumen completo como un vector único de características.

Las aproximaciones basadas en búsqueda de ROIs surgen intuitivamente en las aplicaciones de detección de caras concluyendo que tiene sentido utilizar sólo algunas partes de la imagen donde se pueda focalizar la información de interés, en lugar de emplear la imagen completa que es más sensible a cambios de iluminación o pose, al mismo tiempo que reducen la dimensión del espacio de entrada, resolviendo así el problema del pequeño tamaño muestral [Segovia et al., 2009a; Salas-Gonzalez et al., 2009c].

Los sistemas basados en componentes deben seleccionar qué componentes o regiones de la imagen se van a usar. Algunos sistemas como los descritos en Heisele et al. [2000]; Leung et al. [1995] emplean aquellas componentes predominantes de forma natural en las imágenes faciales, tales como ojos, nariz y boca. Otros sistemas se diseñan para aprender estas regiones de manera automática a partir de las imágenes de entrenamiento [Heisele et al., 2001b; Viola, 1996]. Una vez que las componentes son detectadas en las imágenes de entrada, y posiblemente etiquetadas con un nivel de confianza, el sistema de

detección de objetos basado en componentes empleará otro clasificador para determinar si efectivamente se trata de componentes útiles en la detección del patrón o no. Éste último método será el empleado en este trabajo: la búsqueda de ROIs se realiza de manera automática generalizando el algoritmo y adecuándolo por tanto para cualquier problema de clasificación binaria de patrones cerebrales.

Un modo de convertir un clasificador en detector de ROIs es mediante enventanado, donde componente de la imagen se trata independientemente para entrenar a un clasificador [Vaillant et al., 1993]. Cuando la salida del clasificador supera cierto umbral, entonces la componente correspondiente se etiqueta como ROI.

5.1.1. Subdivisión de una imagen en componentes

Cada imagen cerebral está representada por un conjunto de valores de niveles de gris. Espacialmente este conjunto de valores tiene una estructura cúbica que denominaremos V, por lo que de modo genérico diremos que una imagen queda representada por la función tridimensional $\mathbf{I}(x, y, z)$, que indica la intensidad registrada en cada vóxel con coordenadas $(x, y, z) \in V$. La subdivisión del volumen V en componentes consiste en la extracción de subgrupos de voxels o componentes $C_i \in V$, $i = 1, ..., s$ de modo que cada vóxel pertenezca al menos a una componente C_i. Todos los voxels de una componente cumplen la propiedad de vecindad con al menos otro voxel de la misma componente, es decir, para cada voxel $(x, y, z) \in C_i$, al menos uno de sus 8 vecinos (x', y', z') también se encuentra contenido en C_i. Puede considerarse la posibilidad de solapamiento entre dos o más componentes contiguas en el espacio, de manera que un voxel pertenezca a más de una componente. Las componentes son de tamaño fijo F y no hacen referencia a priori a la forma de las mismas.

Por tanto, la imagen cerebral completa $\mathbf{I}(V)$ queda subdividida en el mismo número de subconjuntos ó componentes $\mathbf{I}(C_1), \mathbf{I}(C_2), ..., \mathbf{I}(C_s)$, donde una *componente* estará constituida por un conjunto de valores de intensidad de manera que:

$$\mathbf{I}(V) = \bigcup_{m=1}^{s} \mathbf{I}(C_m) - \bigcap_{m=1}^{s} \mathbf{I}(C_m) \tag{5.1}$$

donde el segundo término de la parte derecha de la igualdad elimina la redun-

dancia del solapamiento de las componentes, por ejemplo, las componentes que se solapan y cubren parcialmente la misma región cerebral. El motivo de considerar este caso es el considerar la búsqueda "fina" de localizaciones de componentes para poder adaptarlas la las regiones cerebrales más relevantes para la clasificación, discutido en la siguiente sección. Cada vector $\mathbf{x} \in \mathbb{R}^F$ constará de una etiqueta con $y \in \pm 1$. Estos vectores etiquetados se usan como vectores de características para la construcción de un clasificador SVM. Existirán tantos vectores por cada imagen cerebral como subdivisiones en componentes, todos ellos compartiendo la misma etiqueta. Sin embargo, la tarea de clasificación se llevará a cabo considerando cada componente individualmente, obteniéndose un numero s de categorizaciones alternativas de una misma imagen. El último paso supondrá el agregado del conjunto de decisiones SVM para obtener una decisión final colectiva [Álvarez et al., 2008; Górriz et al., 2008].

5.2. Conjunto de SVMs

El agregado o conjunto de SVMs, como orientación alternativo al estudio de SVM, está especializado en combinar una familia de SVMs actuales para inteligencia artificial avanzada. Los ya bien conocidos métodos de agregación de SVMs son el Uno-contra-Todos y el Uno-Contra-Uno. El propósito de dichos conjuntos es extender el SVM binario a una clasificación multiclase. Un proceso típico de agregado puede resumirse en tres pasos: selección del modelo SVM, agregado convexo, y entrenamiento agregado.

5.2.1. Métodos para construir conjuntos de SVM

En SVM agrupados, SVMs individuales se agregan para realizar una decisión colectiva de varios modos posibles como el voto por mayoría, o la ponderación de mínimos cuadrados basada en estimación. El entrenamiento del grupo de SVM puede llevarse a cabo mediante los llamados métodos "bagging" o "boosting". En bagging, cada SVM individual se entrena independientemente usando un conjunto de entrenamiento elegido de forma aleatoria mediante la técnica de bootstrap. En boosting, cada SVM individual se entrena usando el conjunto de entrenamiento elegido de acuerdo con la distribución de probabilidad de las muestras, la cual se actualiza en relación al error de la muestra. SVM agrupado es en esencia un tipo optimización de validación cruzada de un único SVM, llevando a cabo una clasificación con

mejores resultados que los otros modelos. Los detalles de la construcción y aplicaciones de SVM agrupado se describen en Hyun-Chul et al. [2003].

5.2.2. Métodos para agregado de SVM

Tras haber entrenado el sistema, es necesario agregar varios SVMs entrenados independientemente con un método de combinación adecuado. Existen dos tipos de técnicas de combinación que son los métodos lineales y los no lineales. Entre los métodos lineales, esto es, combinaciones lineales de varios SVMs, se encuentran el "Voto por mayoría", la "Ponderación basada en LSE (Least Squares Estimation)" y el "Pegado de votos", descritos abajo. El voto por mayoría y la ponderación basada en LSE se usan habitualmente con bagging y boosting respectivamente. La idea del pegado de votos pretende aliviar problemas de requisitos de memoria para almacenar la base de datos. Por otro lado un método no lineal, es decir, combinaciones lineales de varios SVMs, incluye la combinación jerárquica de doble capa que usa otro SVM de capa superior para combinar varios SVMs de capas más bajas.

Voto por Mayoría

El voto por mayoría es el método más simple para combinar varios clasificadores SVM [Breiman, 1999]. Sea $f_k, (k = 1, 2, ..., s)$ la función de decisión del k-ésimo SVM en el conjunto de SVMs y $y_j, (j = 1, 2, ..., c)$ la etiqueta para la j-ésima clase. Sea $N_j = \#\{k|f_k(\mathbf{x}) = y_j\}$, es decir, el número de SVMs cuyas decisiones son sobre la j-ésima clase. Entonces, la decisión final del conjunto SVM $f_{vm}(\mathbf{x})$ para un vector de prueba $\mathbf{x}$ debida al voto por mayoría viene determinada por

$$f_{vm}(\mathbf{x}) = \arg \max_j N_j \tag{5.2}$$

En el caso de clasificación binaria, consideremos las etiquetas $y = 1$ e $y = -1$ correspondientes a las clases AD y NORMAL respectivamente. Entonces la suma no pesada de los s votos que proyecta cada componente sobre un paciente concreto. Explícitamente, esta función $\mathcal{F}$ de decisión colectiva se

define como

$$\mathcal{F}(\mathbf{x}) = \frac{1}{s}\sum_{k=1}^{s} f_k \tag{5.3}$$

clasificando el paciente como AD si $\mathcal{F}(\mathbf{x}) > 0$, y como NORMAL si $\mathcal{F}(\mathbf{x}) < 0$. Este proceso se itera N veces, siendo N el número de pacientes en la base de datos, hasta que cada paciente de la base de datos es usado una vez como test (técnica de validación dejar uno fuera).

El carácter democrático de este método de agregación se plasma en el hecho de que el paciente será clasificado como perteneciente a una clase si la *mayoría* de las componentes de ese paciente son clasificadas como pertenecientes a esa clase, a pesar de que una cantidad destacable de componentes se clasifiquen erróneamente. Las regiones del cerebro irrelevantes para la detección de la enfermedad serán clasificadas de manera aleatoria en una u otra clase, mientras que las relevantes categorizarán el paciente de la misma manera. Si el número de componentes es suficientemente grande, los votos emitidos aleatoriamente tenderán a compensarse, haciendo que la suma total de los votos de regiones irrelevantes sea nula. De esta manera, unos pocos votos emitidos de manera no aleatoria harán inclinar la decisión final hacia una clase concreta, haciendo que se seleccionen de forma indirecta los votos de las regiones de interés.

El método de voto por mayoría destaca por su sencillez y efectividad, con la interesante propiedad añadida de ser muy robusto, ya que debido a lo anteriormente expuesto, los errores tienden a compensarse y es débilmente afectado por la modificación en cualquiera de los parámetros.

Voto por Relevancia

En este esquema permitimos que solo las componentes más relevantes proporcionen su voto. La relevancia se define en cada componente en función del rendimiento que proporciona el clasificador sobre esa componente, siendo las más relevantes aquellas en las que el proceso de clasificación sea más eficaz.

Para ordenar las componentes de una imagen según un criterio de relevancia, será necesario tener una estimación del rendimiento del clasificador sobre ellas. Puesto que se dispone de una base de datos etiquetada, la pre-

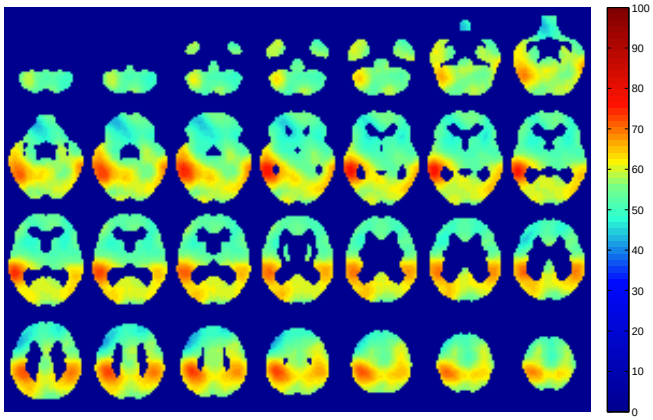


Figura 5.1: Mapa de precisión de componentes cúbicas para imágenes SPECT.

cisión de cada componente puede computarse mediante un proceso previo de clasificación interna, evaluado mediante la técnica Leave-N-Out. Una vez asignado un valor de precisión a cada componente, se descartan aquellas que empobrecen las tasas de clasificación. Este proceso sólo es necesario realizarlo una vez antes de iniciar la evaluación del sistema. La Figura 5.1 muestra el mapa de precisión obtenido sobre las imágenes SPECT mediante la subdivisión de las imágenes en componentes cúbicas de tamaño $4 \times 4 \times 4$. Las componentes con mayor tasa de precisión determinan las ROIs. En particular, existen tres zonas claras mencionadas en la literatura para el diagnóstico de la EA [Goethals et al., 2002] como son la región temporo-parietal, el cingulado posterior y el lóbulo temporal. Estas regiones coinciden con las destacadas mediante la técnica de componentes.

Una vez determinado el mapa de precisión que proporciona cada componente, se seleccionan aquellas cuyo valor de precisión $A^{(i)}$ sea suficientemente alto. El criterio para determinar si A es o no alto será compararlo con un valor estándar T considerado como alto. Así, todas las componentes cuyo valor de A sea superior a T serán consideradas relevantes, y pertenecerán al conjunto de componentes relevantes.

$$\mathcal{C}_T = \{C_i \setminus A^{(i)} > T\} \tag{5.4}$$

Establecido un criterio de relevancia, se define la función de clasificación

$\mathcal{R}$ (que denominaremos *voto por relevancia*) sobre un paciente, como la suma ponderada de las componentes relevantes, donde el peso que pondera cada componente es precisamente la precisión individual $A^{(i)}$ de dicha componente y donde el criterio de relevancia viene determinado por 5.2.2.

Matemáticamente, la función que evalúa el voto por relevancia para un paciente de test $\mathbf{x}$ viene determinada por:

$$\mathcal{R}(\mathbf{x}) = \sum_{k \in \mathcal{C}_T} f_k A^{(k)} \tag{5.5}$$

Así, sólo un subconjunto de componentes $\mathcal{C}_T$ con valores $A^{(i)}$ mayores de T se tienen en cuenta para la votación, ponderando además el voto por la fiabilidad que representa cada componente $A^{(i)}$. Como para el caso de la función de decisión $\mathcal{F}$ de voto por mayoría, el valor del signo de la salida en $\mathcal{R}$ definirá la pertenencia de clases y dependerá del valor de T elegido.

La selección de características generalmente conduce a la obtención de clasificadores con un alto poder de generalización, y mediante la extracción de estas características relevantes podemos localizar los atributos que son responsables de las diferencias entre las clases bajo estudio. Por lo tanto, el voto por relevancia puede proporcionar una solución más precisa [Kohavi and John, 1995] que otros métodos, pero en general la complejidad computacional aumenta de manera sensible. Este aumento sugiere el uso del método de K-pliegues como validación, testando el rendimiento de la función de decisión (5.5) sobre el conjunto de los restantes $N - l$ sujetos, e iterando K veces el proceso.

Es interesante destacar que el subconjunto $\mathcal{C}_T$ dado por la ecuación define una máscara que selecciona las regiones de mejor rendimiento en la clasificación, que corresponderán con las regiones relevantes para el diagnóstico de la enfermedad de Alzheimer. Esta máscara se puede asumir independiente del conjunto muestral, a pesar de que el patrón de Alzheimer es variable, y que caracteriza a la propia enfermedad. Debido a la independencia de esta máscara con la base de datos, esta puede ser computada "off-line", reduciendo el coste computacional de la aproximación.

5.3. Experimentos

Para las imágenes SPECT y PET *Cartuja*, la descomposición de las imágenes en componentes alargadas se lleva a cabo mediante la subdivisión del volumen completo en 15, 20 ó 25 componentes tomando voxels consecutivos a lo largo de los tres ejes de exploración (axial, coronal y sagital) independientemente. Cada componente es desplazada una posición eje de exploración dando lugar a otra componente tratada de manera independiente de la anterior, a pesar de existir un solapamiento entre ellas. Este desplazamiento se realiza 30 veces, es decir, cada componente da lugar a otras 30 más donde cada una de ellas solapa con la anterior en todos los voxels menos en uno. Así, si el volumen se divide en 15, 20 ó 25 componentes iniciales, el número final de votos emitidos al desplazarlas por cada eje será $15 \times 3 \times 30 = 1350$, $20 \times 3 \times 30 = 1800$ y $25 \times 3 \times 30 = 2250$ respectivamente. Las componentes cúbicas por su parte están compuestas por grupos de tamaño $4 \times 4 \times 4$ que se desplazan igualmente a lo largo del todo el volumen. El número de componentes cúbicas resultante depende del factor de submuestreo $v \times v \times v$ aplicado al volumen antes de iniciar la exploración, donde v ha sido variado desde 4 hasta 7. De esta forma se pretende localizar subgrupos de voxels con formas diversas que permitan adaptar las componentes a la forma de las ROIs.

Una vez emitidos todos los votos, el decisión final se lleva a cabo mediante las dos técnicas de recuento correspondientes a las ecuaciones 5.3 de voto por mayoría, y 5.5 donde el valor $A^{(i)}$ proporcionado por cada componente se emplea como criterio de selección a la vez que como pesos para ponderar la suma y darle por tanto mayor importancia a los votos emitidos por las componentes de mayor precisión.

Las bases de datos sobre las que se ha aplicado el método por componentes son:

- Imágenes SPECT: 79 pacientes: 41 etiquetados como NORMAL y 38 etiquetados como AD.
- Imágenes PET *Cartuja*: 60 pacientes: 18 etiquetados como NORMAL y 42 etiquetados como AD.

Para la validación de este método de componentes sobre una base de datos con mayor reconocimiento se han llevado a cabo algunos experimentos sobre la base de datos ADNI. Debido al coste computacional que supone este método, las imágenes se exploraron mediante componentes cúbicas de tamaño $5 \times 5 \times 5$. Las imágenes que componen este grupo son

- Imágenes ADNI: 192 pacientes: 97 etiquetados como NORMAL y 95 etiquetados como AD.

5.4. Resultados

5.4.1. Bases de datos SPECT y PET *Cartuja*

Voto por mayoría

Tras la división de los volúmenes cerebrales en 10, 15 y 20 componentes alargadas y en componentes cúbicas de tamaño $4 \times 4 \times 4$, y mediante una decisión final evaluada por la función $\mathcal{F}$ de voto por mayoría se obtienen las gráficas de la Figura 5.2. Los pacientes etiquetados como AD se sitúan en las primera posiciones del eje horizontal seguidos de los controles. Una unanimidad de todas las componentes en el diagnóstico correcto del paciente produciría un 1 para pacientes AD y un 0 para pacientes normales. Si el número de errores aumenta, el recuento total se alejará del valor correcto. El umbral se sitúa justo en el valor medio, 0,5, de manera que si existen más componentes con voto erróneo que correcto, entonces el cómputo final sobrepasará este umbral clasificando el paciente con la etiqueta equivocada. Estos pacientes se marcan en las gráficas con un punto rojo, y verde en caso de ser clasificados correctamente. En estas gráficas se observa cómo para la base de datos PET el método de voto por mayoría consigue un alto porcentaje de precisión con alta fiabilidad de los votos emitidos, siendo el voto erróneo para dos pacientes en caso de emplear componentes alargadas y de uno sólo cuando se emplean componentes cúbicas. El etiquetado de este paciente en concreto fue revisado por los expertos clínicos y se confirmó que este paciente no estaba afectado por la EA, sin embargo se detectaron anomalías en el funcionamiento del hipocampo. Para la base de datos SPECT la fiabilidad de las componentes es menor así como la tasa de precisión total alcanzada al evaluar el método para esta base de datos, consiguiendo un máximo de 88,6 %, aun así superando el máximo de 78,4 % alcanzado por la técnica VAF consiederada como referencia.

La Figura 5.3 refleja cómo afecta el tamaño del factor de submuestreo v a la aplicación del método de componentes. Para la base de datos SPECT, el factor $7 \times 7 \times 7$ consigue los mejores resultados mientras que para PET este parámetro afecta mayormente a las componentes cúbicas, alcanzándose el mejor comportamiento para el factor $4 \times 4 \times 4$. Un conjunto de resultados

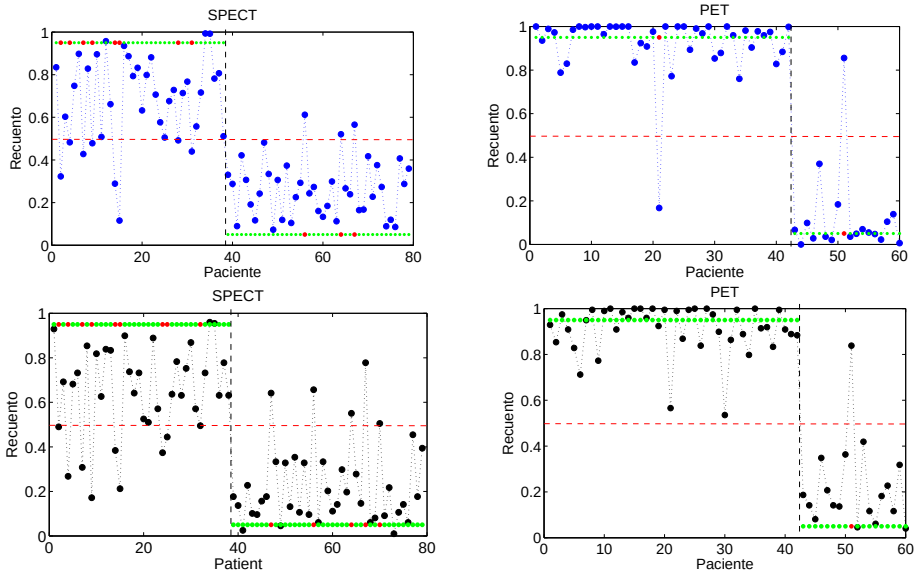
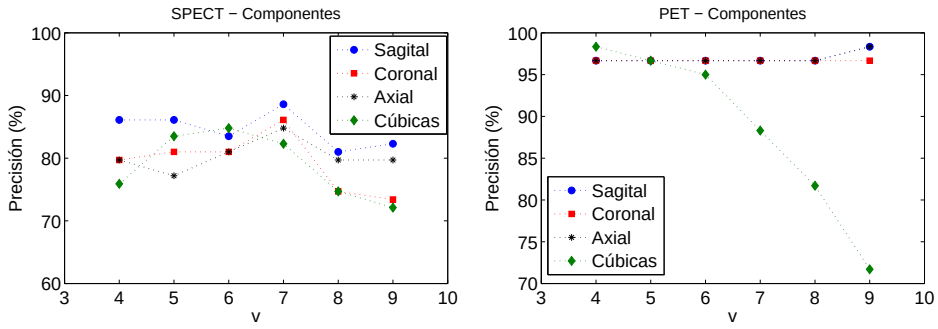


Figura 5.2: Recuento de votos emitidos por las componentes alargadas (arriba) y cúbicas (abajo) para las imágenes SPECT (izquierda) y PET (derecha) para cada uno de los pacientes. Los pacientes clasificados correctamente se marcan con un punto rojo y verde en caso contrario.



(a) (b)

Figura 5.3: Precisión alcanzada por las componentes alargadas en cada uno de los ejes del volumen y las componentes cúbicas en función del factor de submuestreo v para las imágenes (a) SPECT y (b) PET.

	v	15 Alargadas	20 Alargadas	25 Alargadas	Cúbicas
SPECT	4	79.75 %	82.28 %	81.01 %	75.95 %
	5	79.75 %	78.48 %	83.54 %	83.54 %
	6	83.54 %	82.28 %	84.81 %	84.81 %
	7	83.54 %	86.08 %	87.34 %	82.28 %
PET	4	96.67 %	96.67 %	96.67 %	98.33 %
	5	96.67 %	96.67 %	96.67 %	96.67 %
	6	96.67 %	96.67 %	96.67 %	95 %
	7	96.67 %	96.67 %	96.67 %	88.33 %

Tabla 5.1: Resultados obtenidos por el método de componentes con recuento de votos por mayoría.

completo obtenido para cada tipo de componente cuando el parámetro v varía se presenta en la Tabla 5.1.

Voto por relevancia

Las gráficas de la Figura 5.4 muestran los resultados obtenidos con el voto por relevancia en función del umbral establecido T y para la combinación de diferentes tipos de componentes y diferentes factores v de submuestreo. Como era de esperar, a medida que el umbral aumenta el número de componentes consideradas en el recuento disminuye según la curva verde representada en cada gráfica a la vez que la precisión total obtenida sobre cada base de datos aumenta. Existe por tanto la necesidad de llegar a un compromiso entre estas dos variables y establecer así un umbral de selección con un nivel de restricción adecuado para las imágenes que se están tratando. A partir de las gráficas se puede establecer empíricamente un valor del umbral adecuado para cada base de datos. Para la base de datos SPECT (gráficas de la columna izquierda) se produce una mejora considerable a partir de un umbral $T = 84\,\%$, alcanzándose una precisión sobre esta base de datos de 97,47 %. Para este valor umbral el número de componentes consideradas es aproximadamente 200. Para la base de datos PET (gráficas de la columna derecha) sólo se producen mejoras con respecto a la votación por mayoría cuando se establece un valor del umbral a partir de $T = 92\,\%$. A pesar de suponer un valor muy alto en comparación con la base de datos SPECT, este umbral supone una restricción similar para la base de datos PET pues el número de componentes seleccionado mediante este valor de umbral es aproximadamente de 220. Nótese además que estos resultados se obtienen también con factores de submuestreo altos (por ejemplo $v = 6$), lo que permite au-

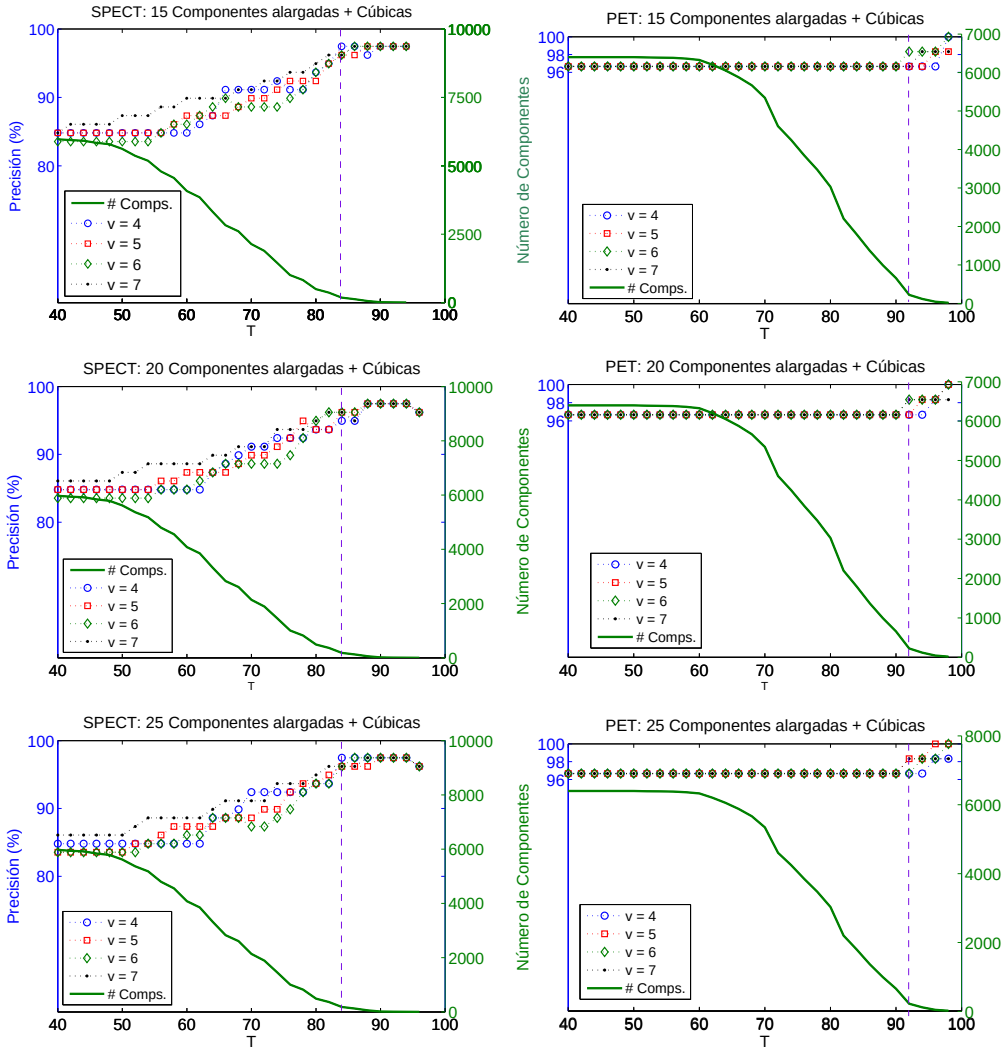


Figura 5.4: Resultados obtenidos mediante la evaluación de la función de voto por relevancia en función del umbral T para la base de datos SPECT (izquierda) y PET (derecha). El aumento del umbral supone una disminución del número de componentes que se consideran en el recuento.

mentar la efieciencia computacional del método y realizar la exploración por componentes a los volúmenes reducidos.

5.4.2. Base de datos ADNI

En una primera exploración, el conjunto de pacientes del grupo ADNI se divide aleatoriamente en 2 subgrupos, uno de ellos será empleado para la obtención del mapa de precisión y el otro compondrá el conjunto de test. Para la estimación de la precisión de cada componente se emplea la estrategia de validación cruzada k-pliegues con $k = 25$, obteniéndose así el mapa de precisión de componentes. En el proceso de test, las componentes seleccionadas (todas en el caso de voto por mayoría y sólo aquellas que superen el umbral T para voto por relevancia) se emplean para clasificación, obteniéndose los resultados de la Figura 5.5. La precisión de las componentes relevantes supera en todo caso al voto por mayoría alcanzando hasta un 89,53 % de precisión cuando se establece el umbral en $T = 78$ y $T = 79$. Para un valor de T superior a 80 el número de componentes que aportan su voto decrece y la variabilidad de los resultados aumenta debido a la fuerte criba que ejerce el umbral.

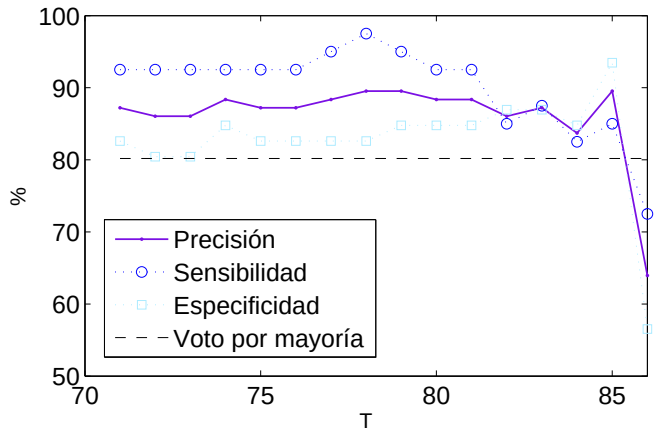
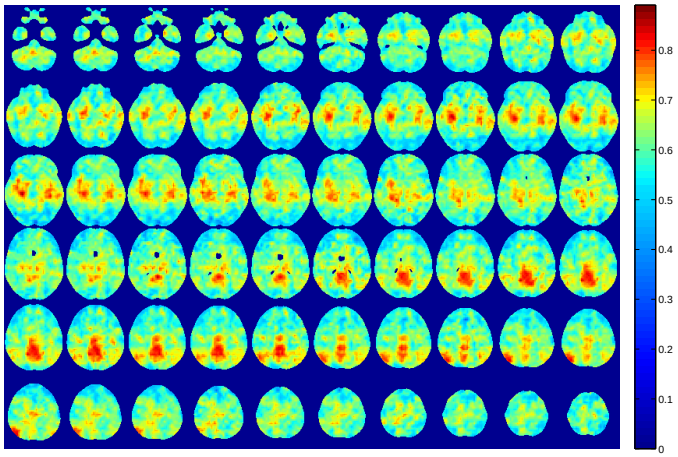


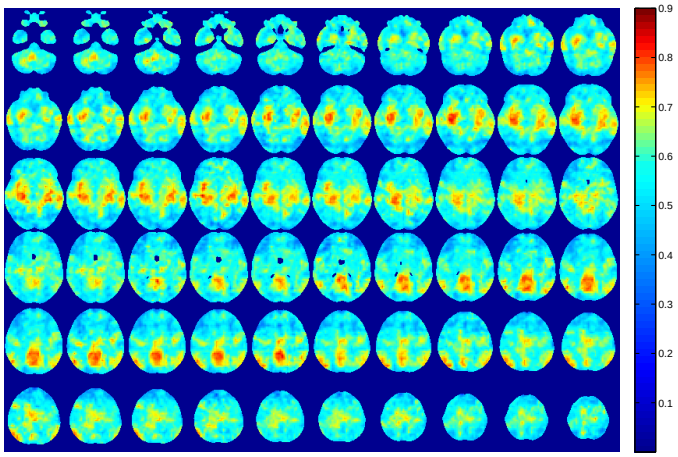
Figura 5.5: ADNI: Precisión, sensibilidad y especificidad a medida que el umbral de selección de componentes aumenta. Comparación con el método de voto por mayoría.

La Figura 5.6 muestra la relevancia de las componentes en términos de sensibilidad y especificidad, dando una noción espacial de la relevancia de las mismas. Los valores de sensibilidad son en regla general mayores que los de

especificidad. Esto tiene sentido si se consideran los criterios de etiquetado de la base de datos ADNI en los que prevalece la identificación de pacientes AD sobre la de los pacientes normales. El etiquetado se basa únicamente en los tests clínicos y no en la información contenida en las propias imágenes, proporcionando una alta sensibilidad en detrimento de la especificidad [Jobst et al., 1998]. Los voxels situados en el precunei proporcionan altos valores tanto de sensibilidad como de especificidad, y darán por tanto como resultado altas tasas de precisión.



(a)



(b)

Figura 5.6: ADNI: Mapas de (a) sensibilidad y (b) especificidad.

CAPÍTULO 6

Análisis de Componentes Principales

El análisis de Componentes Principales (PCA, de las siglas en inglés Principal Component Analysis) [Jolliffe, 1986], también conocido como la transformación de Karhunen-Loeve (KL), es uno de los resultados más valiosos del álgebra lineal aplicada. PCA se usa comúnmente en todas las formas de análisis porque es un método simple y no paramétrico para extraer información relevante de grupos de datos confusos. Con un mínimo esfuerzo adicional, PCA proporciona una vía para reducir un conjunto de datos complejo a una dimensión menor que revela la dinámica simplificada, y a veces oculta, que a menudo subyace bajo éste. La idea clave es proyectar los datos en un subespacio ortogonal que proporcione una representación compacta de los mismos, transformando las variables iniciales correladas en un pequeño número de variables no correladas, llamadas componentes principales (CPs). Esta técnica y otros métodos basados en PCA se han aplicado con éxito en diferentes problemas de clasificación de imagen [Kirby and Sirovich, 1990; Spetsieris et al., 2009; Turk and Pentland, 1991], y en concreto en clasificación de imágenes neurológicas [López et al., 2009c,a].

6.1. Análisis de Componentes Principales

Una aproximación simple para extraer la información contenida en una imagen es capturar de algún modo la variación en una colección de imágenes y usar esa información para codificar y comparar imágenes individuales. Desde el punto de vista de la teoría de la información, PCA pretende extraer la información relevante contenida en una imagen, codificarla tan eficientemente como sea posible y comparar la codificación de la imagen con una base de datos de modelos codificados de modo similar.

Aplicado al análisis de imágenes neurológicas, con PCA se desea encontrar las componentes principales (CPs) de la distribución de las imágenes, o lo que es lo mismo, los autovectores de la matriz de covarianza de un conjunto de imágenes tratando cada imagen como un punto (o vector) en un espacio de alta dimensión. Los autovectores quedan ordenados en función de la cantidad de variación entre las imágenes que representan.

6.1.1. Aspectos matemáticos: La transformación de Karhunen-Loéve

Cualquier vector n-dimensional $\mathbf{x}$ se puede representar de manera exacta a través de un conjunto de n vectores linealmente independientes $\mathbf{u}_k \in \mathbb{R}^n$ como:

$$\mathbf{x} = \sum_{i=1}^{n} \omega_i \mathbf{u}_i \tag{6.1}$$

donde se cumple que

$$\mathbf{u}_l^T \mathbf{u}_k = \delta_{lk} = \begin{cases} 1, & \text{si } l = k \\ 0, & \text{en caso contrario} \end{cases} \tag{6.2}$$

Supongamos que, en lugar de una representación fiel de $\mathbf{x}$ como en la ecuación 6.1, estamos interesados en aproximar $\mathbf{x}$ usando un número reducido ($m < n$) de vectores de la base $\{\mathbf{u}_i\}$. Una forma de hacerlo sería sustituir algunas componentes ω_i, cuyos valores no calculamos, por constantes arbitrarias b_i,

de manera que se construye la siguiente aproximación de $\mathbf{x}$:

$$\hat{\mathbf{x}} = \sum_{i=1}^{m} \omega_i \mathbf{u}_i + \sum_{i=m+1}^{n} b_i \mathbf{u}_i \tag{6.3}$$

El error que se comente al aproximar $\mathbf{x}$ por $\hat{\mathbf{x}}$ vendrá dado por:

$$\begin{aligned} \Delta\mathbf{x} =& \mathbf{x} - \hat{\mathbf{x}} \\ =& \sum_{i=1}^{n} \omega_i \mathbf{u}_i - \sum_{i=1}^{m} \omega_i \mathbf{u}_i - \sum_{i=m+1}^{n} b_i \mathbf{u}_i = \\ =& \sum_{i=m+1}^{n} (\omega_i - b_i)\mathbf{u}_i \end{aligned} \tag{6.4}$$

Seguiremos un criterio de mínimos cuadrados para obtener una solución óptima al problema de la aproximación, buscado aquel valor de las constantes b_i que minimice el error cuadrático medio (o mse de sus siglas en inglés *mean squared error*):

$$mse = E\left\{\Delta\mathbf{x}^2\right\} = \sum_{i=m+1}^{n} E\left\{(\omega_i - b_i)^2\right\} \tag{6.5}$$

Por lo tanto minimizar el error cuadrático medio equivale a buscar una solución a:

$$\frac{\partial}{\partial b_i} E\left\{(\omega_i - b_i)^2\right\} = -2(E\{\omega_i\} - b_i) = 0 \tag{6.6}$$

que sencillamente conduce a:

$$b_i = E\{\omega_i\} = \mathbf{u}_i^T E\{\mathbf{x}\} \tag{6.7}$$

quedando determinadas las constantes b_i al valor esperado de las compo-

nentes ω_i. Ahora, se puede reescribir el error cuadrático medio como:

$$\begin{aligned} mse &= \sum_{i=m+1}^{n} E\left\{(\omega_i - E\{\omega_i\})^2\right\} = \\ &= \sum_{i=m+1}^{n} \mathbf{u}_i^T E\left\{(\mathbf{x} - E\{\mathbf{x}\})\right\} E\left\{(\mathbf{x} - E\{\mathbf{x}\})\right\}^T \mathbf{u}_i = \\ &= \sum_{i=m+1}^{n} \mathbf{u}_i^T \mathbf{\Sigma_X} \mathbf{u}_i \end{aligned} \tag{6.8}$$

donde $\mathbf{\Sigma_X}$ es, por definición, la matriz de covarianza de $\mathbf{x}$. Se puede demostrar [Fukunaga, 1990; Miranda et al., 2008] que la elección optima para $\mathbf{u}_i$ es aquella que satisface:

$$\mathbf{\Sigma_X} \mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{6.9}$$

o dicho de otro modo, aquella en la que $\mathbf{u}_i$ y λ_i son los autovectores y autovalores de la matriz de covarianza. La expansión de un vector en términos de autovectores de la matriz de covarianza se denomina expansión de Karhunen-Loéve.

6.2. *Eigenbrains*

Los autovectores de la matriz de covarianza de un conjunto de imágenes puede entenderse como un conjunto de vectores que caracterizan la variación entre las imágenes. La localización de cada imagen contribuye más o menos a cada autovector, de modo que se puede representar cada autovector como una imagen propia. En el problema concreto de reconocimiento de caras, la representación de dichas imágenes fue denominada *eigenface* [Turk and Pentland, 1991]. Derivado de dicho término y debido a su apariencia de cerebro, los autovectores o CPs $\mathbf{u}_i, i = 1, ..., N$ de la matriz de covarianza se denominan *eigenbrains* [Álvarez et al., 2009a].

Cada imagen cerebral se puede representar exactamente en términos de una combinación lineal de eigenbrains, pero también puede aproximarse usando solo los "mejores" eigenbrains, considerados en una primera aproximación como aquéllos con mayor autovalor asociado y que por tanto cuentan con

la mayor parte de la varianza dentro del grupo de imágenes. Los mejores m eigenbrains extienden un subespacio m-dimensional de todas las posibles imágenes. Cada individuo por tanto se podría caracterizar por el pequeño conjunto de pesos asociados a las imágenes propias necesario para describirlo y reconstruirlo. Ésta es una representación extremadamente compacta en comparación con las imágenes en sí mismas.

6.2.1. Cálculo efectivo de los eigenbrains

Tras los pasos de preprocesado se obtiene una representación $3D$ de cada sujeto de tamaño $n = 79 \times 95 \times 69 \sim 5 \cdot 10^5$ voxels. Sea $\mathbf{I} = [\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N]$ el conjunto de estas representaciones recolocadas en forma de vectores n-dimensionales, donde N es el número de muestras o pacientes. La media de estos vectores se define como $I_M = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}_i$. Tras normalizar los vectores a norma unidad y restarles la media, se obtiene un nuevo conjunto de vectores $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, donde cada $\mathbf{x}_i$ representa un vector normalizado $n-$dimensional, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})^T$, $i = 1, 2, ..., N$. La matriz de covarianza se define como

$$\mathbf{\Sigma_X} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{N}\mathbf{X}\mathbf{X}^T \tag{6.10}$$

La matriz $\mathbf{\Sigma_X}$ es una matriz de $n \times n$, y determinar los n autovectores y autovalores puede llegar a ser una tarea intratable debido a los típicos tamaños n. Sin embargo existe un método viable para encontrar estos autovectores.

Nótese que si el tamaño de la muestra N es más pequeño que la dimensión del espacio de entrada n, existirán solo $N - 1$ (en lugar de n) autovectores significativos. Los restantes autovectores estarán asociados a autovalores de valor cero. Afortunadamente se puede resolver el problema de autovectores n-dimensionales mediante primero la resolución de los autovectores de una matriz $N \times N$ y después tomando las combinaciones lineales apropiadas de las imágenes $\mathbf{x}_i$. Consideremos que $\mathbf{v}_i$ son los autovectores de la matriz $\mathbf{X}^T\mathbf{X}$, es decir

$$\mathbf{X}^T\mathbf{X}\mathbf{v}_i = \mu_i \mathbf{v}_i \tag{6.11}$$

Multiplicando ambos lados de la ecuación por la izquierda por la matriz $\mathbf{X}$

se obtiene

$$\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{v}_i = \mu_i\mathbf{X}\mathbf{v}_i \tag{6.12}$$

de lo que se deduce que $\mathbf{X}\mathbf{v}_i$ son los autovectores de $\mathbf{\Sigma_X} = \mathbf{X}\mathbf{X}^T$.

Continuando con el análisis, se construye la matriz $\mathbf{L} = \mathbf{X}^T\mathbf{X}$ de dimensión $N \times N$, donde $\mathbf{L}_{ij} = \mathbf{x}_i^T\mathbf{x}_j$, y se calculan los autovectores, $\mathbf{v}_l$, de $\mathbf{L}$. Estos vectores determinan combinaciones lineales del conjunto de N imágenes de entrenamiento para formar los eigenbrains $\mathbf{u}_l$

$$\mathbf{u}_l = \sum_{k=1}^{N} \mathbf{v}_{lk}\mathbf{x}_k, \quad l = 1, ..., N \tag{6.13}$$

Con este análisis, los cálculos se reducen enormemente del orden del número de voxels n al orden del número de imágenes del conjunto de entrenamiento N. En la práctica, el tamaño del conjunto de entrenamiento será relativamente pequeño ($N \ll n$) haciendo los cálculos más manejables.

Los autovalores nos permiten ordenar sus autovectores asociados de acuerdo a la utilidad de los mismos a la hora de caracterizar la variación entre imágenes. Por ejemplo, en la Figura 6.1 se muestra la contribución de los primeros 30 eigenbrains obtenidos de la base de datos SPECT. Estos primeros 30 eigenbrains explican más del 90 % de la variación total contenida en la base de datos completa.

6.2.2. Uso de los eigenbrains para clasificación

En la práctica, sólo un número $m < N$ de eigenbrains es suficiente para la clasificación de imágenes, ya que no se requiere una reconstrucción precisa. En este contexto, la clasificación se convierte en una tarea de reconocimiento de patrones. Los m eigenbrains componen un subespacio m-dimensional del espacio original de entrada de dimensión n. Los m autovectores más significativos de la matriz $\mathbf{L}$ se eligen en principio como aquéllos con mayores autovalores asociados (ver Sección 6.2.3).

Una nueva imagen $\mathbf{I}$ se transforma en componentes eigenbrains o coeficientes PCA al proyectarla en el espacio "cerebro" mediante una simple

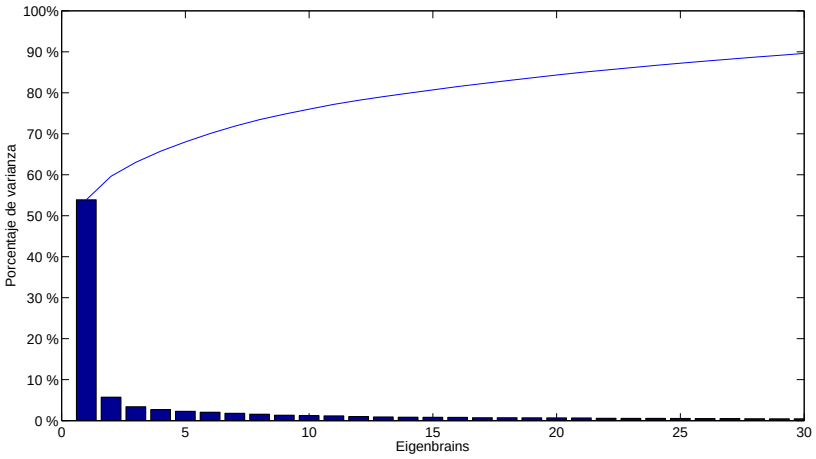
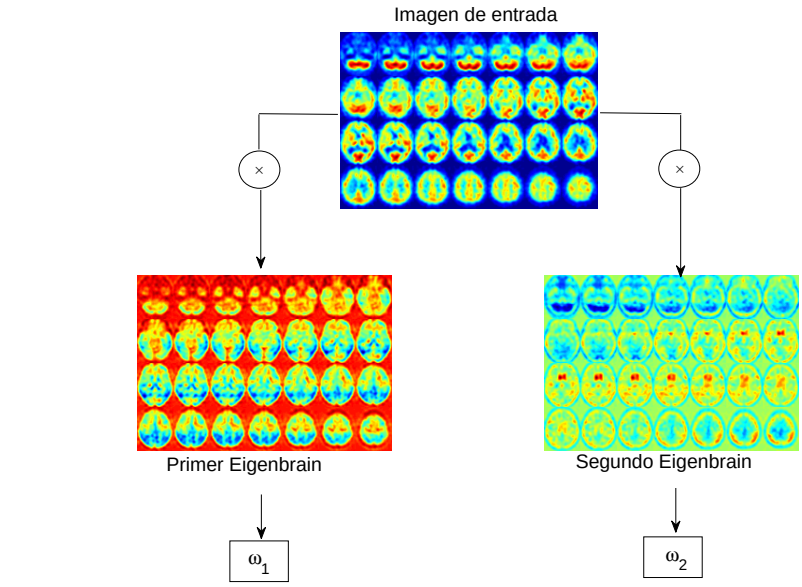


Figura 6.1: Porcentaje de la varianza total contenida en los primeros 30 eigenbrains obtenidos a partir de una base de datos real SPECT.

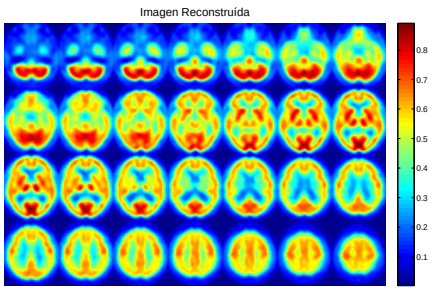
operación

$$\omega_k = \mathbf{u}_k^T(\mathbf{I} - I_M) \tag{6.14}$$

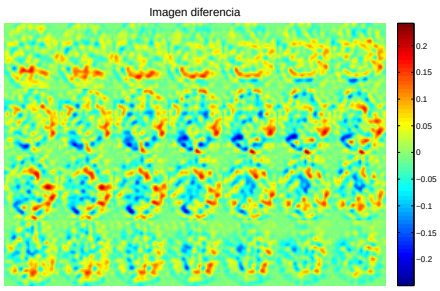
para $k = 1, ..., m$. Esto describe un conjunto de multiplicaciones y sumas de imágenes punto a punto. Los pesos o coeficientes ω_i forman un vector $\mathbf{\Omega}^T = [\omega_1, \omega_2, ..., \omega_m]$ que describe la contribución de cada eigenbrain en la representación de la imagen de entrada, tratando los eigenbrains como el conjunto de vectores base para las imágenes. El vector de pesos por tanto puede ser utilizado en un algoritmo estándar de reconocimiento de patrones para encontrar, dentro de un conjunto de clases predefinidas, aquélla que mejor describe la imagen. La Figura 6.2(a) muestra un esquema gráfico del proceso de extracción de los coeficientes ω_i, $i = 1, 2$. La imagen de entrada con la media substraída se proyecta en el espacio eigenbrain expandido por los dos primeros autovectores $\mathbf{u}_1$ y $\mathbf{u}_2$. En caso de ser necesaria una recuperación de la imagen de entrada, ésta podría ser reconstruida a partir de dichos coeficientes y los eigenbrains empleados en la proyección. Así, la aproximación de la ecuación 6.3 vendrá dada por la combinación lineal de los m autovectores $\mathbf{u}_i$ cuyas componentes principales tengan mayor varianza, recogiéndose las características de mayor variabilidad de los datos en un número $m < N$ de



(a)



(b)



(c)

Figura 6.2: (a) Esquema de la extracción de los dos coeficientes principales. La imagen de entrada normalizada se proyecta sobre los dos primeros eigenbrains, $\mathbf{u}_1, \mathbf{u}_2$ obteniéndose un coeficiente ω para cada proyección. La imagen de entrada se puede recuperar parcialmente mediante la suma de dichos eigenbrains ponderados por su correspondientes coeficientes. (b) Imagen recuperada a partir de los dos primeros eigenbrains. (c) Imagen diferencia entre la imagen original de entrada y la imagen (b) recuperada.

variables:

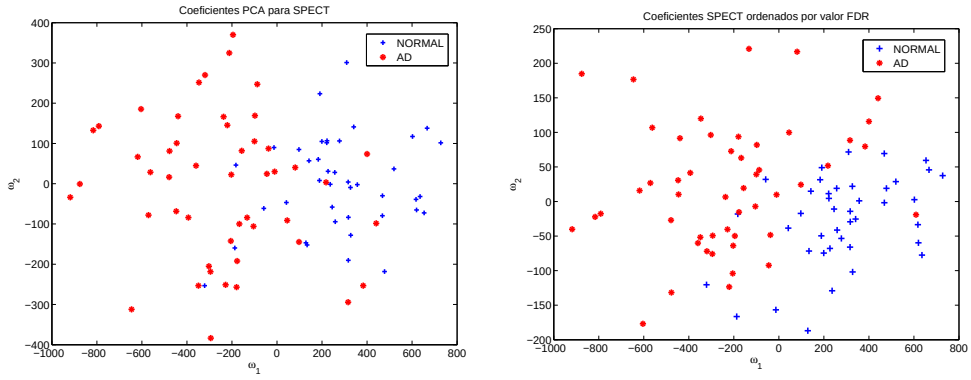
$$\hat{\mathbf{x}} = \sum_{i=1}^{m} \mathbf{u}_i \omega_i \tag{6.15}$$

La imagen recuperada $\hat{\mathbf{x}}$ con media cero se muestra en la Figura 6.2(b), y la imagen diferencia de la ecuación 6.4 obtenida al aproximar $\hat{\mathbf{x}}$ por $\mathbf{x}$ se representa en la Figura 6.2(c).

6.2.3. Selección de eigenbrains mediante el criterio de Fisher

La descripción de la base de datos en términos de variabilidad puede entenderse calculando la importancia de cada vector de proyección o CP mediante la ecuación 6.13, como muestra la Figura 6.1. Habitualmente, la importancia de las CPs viene determinada por sus autovalores asociados, proporcionando mayor aportación aquellos autovectores con autovalores mayores. Como método de aprendizaje no supervisado (es decir, el etiquetado de las imágenes no se requiere ni se usa), PCA es puramente descriptivo o representativo de los datos originales mediante un número sustancialmente menor de variables. Por tanto, el criterio de variabilidad no nos garantiza la máxima eficiencia de dichas variables en un marco de clasificación o discriminación entre clases, puesto que las variabilidades contenidas en ellas podrían responder a características comunes presentes en ambas clases.

La Figura 6.3(a) muestra gráficamente los coeficientes ω_1 y ω_2 extraídos mediante la proyección de las imágenes sobre el primer y segundo eigenbrains distinguidos por clases. Aproximadamente, se puede determinar de manera visual que valores de ω_1 mayores que cero corresponden a pacientes normales mientras que los valores negativos de ω_1 se pueden asociar a pacientes etiquetados como AD. Este hecho revela que la información contenida en este coeficiente es útil a la hora de distinguir sujetos AD de aquéllos normales. Sin embargo, esto no ocurre para el segundo coeficiente ω_2, cuya variabilidad contenida no parece estar relacionada con el AD, ya que toma valores en el mismo rango para pacientes de ambas clases. Este hecho motiva la reordenación de los coeficientes PCA ω_i mediante un criterio más útil para propósitos de clasificación, como el Factor Discriminante de Fisher (Fisher Discriminant Ratio FDR, [Fisher, 1936]). El uso de este criterio para tareas de clasificación tiene sentido puesto que se calcula considerando las etiquetas



(a) (b)

Figura 6.3: Coeficientes tras la proyección de la base de datos en los dos primeros eigenbrains ordenados por el criterio de (a) varianza y (b) FDR.

previamente asignadas a las muestras. Su expresión es:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{6.16}$$

donde μ_i and σ_i^2 denotan la media y varianza muestral de la clase i respectivamente. Así, en la fase de clasificación, el conjunto de entrenamiento se usa para el cómputo de las CPs y de los valores FDR asociados a los coeficientes resultantes de la proyección PCA. La imagen de test se proyecta sobre dichos CPs, y el conjunto de coeficientes Ω_{test} que se obtienen se reordenan en orden decreciente de acuerdo a los valores FDR calculados anteriormente a partir del conjunto de entrenamiento. En algunos casos, esta reordenación mejorará los resultados de clasificación, como se muestra en la Sección 6.5. Otras métricas para determinar la robustez de las CPs para propósitos de clasificación se proponen en Markiewicz et al. [2009]. La Figura 6.3(b) muestra los dos primeros coeficientes cuando se emplea el criterio de reordenación FDR. En este caso, el segundo coeficiente aporta más información de clase en comparación con el criterio de variabilidad, tomando valores menores de 50 aproximadamente para los controles y ampliándose el rango hasta 200 para los pacientes con AD.

6.3. PCA localizado

PCA puede aplicarse al volumen entero formado por los $n = 79 \times 95 \times 69 \sim 5 \cdot 10^5$ voxels consiguiendo una notable reducción de la dimensión. Teóricamente, las CPs obtenidas a partir de este proceso contienen las principales diferencias entre los sujetos que componen la base de datos, entre las cuales se espera encontrar también aquellas que distinguen sujetos normales frente a sujetos AD. Sin embargo, cuando se tratan datos de dimensión tan elevada, PCA es sensible a cualesquiera otras diferencias estadísticas, quizás derivadas del proceso de normalización o ruido y que no son útiles para clasificación.

Como método descriptivo de una base de datos, PCA puede proporcionar información sobre las características más relevantes en el espacio de características n-dimensional. En la secciones anteriores se proponen métodos para seleccionar aquéllas CPs de interés en el problema de clasificación que nos concierne. Estas selección de características en el espacio transformado PCA puede ser complementado mediante una previa selección de características en el espacio original de características.

Un método de exploración de las imágenes es las descomposición de las mismas en cortes $2D$ a lo largo de las tres dimensiones axial, sagital y coronal. Cada imagen se puede descomponer en:

$$\mathbf{x} = S_1 \cup S_2 \cup ... \cup S_s \tag{6.17}$$

donde $s = 69$, 79 ó 95 en función del eje de exploración axial, sagital y coronal respectivamente. Mediante la pre-selección de los voxels contenidos en el corte S_i el espacio de características se reduce en dos órdenes de magnitud con respecto al volumen completo. En este caso cada imagen la componen el conjunto de vectores sobre el que actuará la transformada PCA es $S = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_s]$, donde $\mathbf{s}_i$ es el vector columna que contiene los voxels del i-ésimo corte. Por tanto la transformación queda:

$$\Omega_j = \mathbf{U}_j^T \mathbf{s}_j = [\mathbf{u}_{1j}, \mathbf{u}_{2j}, ..., \mathbf{u}_{mj}]^T \mathbf{s}_j \quad j = 1, ..., s. \tag{6.18}$$

donde $\mathbf{u}_{ij}$ es el i-ésimo autovector extraído de la matriz de covarianza construída con el j-ésimo corte de cada una de las muestras disponibles, m es el número de autovectores representativos elegido y s es el número de cortes. Los vectores $\mathbf{u}_{ij}$ son r-dimensionales, donde $r = 69 \times 79, 69 \times 95$ ó 95×79 en función del eje de exploración. Por tanto, en este tratamiento por cortes,

de cada sujeto se extraen tres conjuntos de vectores de pesos o coeficientes m-dimensionales $\mathbf{\Omega} = \{\mathbf{\Omega}_{axial}, \mathbf{\Omega}_{coronal}, \mathbf{\Omega}_{sagital}\}$, uno por cada eje de estudio, donde m puede variar de un conjunto a otro y entre los vectores que componen un mismo conjunto.

Haciendo uso de estos vectores de pesos y mediante un aprendizaje supervisado llevado a cabo con clasificadores SVM se puede obtener una representación de las bases de datos que revela la importancia de los cortes que componen las imágenes. Se trata de un método de búsqueda de cortes de interés (CDIs) que puede ser utilizada en posteriores tareas de clasificación como método de selección previa de voxels en el espacio de características. Éste método agiliza el proceso de obtención de componentes principales de los datos, siendo a su vez dichas componentes más significativas para propósitos de clasificación.

6.4. Experimentos

La técnica de extracción de características basada en PCA en combinación con diferentes métodos de clasificación ha sido probada sobre las bases de datos tanto ADNI y PET *Cartuja* de imágenes PET como la base de datos SPECT. En todos los experimentos las imágenes fueron submuestreadas por un factor de $2 \times 2 \times 2$. Por tanto, la aplicación de los algoritmos de extracción de características se llevan a cabo en volúmenes de voxels de tamaño $34 \times 47 \times 39 \sim 6 \cdot 10^4$. Éste es por tanto el tamaño del espacio de entrada en todos los experimentos donde PCA se aplica al volumen completo. Para la base de datos SPECT se realiza además un estudio de los volúmenes por cortes para localizar aquéllos con poder discriminante mayor. Por tanto en este estudio por cortes la dimensión del espacio de entrada se reduce a 34×39, 34×47 o 47×39 ó , dependiendo de si el eje bajo estudio es el axial, coronal o sagital respectivamente.

El objetivo principal de estos experimentos es determinar en un primer estudio preliminar la capacidad del análisis PCA como técnica de reducción de la dimensión del espacio de características y de discriminación de dichas características para distinguir entre pacientes etiquetados como NORMAL y pacientes etiquetados como AD. Asimismo, se pretende determinar la conveniencia del criterio Fisher como criterio de reordenación y selección de CPs. Una vez extraídas las características y reordenadas según el criterio más conveniente (varianza o Fisher), se pretende comprobar cómo se ajustan a ellas distintos clasificadores, evaluando su rendimiento mediante la variación de

parámetros inherentes a ellos. El número máximo de coeficientes PCA que se pueden extraer de una base de datos de N muestras es $N-1$. Los experimentos se llevan a cabo variando el número de coeficientes m empleados en las tareas de entrenamiento y test desde $m=1$ hasta $m=N-1$, para determinar en cada caso el número adecuado de coeficientes que es necesario para una exitosa separación de clases. Por otro lado, se evalúan clasificadores SVM con kernel lineal, cuadrático, polinómico y de tipo RBF ($\sigma=1$ cuando no se especifica su valor), clasificadores NN variando el número de neuronas de la capa oculta, y finalmente el clasificador bayesiano lineal y cuadrático según el modelo adoptado para el cálculo de las matrices de covarianza.

La descripción de los grupos sobre los que se aplica el proceso completo son:

- SPECT: 91 pacientes, de los cuales 50 están etiquetados como NORMAL y 41 como AD.
- PET *Cartuja*: 60 pacientes, de los cuales 18 están etiquetados como NORMAL y 42 están etiquetados como AD.
- ADNI (Grupo 1): 105 pacientes, de los cuales 52 están etiquetados como NORMAL y 53 como AD.

6.5. Resultados

6.5.1. Base de datos SPECT

Exploración por cortes

En primer lugar se muestra la capacidad descriptiva y a la vez discriminante de los coeficientes PCA cuando se aplica la transformada a los volúmenes divididos por cortes. La Figura 6.4 muestra la precisión obtenida cuando se usan los coeficientes PCA extraídos de cada corte, tomándose éstos en las tres direcciones (axial, sagital y coronal) de forma independiente. Como era de esperar, los CDIs corresponden con las regiones del cerebro donde principalmente se manifiesta la EA, es decir, el cíngulo posterior y el precunei, así como la región temporo-parietal. Además de la capacidad descriptiva de los coeficientes extraídos, éstos pueden ser usados posteriormente en

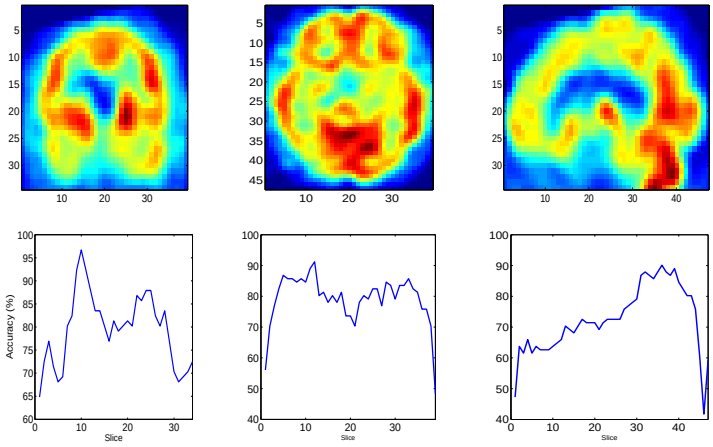


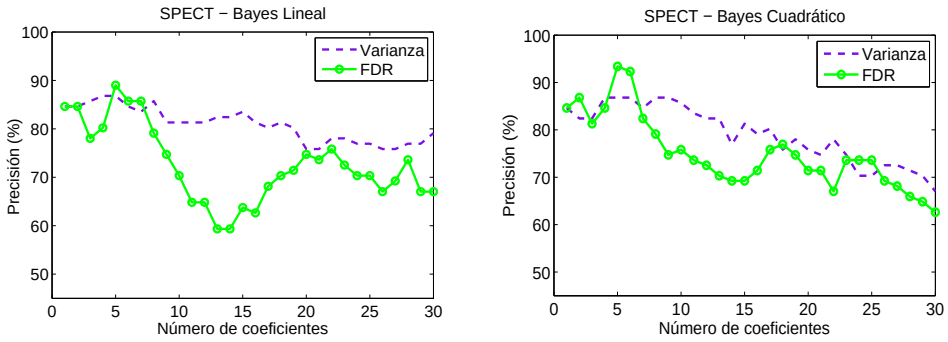
Figura 6.4: Cortes de interés (CDIs) localizados por medio de la transformación PCA y clasificadores SVM para cada eje del volumen.

clasificación para distinguir entre controles normales y sujetos AD en combinación con un clasificador, en este caso, basado en SVM con kernel lineal. Un conjunto completo de resultados de precisión obtenidos con este esquema de clasificación se muestra en la Tabla 13.1, donde m indica el número de coeficientes utilizados en la tarea de clasificación. El corte más discriminante resultó ser el décimo corte en la dirección axial, que haciendo uso de tan sólo 3 coeficientes extraídos de los voxels que lo componen dio lugar a un pico de precisión de 96,7 %.

SPECT

Eje	m/ CDI	SVM Lineal	SVM Cuadrático	SVM Polinómico	SVM RBF
Axial	3/10	87.91 %	**96.7 %**	90.11 %	93.41 %
	3/11	87.91 %	92.31 %	90.11 %	93.41 %
	4/10	86.81 %	90.11 %	85.71 %	92.31 %
	4/11	87.91 %	90.11 %	89.01 %	89.01 %
Coronal	3/11	89.01 %	91.21 %	84.62 %	86.81 %
	3/12	91.21 %	89.01 %	85.71 %	84.62 %
	4/11	89.01 %	87.91 %	81.32 %	85.71 %
	4/12	91.21 %	89.01 %	78.08 %	85.71 %
Sagital	1/35	89.01 %	86.81 %	86.81 %	87.91 %
	1/36	87.91 %	89.01 %	90.11 %	90.11 %

Tabla 6.1: Resultados obtenidos de la evaluación de SVM y coeficientes PCA extraídos por cortes.



(a) (b)

Figura 6.5: SPECT: Resultados de precisión frente al número de coeficientes PCA empleados en la clasificación cuando éstos se reordenan mediante los criterios de varianza y FDR, para un clasificador bayesiano (a) lineal y (b) cuadrático.

Volumen completo

Tras el estudio de los volúmenes divididos en cortes, se realiza la evaluación de los coeficientes PCA como características para clasificación cuando dichos coeficientes se extraen del volumen completo. Estas características se combinan con diferentes clasificadores obteniéndose los resultados de la Tabla 6.2. Para todos los datos de precisión de los clasificadores se presentan los correspondientes valores de sensibilidad y especificidad.

En el caso del clasificador Bayesiano, la reordenación de los coeficientes PCA según el criterio de Fisher (FDR) produce notables mejoras en los resultados de clasificación con respecto al clásico criterio de varianza, y en particular para el modelo cuadrático, como se muestra en la Figura 6.5. En esta figura los resultados de precisión se representan en función del número de coeficientes PCA empleado en la fase de entrenamiento y test. Para ambos modelos, existe un máximo en la precisión obtenida cuando se entrena el sistema con $m = 5$ coeficientes, alcanzando tasas de precisión de 89,01 % y 93,41 % para los modelos lineal y cuadrático respectivamente.

La reordenación de los coeficientes por criterio de Fisher no mejora cuantitativamente los resultados de precisión obtenidos con respecto al criterio de la varianza cuando se emplean clasificadores SVM. Sin embargo, las curvas de precisión obtenidas presentan menos variabilidades presentando formas muy similares cuando se varían los parámetros del clasificador. Esto implica mayor robustez de los coeficientes PCA empleados en la clasificación al ser

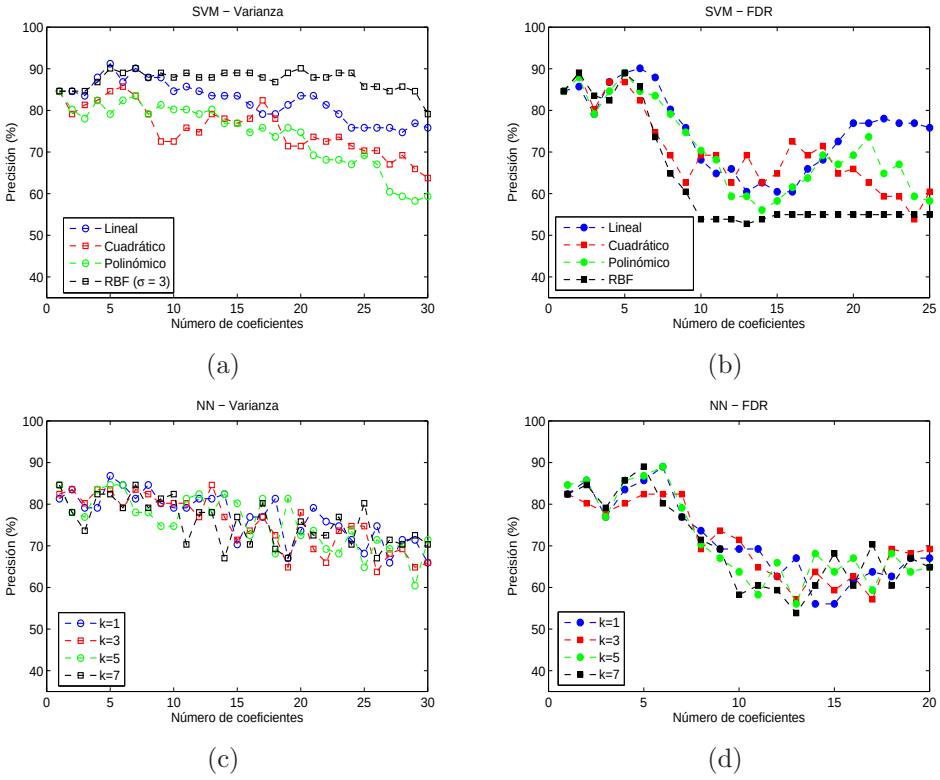


Figura 6.6: SPECT: Resultados de precisión frente al número de coeficientes PCA empleados en la clasificación para SVM y NN cuando los coeficientes se reordenan mediante los criterios de varianza y FDR.

los resultados menos dependientes del clasificador empleado.

Para el caso de clasificadores NN, de nuevo la aplicación del criterio de selección FDR mejora los resultados en este caso tanto cuantitativamente como cualitativamente, al presentar curvas menos variables y picos más claros. Al igual que con clasificadores bayesianos y SVM, para NN queda más claramente determinado el número óptimo de coeficientes PCA necesario para llevar a cabo la clasificación, como se muestra en la Figura 6.6. Estos picos se producen cuando el número de coeficientes seleccionado es $m = 5$ y $m = 6$.

6.5.2. Base de datos PET *Cartuja*

Las curvas de precisión obtenidas mediante la técnica PCA y los distintos clasificadores aplicada a las imágenes PET *Cartuja* se muestran en la Figura 6.7. Para esta base de datos, la extracción de características basada en la

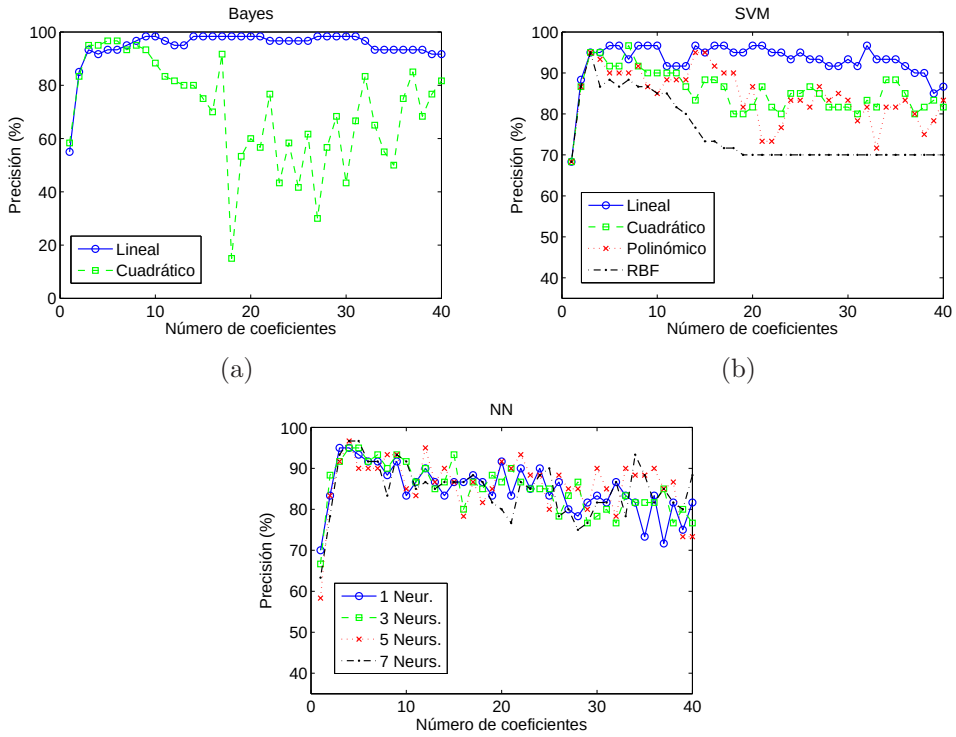


Figura 6.7: PET *Cartuja*: Resultados de precisión frente a número de coeficientes obtenidos mediante la evaluación de clasificadores basados en (a) Regla de Bayes, (b) SVM y (c) NN.

transformada PCA demuestra ser adecuada para lograr una separación de clases, y más concretamente cuando estas características se combinan con métodos lineales de separación, como son el clasificador bayesiano lineal y SVM con kernel lineal. En concreto el clasificador bayesiano lineal alcanza una precisión de 98,33 % a partir de 14 coeficientes ordenados por varianza y empleados para el entrenamiento y test, y mantiene estos resultados de forma estable hasta 21 coeficientes. SVM y NN alcanzan una precisión de 96,67 %, sin embargo esta tasa aparece más como un pico de precisión a lo largo de la evaluación de los clasificadores. De estas gráficas se puede extraer la conclusión de la idoneidad de usar Bayes lineal para la separación de las clases NORMAL y AD de esta base de datos.

Para la base de datos PET también se produce una mejora al reordenar las características mediante el criterio FDR, aunque menos significativa que para la base de datos SPECT. Aún así, SVM con kernel lineal alcanza un pico de precisión de hasta 100 % para cuando se seleccionan $m = 7$ coeficientes.

Éstos y otros resultados se recogen en la Tabla 6.2.

6.5.3. Base de datos ADNI

Los resultados de clasificación haciendo uso del grupo 1 (NORMAL frente a AD) de la base de datos ADNI se muestran en la Figura 6.8. De nuevo las técnicas lineales de separación muestran un comportamiento más estable a la vez que producen las tasas de precisión mayores. Cuando el criterio de varianza es el empleado en el orden de selección de los coeficientes, las tasas máximas de precisión alcanzadas son 86,67 %, 85,71 % y 86,67 % para los clasificadores SVM con kernel lineal, NN con 1 neurona en la capa oculta y Bayes lineal (Figuras 6.8(a), 6.8(c) y 6.8(e) respectivamente). Estos resultados se mejoran cuando se seleccionan los coeficientes de acuerdo al criterio de Fisher, obteniendose entonces 91,43 %, 87,62 % y 88,57 % para los mismos clasificadores. El empleo de FDR además aporta la ventaja de alcanzar estos picos de precisión con un número menor de coeficientes. Además, el comportamiento de los clasificadores se uniformiza encontrando los picos para una cantidad similar de coeficientes PCA. Así, observando las curvas de precisión de las gráficas 6.8(b), 6.8(d) y 6.8(f) se observa que los picos de precisión se encuentran en torno a 3, 8 y 15 coeficientes para todos los clasificadores. Las superficies de decisión diseñadas por varios clasificadores se muestran en el conjunto de gráficas de la Figura 6.9, donde se representan los tres primeros coeficientes PCA como puntos tridimensionales. En todas ellas los pacientes normales se vienen indicados con cruces azules, los pacientes AD con asteriscos rojos, y en verde las fronteras de separación que cada regla de clasificación proporciona. Las reglas más sencillas vienen determinadas por los clasificadores lineales mientras que otros como SVM RBF o NN con $k = 3$ determinan reglas complejas de separación, en muchos casos más adecuadas que una simple separación lineal.

6.5.4. Resumen y comparación con VAF

En la Tabla 6.2 se muestra una recopilación de los mejores resultados obtenidos mediante la técnica PCA empleada en este capítulo con la finalidad de compararlos con el método VAF tomado como referencia. Todos los resultados de precisión se complementan con los respectivos valores de sensibilidad y especificidad. Independientemente de la base de datos probada y del clasificador empleado, el uso de los coeficientes PCA como características

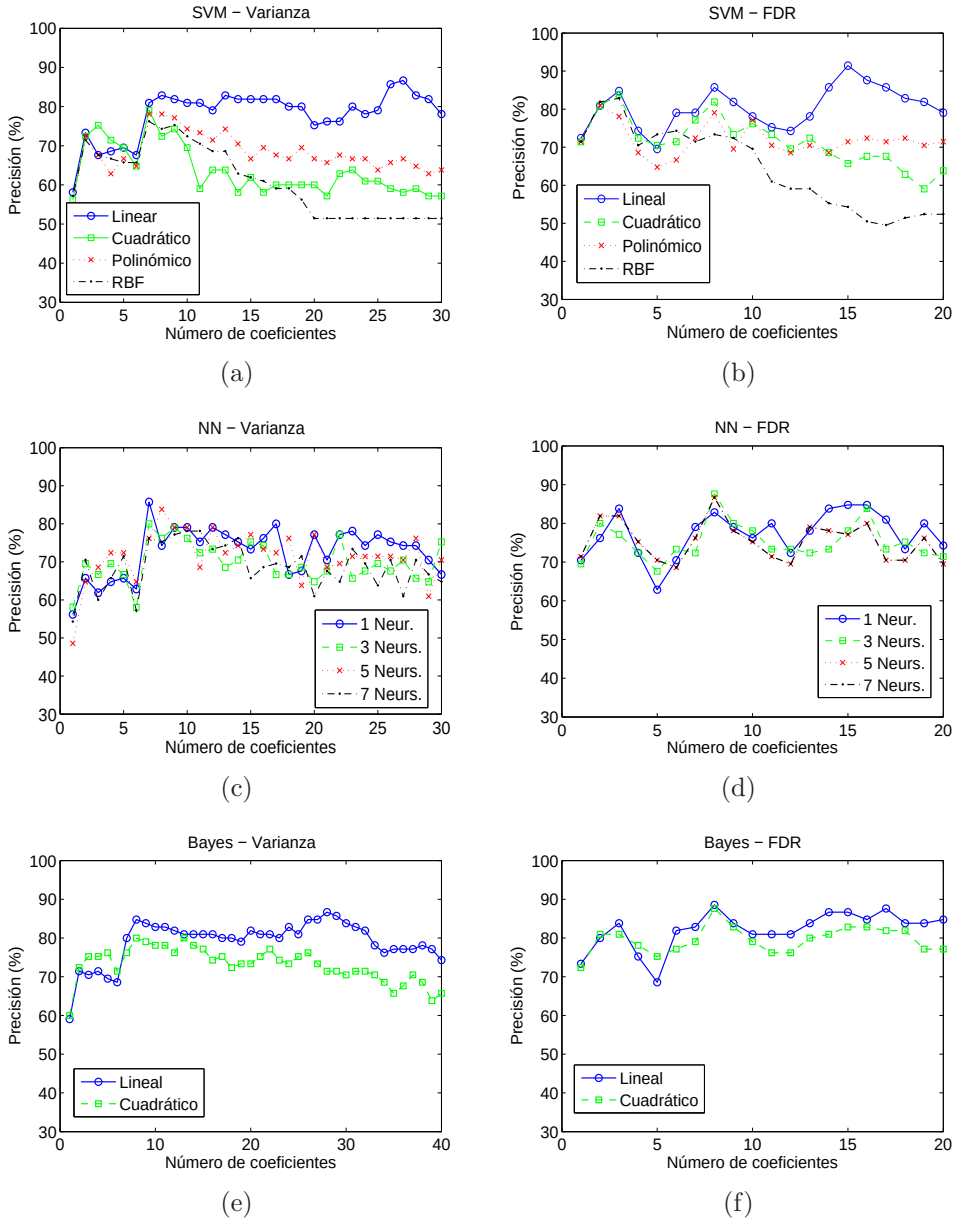


Figura 6.8: Grupo 1 de ADNI: Resultados de la clasificación para distintos clasificadores y criterios de selección de coeficientes por varianza y FDR.

supone una mejora en la mayoría de los casos con respecto a la técnica VAF. Además, la reordenación de dichos coeficientes mediante el criterio de Fisher proporciona mejoras en el rendimiento de los sistemas dando consistencia a las características para todas las bases de datos y aumentando por tanto los valores de precisión alcanzados.

	SPECT		PET *Cartuja*		ADNI (Grupo 1)	
	Varianza	FDR	Varianza	FDR	Varianza	FDR
VAF	85.71 % (88/82.93) %	-	96.67 % (97.62/94.44) %	-	83.81 % (84.91/82.69) %	-
SVM Lineal	91.21 % (90/92.68) %	90.11 % (92/87.80) %	96.67 % (97.62/94.44) %	100 % (100/100) %	86.67 % (86.79/86.54) %	91.43 % (94.34/88.46) %
SVM Cuad.	85.71 % (84/87.80) %	89.01 % (88/90.24) %	96.67 % (97.62/94.44) %	96.67 % (100/88.89) %	79.05 % (79.24/78.85) %	83.81 % (83.02/84.61) %
SVM Polin.	83.52 % (78/90.24) %	89.01 % (94/82.93) %	95 % (95.24/94.44) %	98.33 % (97.62/100) %	78.1 % (77.36/78.85) %	80.95 % (83.02/87.85) %
SVM RBF	90.11 % (92/85.37) %	89.01 % (90/87.81) %	95 % (97.62/88.89) %	95 % (97.62/88.89) %	76.19 % (90/90.24) %	82.86 % (84.91/80.77) %
Bayes Lineal	86.81 % (82/92.68) %	89.01 % (86/92.68) %	98.33 % (97.62/100) %	98.33 % (97.62/100) %	86.67 % (88.68/84.61) %	88.57 % (86.79/90.38) %
Bayes Cuad.	86.81 (88/85.36) %	93.41 % (94/92.68) %	96.67 % (97.62/94.44) %	95 % (95.24/94.44) %	80 % (81.13/78.85) %	87.62 % (90.57/84.61) %
NN 1 Neur.	86.81 % (82/92.68) %	89.01 % (92/85.36) %	95 % (95.24/94.44) %	96.67 % (97.62/94.44) %	85.71 % (86.79/84.61) %	84.76 % (80.79/82.69) %
NN 3 Neurs.	84.62 % (82/87.80) %	82.42 % (84/80.49) %	95 % (92.86/100) %	95 % (95.24/94.44) %	80 % (71.7/88.46) %	87.62 % (86.79/88.46) %
NN 5 Neurs.	84.62 % (86/80.49) %	89.01 % (92/85.36) %	96.67 % (97.62/94.44) %	96.67 % (97.62/94.44) %	83.81 % (84.91/82.69) %	86.67 % (88.67/88.61) %
NN 7 Neurs.	84.61 % (84/85.36) %	89.01 % (90/87.80) %	96.67 % (95.24/100) %	96.67 % (97.62/94.44) %	78.1 % (79.24/79.92) %	86.67 % (88.67/88.61) %

Tabla 6.2: Resumen de los mejores resultados obtenidos y comparación con el método VAF tomado como referencia.

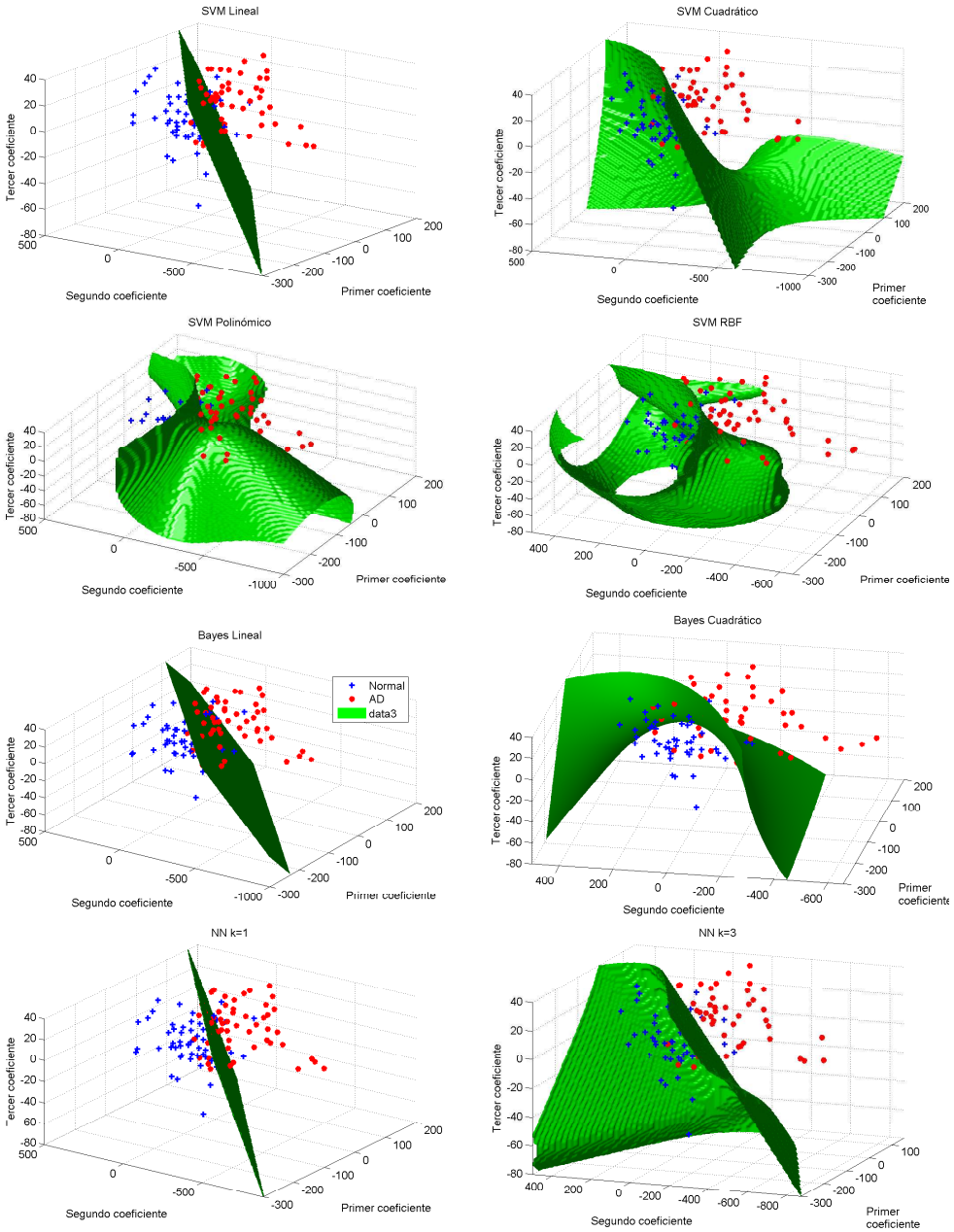


Figura 6.9: Superficies de separación definidas por diferentes clasificadores. Cada paciente viene representado por sus tres coeficientes principales PCA como puntos tridimensionales.

CAPÍTULO 7

Análisis Discriminante Lineal

El análisis discriminante lineal (Linear Discriminant Analysis o LDA) [Fisher, 1936] es una de las técnicas de proyección lineales más populares usada como técnica de extracción de características. El objetivo del análisis discriminante puede resumirse como encontrar una función que devuelva valores escalares que permitan una buena discriminación entre diferentes clases de los datos de entrada. Estos discriminantes se usan posteriormente para entrenar por ejemplo un clasificador o para visualizar ciertos aspectos de los datos. En este sentido, el análisis discriminante puede entenderse como un método de procesado o de extracción de características supervisado, ya que se le proporciona al algoritmo de aprendizaje información acerca de la conexión que existe entre las características y los ejemplos de entrenamiento. En este capítulo se muestra la eficiencia de esta técnica al aplicarla sobre imágenes neurológicas para la extracción de información así como su poder de discriminante en la tarea de detección de la EA.

7.1. Análisis Discriminante Lineal

Un modo de ver un modelo de clasificación lineal es en términos de reducción de la dimensión. Consideramos el caso de dos clases, y supongamos que se toma el vector de entrada $\mathbf{x}$ n-dimensional y se proyecta en una dimensión usando

$$y = \mathbf{w}^T\mathbf{x} \tag{7.1}$$

Si se establece un umbral sobre y y se clasifica $y \geq -w_0$ como clase ω_1 y como clase ω_2 en caso contrario, entonces se obtiene un clasificador lineal estándar como el de la ecuación 4.1. En general, la proyección en una sola dimensión conduce a una considerable pérdida de información, y las clases que se pueden separar correctamente en el espacio n-dimensional pueden llegar a estar fuertemente solapadas en una dimensión. Sin embargo, mediante el ajuste de las componentes del vector $\mathbf{w}$ se puede seleccionar una proyección que maximice la separación de clases. Consideramos el caso de dos clases en las que existen N_1 elementos en la clase ω_1 y N_2 elementos en la clase ω_2, de modo que los vectores media de ambas clases vienen dados por:

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n\in\omega_1}\mathbf{x}_n, \quad \boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n\in\omega_2}\mathbf{x}_n \tag{7.2}$$

La medida más simple de separación de clases cuando se proyecta sobre $\mathbf{w}$ es la separación entre las proyecciones de las medias. Esto sugiere elegir $\mathbf{w}$ de modo que maximice

$$\mu_1 - \mu_2 = \mathbf{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{7.3}$$

donde

$$\mu_k = \mathbf{w}^T\boldsymbol{\mu}_k \tag{7.4}$$

es la media de los datos proyectados de la clase ω_k. Sin embargo, ésta puede hacerse arbitrariamente grande simplemente incrementando la magnitud de $\mathbf{w}$. Para resolver este problema se puede imponer la condición de que $\mathbf{w}$ tenga norma unidad, de modo que $\sum_i {w_i}^2 = 1$. Usando un multiplicador de Lagrange para llevar a cabo la maximización con restricciones se llega a que

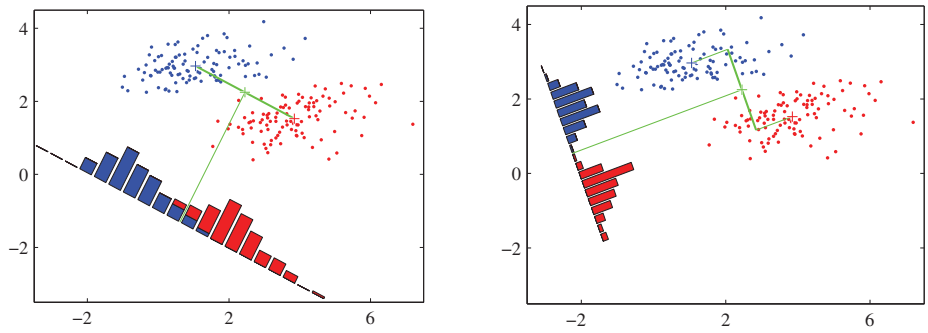


Figura 7.1: Gráficos ilustrativos de la proyección LDA. En el gráfico de la izquierda se representan muestras de dos clases (pintadas en rojo y azul) junto con sus histogramas resultantes de la proyección en la línea que une las medias. Nótese que existe un solapamiento considerable en el espacio proyectado. El gráfico de la derecha muestra la proyección correspondiente basada en el discriminante lineal de Fisher, mostrando que se mejora en gran medida la separación de clases.

$\mathbf{w} \propto (\mu_1 - \mu_2)$. Existe todavía un problema con esta aproximación, como se muestra en la Figura 7.1. Esta figura muestra dos clases que se pueden separar correctamente en el espacio original bidimensional (x_1, x_2) pero existe un solapamiento considerable cuando se proyectan en la línea que une sus medias. Esta dificultad surge debido a las covarianzas fuertemente no diagonales de las distribuciones de las clases. La idea propuesta por Fisher es maximizar una función que dé una separación alta entre las medias de las clases proyectadas a la vez que una pequeña varianza intra-clase, minimizando por tanto el solapamiento entre clases [Bishop, 2006].

La fórmula de proyección 7.1 transforma el conjunto de datos etiquetados X en un espacio unidimensional etiquetado Y.

La varianza intra-clase de los datos transformados pertenecientes a la clase ω_k definida en la ecuación 14.11, viene dada con esta nueva notación por

$$s_k^2 = \sum_{n \in \omega_k} (y_n - \mu_k)^2 \tag{7.5}$$

donde $y_n = \mathbf{w}^T x_n$. Para dos clases se puede definir la varianza total intra-clase para el conjunto completo de datos como $s_1^2 + s_2^2$. El criterio de Fisher se define como la razón entre la varianza inter-clase y la varianza intra-clase

y viene dado por:

$$J(\mathbf{w}) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2} \tag{7.6}$$

Se puede hacer explícita la dependencia con respecto a $\mathbf{w}$ usando las ecuaciones 7.1, 7.4 y 7.5 para reescribir el criterio de Fisher de la forma

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} \tag{7.7}$$

donde $\mathbf{S}_b$ y $\mathbf{S}_w$ son las matrices de covarianza inter-clase e intra-clase respectivamente, y se definen como

$$\mathbf{S}_b = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \tag{7.8}$$

$$\mathbf{S}_w = \sum_{n \in \omega_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T + \sum_{n \in \omega_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T \tag{7.9}$$

Derivando la ecuación 7.7 con respecto a $\mathbf{w}$ se llega a que $J(\mathbf{w})$ se maximiza cuando

$$(\mathbf{w}^T \mathbf{S}_b \mathbf{w}) \mathbf{S}_w \mathbf{w} = (\mathbf{w}^T \mathbf{S}_w \mathbf{w}) \mathbf{S}_b \mathbf{w} \tag{7.10}$$

De la ecuación 14.11 se puede ver que $\mathbf{S}_b \mathbf{w}$ siempre está en la dirección de $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Además, no afecta la magnitud de $\mathbf{w}$, sino su dirección, y por tanto se pueden eliminar los factores escalares $(\mathbf{w}^T \mathbf{S}_b \mathbf{w})$ y $(\mathbf{w}^T \mathbf{S}_w \mathbf{w})$. Multiplicando ambos lados de la ecuación 7.10 por $\mathbf{S}_w^{-1}$ se obtiene

$$\mathbf{w} \propto \mathbf{S}_w^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \tag{7.11}$$

Nótese que si la matriz de covarianza inter-clase es isotrópica, de modo que $\mathbf{S}_w$ es proporcional a la matriz unidad, entonces $\mathbf{w}$ es proporcional a la diferencia entre las medias de las clases, como se discutió anteriormente.

El resultado 7.11 se conoce como el *discriminante lineal de Fisher*, aunque estrictamente no es un discriminante sino más bien la elección de una dirección de proyección específica de los datos en una dimensión. Sin embargo,

los datos proyectados pueden usarse posteriormente para construir un discriminante, mediante la elección de un umbral y_0 de modo que un nuevo punto se clasifica como perteneciente a la clase ω_1 si $y(\mathbf{x}) \geq y_0$ y se clasifica como perteneciente a la clase ω_2 en caso contrario. Para la regla de Bayes por ejemplo, se pueden modelar las densidades de clase condicionales $p(y|\omega_k)$ usando distribuciones Gaussianas mediante máxima verosimilitud (la asunción de distribuciones Gaussianas se justifica con el teorema central del límite observando que $y = \mathbf{w}^T\mathbf{x}$ es la suma de un conjunto de variables aleatorias) y fijar este umbral y_0 como aquel que minimiza la probabilidad del error.

7.2. LDA para múltiples clases

Consideremos ahora la generalización del discriminante de Fisher para un número de clases $c > 2$ asumiendo que la dimensión n del espacio de entrada es mayor que el número de clases c. A continuación, se introducen $l > 1$ características lineales $y_k = \mathbf{w}_k^T\mathbf{x}$, donde $k = 1, ..., l$. Estos valores de las características se pueden agrupar para formar el vector $\mathbf{y}$. De igual modo, los vectores de pesos $\{\mathbf{w}_k\}$ se pueden considerar columnas de la matriz $\mathbf{W}$ de modo que

$$\mathbf{y} = \mathbf{W}^T\mathbf{x} \tag{7.12}$$

La generalización de la matriz de covarianza intra-clase para el caso de c clases lleva la ecuación 14.12 a

$$\mathbf{S}_w = \sum_{k=1}^{c} \mathbf{S}_k \tag{7.13}$$

donde

$$\mathbf{S}_k = \sum_{n \in \omega_k} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{7.14}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in \omega_k} \mathbf{x}_n \tag{7.15}$$

y N_k es el número de patrones en la clase ω_k. Con el fin de generalizar la matriz de covarianza inter-clase y siguiendo Duda and Hart [1973], consideremos primero la matriz de covarianza total

$$\mathbf{S}_T = \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T \tag{7.16}$$

donde $\boldsymbol{\mu}$ es la media del conjunto total de datos

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \tag{7.17}$$

La matriz de covarianza total puede descomponerse en la suma de la matriz de covarianza intra-clase e inter-clase dadas por las ecuaciones 7.13 y 7.19 respectivamente,

$$\mathbf{S}_T = \mathbf{S}_w + \mathbf{S}_b \tag{7.18}$$

donde

$$\mathbf{S}_b = \sum_{k=1}^{c} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T \tag{7.19}$$

Estas matrices se pueden definir igualmente en el espacio proyectado. De nuevo, la idea es construir un escalar que sea grande cuando la covarianza inter-clase sea grande y cuando la covarianza inter-clase sea pequeña en el espacio proyectado. Existen varias posibilidades [Fukunaga, 1990]. Un criterio posible, reescrito de forma explícita en función de la matriz de proyección $\mathbf{W}$ es

$$\mathbf{W}_{lda} = \arg \max_{\mathbf{W}} \frac{\left|\mathbf{W}^T \mathbf{S}_b \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_w \mathbf{W}\right|} = [\mathbf{w}_1 \mathbf{w}_2 ... \mathbf{w}_l] \tag{7.20}$$

La maximización de este criterio es bastante directo, aunque a veces complicado como se discute en Fukunaga [1990]. Los vectores $\{\mathbf{w}_k\}$ quedan determinados por los autovectores de $\mathbf{S}_w^{-1}\mathbf{S}_b$ que correspondientes a los l autovalores mayores asociados.

7.3. Limitaciones de LDA

En problemas de reconocimiento de patrones y clasificación de imágenes uno debe hacer frente a la dificultad que presenta la matriz de dispersión intra-clase $\mathbf{S}_w$ ya que suele ser una matriz singular. Esto surge del hecho de que el rango de $\mathbf{S}_w$ es como mucho $N - c$, y en general, el número N de imágenes del conjunto de entrenamiento es mucho más pequeño que el número de voxels n que representa a cada sujeto. Esto implica que es posible elegir una matriz $\mathbf{W}$ tal que la matriz de dispersión intra-clase de las imágenes proyectadas sea exactamente cero. Para solventar este problema, Belhumeour et al. [1997] propone el uso de un espacio intermedio. El espacio propuesto es justamente el espacio PCA [Yang and Yang, 2003]. Este método evita el problema mediante la proyección del conjunto de imágenes en un espacio de dimensión menor de modo que la matriz de dispersión intra-clase $\mathbf{S}_w$ sea no singular. Esto se consigue usando PCA para reducir la dimensión del espacio de características hasta $N - c$ y después aplicando la transformada LDA definida en la ecuación 13.8 para reducir la dimensión a $c - 1$ [Duda and Hart, 1973]. Más formalmente, $\mathbf{W}_{opt}$ sería

$$\mathbf{W}_{opt}^T = \mathbf{W}_{lda}^T \mathbf{W}_{pca}^T \tag{7.21}$$

donde $\mathbf{W}_{pca}$ es la matriz de proyección definida por la transformada PCA, es decir, $\mathbf{W}_{pca} = [\mathbf{u}_1 \mathbf{u}_2 ... \mathbf{u}_m]$, donde los vectores $\mathbf{u}_k$, $k = 1, ..., m$ son los autovectores de la matriz de covarianza de los vectores del espacio original, como se definió en la ecuación 6.13, o expresado de otro modo como

$$\mathbf{W}_{pca} = \arg\max_{\mathbf{W}} \left| \mathbf{W}^T \mathbf{S}_T \mathbf{W} \right| \tag{7.22}$$

Por tanto, la matriz de proyección LDA queda finalmente

$$\mathbf{W}_{lda} = \arg\max_{\mathbf{W}} \frac{\left| \mathbf{W}^T \mathbf{W}_{pca}^T \mathbf{S}_b \mathbf{W}_{pca} \mathbf{W} \right|}{\left| \mathbf{W}^T \mathbf{W}_{pca}^T \mathbf{S}_w \mathbf{W}_{pca} \mathbf{W} \right|} \tag{7.23}$$

Nótese que la optimización para $\mathbf{W}_{pca}$ se realiza sobre matrices de dimensión $n \times (N - c)$ con columnas ortonormales mientras que la optimización para $\mathbf{W}_{lda}$ se lleva a cabo sobre matrices de tamaño $(N - c) \times m$ con columnas ortonormales. En el cálculo de $\mathbf{W}_{pca}$ se descartan sólo las $c - 1$ componentes principales más pequeñas. Otra alternativa al problema de la no singularidad

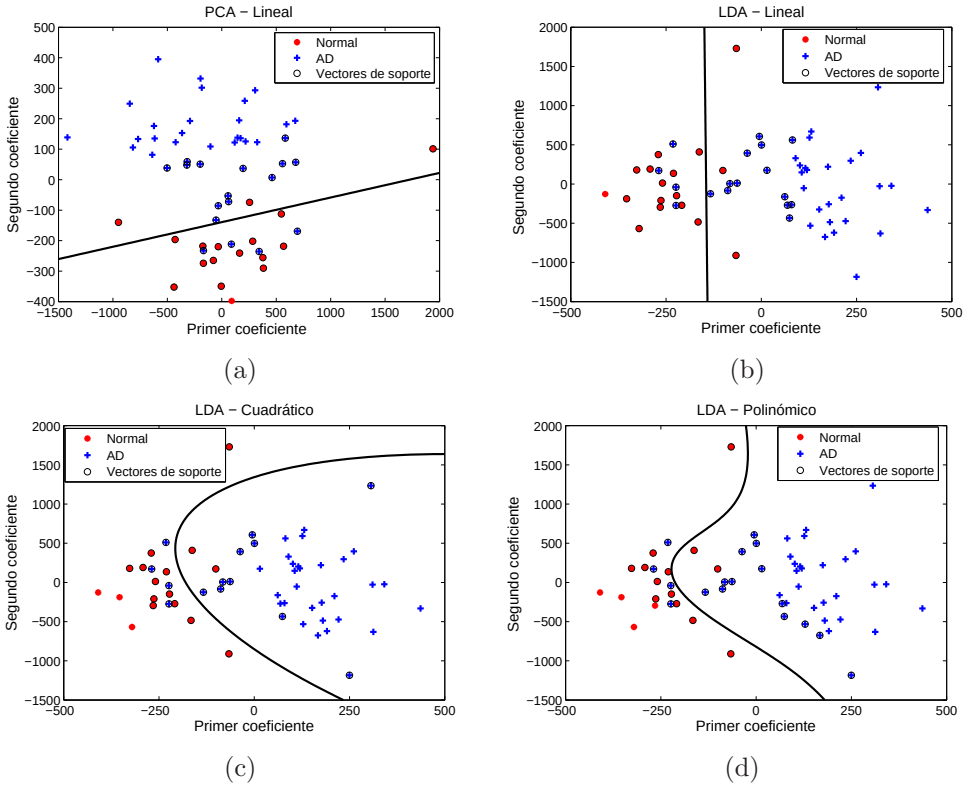


Figura 7.2: Distribución de los dos primeros coeficientes (a) PCA y (b), (c) y (d) LDA para las imágenes PET *Cartuja* y fronteras de decisión establecidas por un clasificador SVM (a) y (b) lineal, (c) cuadrático y (d) polinómico. El gráfico (b) muestra cómo el primer coeficiente extraído mediante la proyección sobre un eje LDA conserva toda la información de clase.

de $\mathbf{S}_b$ se propone en Yu and Yang [2000], que resuelve el problema mediante la aplicación directa de LDA a datos de alta dimensión sin la necesidad de aplicar pasos de reducción de la dimensión.

La Figura 7.2 muestra cómo se transforman las dos primeras características extraídas de las imágenes PET *Cartuja* mediante la aplicación de la transformadas $\mathbf{W}_{pca}$ (Figura 7.2(a)) y $\mathbf{W}_{lda}$. Una vez aplicada la proyección sobre el eje LDA, el primer coeficiente se convierte en la característica más relevante para la clasificación, como muestra la Figura 7.2(b), donde un clasificador SVM con kernel lineal establece la frontera de decisión como un umbral únicamente dependiente de este primer coeficiente. La variedad de clasificadores no obliga a emplear una regla de decisión lineal, sino que estas características pueden combinarse con otros kernels como muestran las

Figuras 7.2(c) y 7.2(d), pudiendo resultar una frontera de separación más adecuada.

7.4. Experimentos

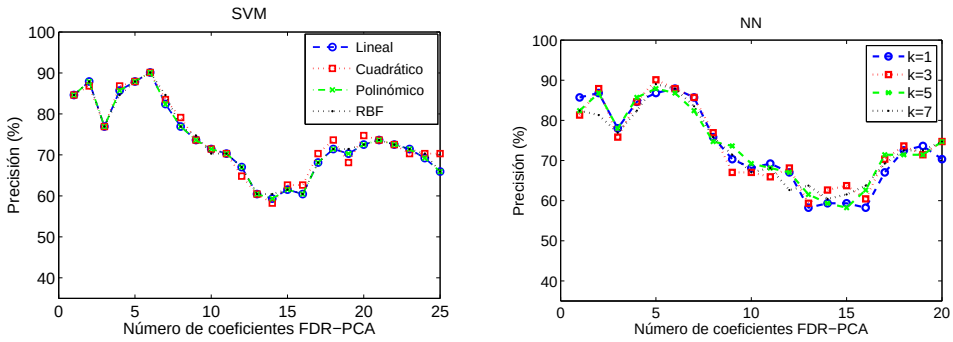
La aplicación de la transformada LDA se lleva a cabo sobre los coeficientes PCA obtenidos según se explica en el Capítulo 6. Sobre las imágenes se realizan los mismos pasos de muestreo descritos en la sección 6.4. En este caso se pretende analizar la posible mejora introducida mediante la transformación LDA sobre los coeficientes PCA ya extraídos en el capítulo anterior. Los grupos sobre los que se experimenta esta transformada son:

- SPECT: 91 pacientes, de los cuales 50 están etiquetados como NORMAL y 41 como AD.
- PET *Cartuja*: 60 pacientes, de los cuales 18 están etiquetados como NORMAL y 42 están etiquetados como AD.
- Grupo 1 de ADNI: 105 pacientes, de los cuales 52 están etiquetados como NORMAL y 53 como AD.
- Grupo 2 de ADNI: 166 pacientes, de los cuales 52 están etiquetados como NORMAL y 114 como MCI.

7.5. Resultados

7.5.1. Base de datos SPECT

Para la clasificación llevada a cabo sobre las imágenes SPECT mediante clasificadores SVM y NN, la Figura 7.3 muestra los resultados de precisión obtenidos mediante LDA en función del número de coeficientes PCA sobre los cuales se aplica la transformada. En todos los experimentos estos coeficientes se seleccionan mediante el criterio FDR y se proyectan mediante LDA en una única dimensión dando lugar a una sola característica empleada para el entrenamiento y test, es decir, el espacio final de características es en todos los casos unidimensional.



(a) (b)

Figura 7.3: SPECT: Resultados de precisión frente a número de coeficientes FDR-PCA sobre los que se aplica LDA obtenidos mediante la evaluación de clasificadores basados en (a) SVM y (b) NN.

Para clasificadores SVM existe un pico de precisión de 90,11 % igualando la precisión alcanzada sin aplicar LDA. Sin embargo, cabe destacar que este máximo se mantiene para todos los kernels cuando se proyectan $m = 6$ coeficientes PCA reordenados mediante el criterio FDR. Esta uniformidad en los resultados implica mayor consistencia de las características extraídas y empleadas en el entrenamiento y test de los clasificadores. Los valores de sensibilidad y especificidad alcanzados para las cuatro modalidades de kernel son 92 y 87,81 % respectivamente, por lo que cualquiera de ellos se puede considerar adecuado.

Para clasificadores NN, el máximo de precisión obtenida es de 91,21 % para $k = 3, k = 5$ y $m = 3, m = 4$ neuronas en la capa oculta y coeficientes FDR-PCA respectivamente, mejorando levemente la precisión de 90,11 % obtenida cuando no se aplica LDA. En este caso los valores de sensibilidad y especificidad correspondientes a estos picos de precisión son 94 y 87.81 % respectivamente.

Para el clasificador basado en la regla de Bayes los resultados empeoran levemente cuando LDA se aplica a los coeficientes PCA, por lo que no se muestran los resultados.

7.5.2. Base de datos PET *Cartuja*

En la sección 6.5.2 se demostró la capacidad de los coeficientes PCA para discriminar entre las muestras de las dos clases contenidas en esta base

de datos, alcanzando en todos los casos más de un 95 % de precisión. La aplicación de LDA todavía implica una mejora cuantitativa pues conduce a resultados de 100 % de precisión usando una única característica descriptora de cada muestra. Esto supone un éxito tanto en la tarea de clasificación como en la resolución del problema del pequeño tamaño muestral, alcanzando el máximo de precisión posible haciendo uso del mínimo número de características. En la Figura 7.4 se muestran las mejores curvas de precisión obtenidas para algunos clasificadores.

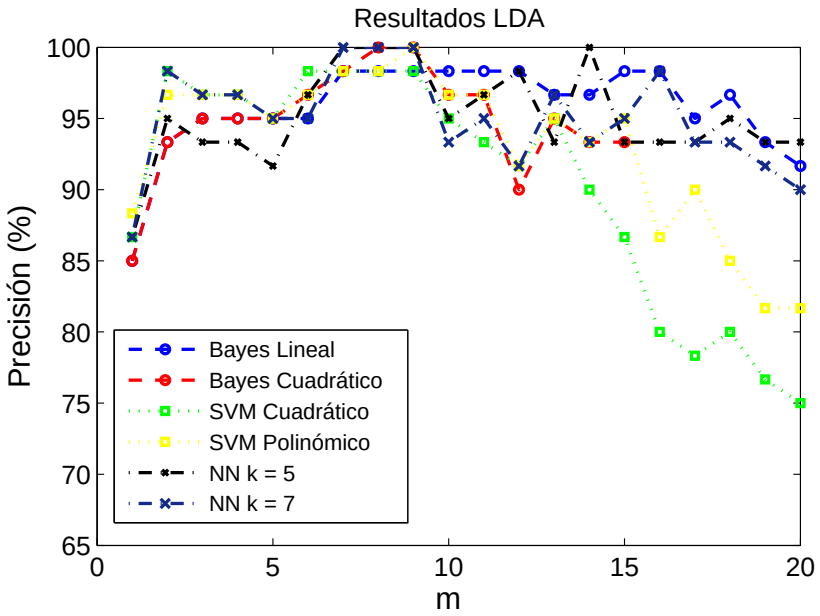


Figura 7.4: PET *Cartuja*: Curvas de precisión obtenidas mediante la proyección LDA cuando el número de coeficientes FDR-PCA aumenta.

En la base de datos PET encontramos dos sujetos cuya clasificación es en la mayor parte de los casos errónea. En el proceso de clasificación de una muestra de test mediante PCA y LDA, dicha muestra sufre dos proyecciones lineales que dan lugar a una característica final, donde los dos conjuntos de vectores base sobre los que se proyecta se han obtenido sin incluir la muestra de test en el cálculo. Por tanto, en ocasiones la proyección final queda alejada del rango expandido por las muestras de su clase. En la Figura 7.5 se representan las características obtenidas para cada paciente y la distancia a la posición media de las características de su clase para diferentes valores de m coeficientes PCA seleccionados. Los pacientes situados en las posiciones 10 y 21 y etiquetados como NORMAL y AD respectivamente tienden a alejarse de la posición media de su clase, y sólo se clasifican correctamente cuando se

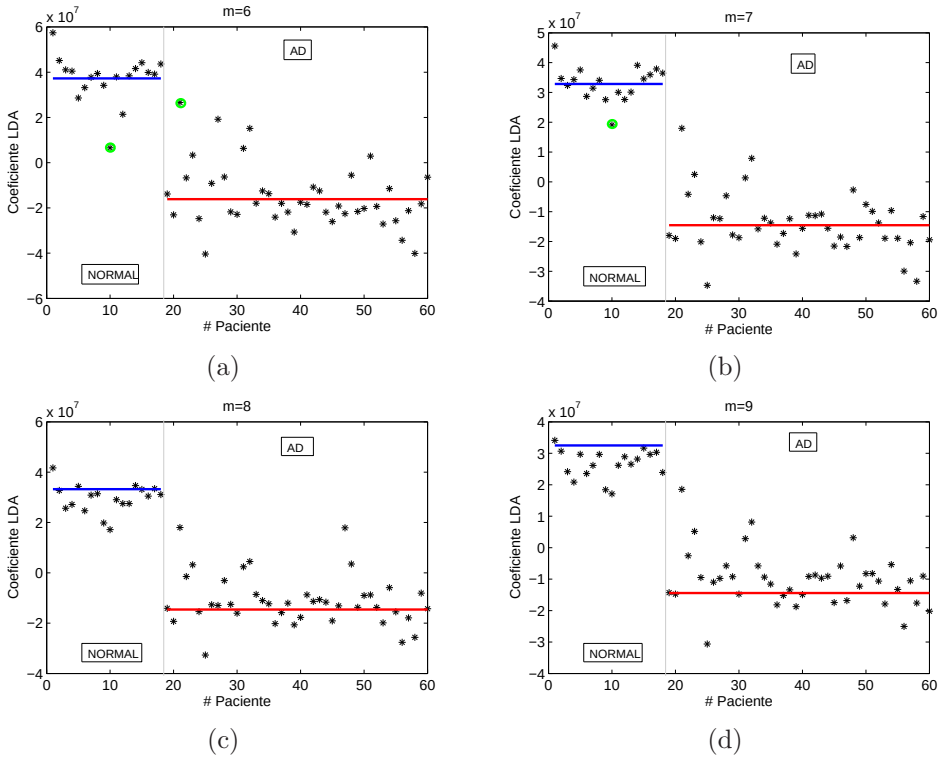


Figura 7.5: PET *Cartuja*: Posición de la característica final LDA que representa a cada paciente y posición media de las características de su clase para diferentes valores de m. En verde se marcan los pacientes mal clasificados.

ajusta de forma precisa el número de coeficientes PCA empleados en la clasificación. En verde se señalan los casos en los que son clasificados erróneamente por un clasificador SVM cuadrático.

7.5.3. Base de datos ADNI

Grupo 1

De nuevo la aplicación del criterio FDR de selección de coeficientes PCA seguido de una transformación LDA mejora los resultados cuando se considera el grupo 1 de la base de datos ADNI para clasificación. Los resultados se muestran en el conjunto de gráficas de la Figura 7.6. Como se puede observar en las gráficas 7.6(a), 7.6(b) y 7.6(c), todos los clasificadores alcanzan el mismo máximo de 89,52 % de precisión cuando $m = 8$. Las curvas ROC representadas en la Figura 7.6(d) que revelan que SVM y Bayes cuadrático

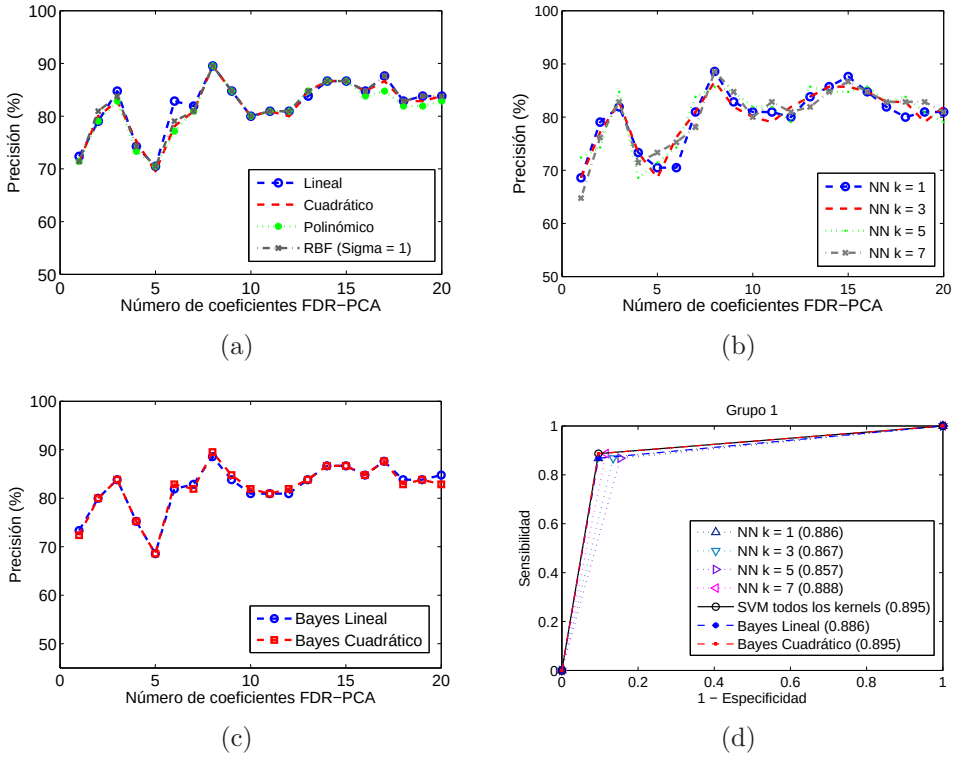


Figura 7.6: Grupo 1 de ADNI: Resultados de precisión alcanzada por clasificadores (a) SVM, (b) NN y (c) Bayesianos. Las curvas ROC correspondientes se representan en (d).

son las mejores opciones como clasificadores para este grupo.

Obsérvese que la forma de las curvas de precisión de la Figura 7.6 (a)-(c) es muy similar para los diferentes clasificadores. Esto es debido al hecho de usar una única característica final para la clasificación, lo que obliga a los clasificadores a determinar el umbral de separación de forma similar a pesar de los diferentes criterios que cada uno impone en el diseño de su regla de clasificación. En todos los casos sin embargo se considera $m = 8$ el número óptimo de coeficientes PCA para llevar a cabo la transformación LDA y la posterior clasificación. Esta unanimidad implica robustez de las características extraídas y empleadas en el entrenamiento para la separación de las clases contenidas en este grupo.

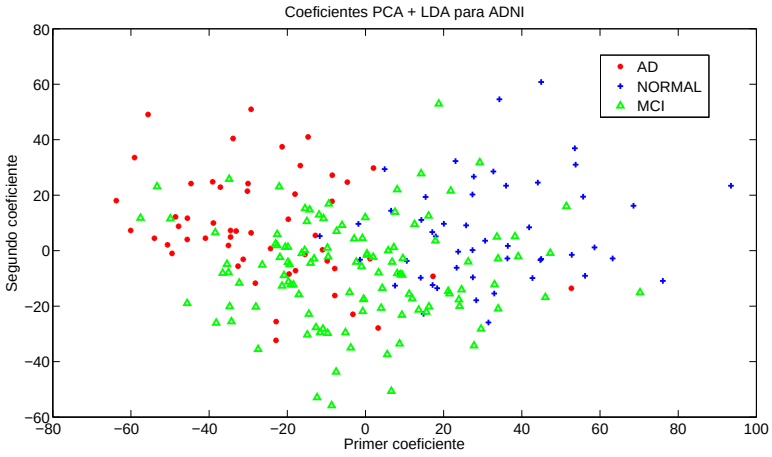
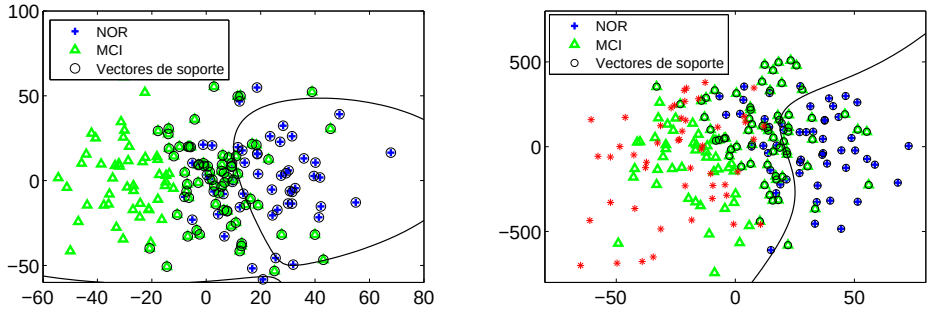


Figura 7.7: Descripción de las clases de la base de datos ADNI en términos de coeficientes PCA+LDA. Las imágenes originales fueron proyectadas en $m = 30$ componentes principales y seguidamente en $l = 2$ ejes de proyección LDA.

Grupo 2

La tarea de clasificación más complicada llevada a cabo sobre la base de datos ADNI es aquella que pretende diferenciar los pacientes etiquetados como NORMAL y los MCI. Lograr una tasa de acierto alta en la clasificación de este grupo significaría la alta precisión a la hora de detectar una anomalía de tipo funcional en el paciente que podría evolucionar hacia el estado AD. Los patrones cerebrales etiquetados como MCI son complejos y con una variabilidad muy alta, que evolucionan en tiempo cuando la enfermedad avanza [Minoshima et al., 1997; Drzezga et al., 2003]. En la Figura 7.7 queda reflejado el amplio rango en el que se expanden las características extraídas de pacientes MCI, quedando solapadas con aquellas extraídas tanto de controles como de pacientes AD, y por tanto dificultando la tarea de clasificación.

En el caso de este grupo y debido a la dispersión de las muestras de la clase MCI, el uso del criterio FDR para la selección de coeficientes PCA no es apropiado. El criterio FDR define un orden de selección muy diferente del criterio de varianza para este grupo, siendo a su vez pequeñas las diferencias entre los coeficientes FDR calculados. La reordenación mediante FDR asigna prioridad a componentes principales de bajo contenido en varianza. Por ejemplo, el criterio FDR sitúa a la primera componente principal según el criterio de varianza en el puesto 100. Según el criterio FDR por tanto es-



(a) (b)

Figura 7.8: Grupo 2 de ADNI: Línea de clasificación definida por un clasificador SVM con kernel polinómico. En (a) se emplean las características extraídas de pacientes NORMAL y MCI y en (b) se incluyen los pacientes AD en la etapa de entrenamiento como muestras adosadas a la clase MCI.

ta componente queda fuera de la clasificación a pesar de ser la que mayor información de variabilidad de clases contiene.

Cuando se llevan a cabo las tareas de clasificación binaria sobre este grupo, la tasa máxima de precisión alcanzada no supera el 74,1 % mediante PCA y LDA como técnicas de extracción de características (es decir, una sóla característica final). A pesar de la imprecisión que caracteriza a la clase MCI, ésta puede considerarse una etapa anterior a la EA [Minoshima et al., 1997; Silverman et al., 2001] aunque no existe certeza de que las anomalías detectadas evolucionen hacia la EA, sino que podrían evolucionar hacia otras demencias o incluso tras un seguimiento, ser finalmente etiquetados como pacientes normales. Con el fin de desarrollar un sistema CAD capaz de diferenciar este tipo de patrones anormales o MIC de aquellos normales, se puede tomar ventaja del hecho de contar con un conjunto de imágenes en la misma base de datos etiquetada como AD. Como se justificó en la sección 7.3, LDA puede encontrar un máximo de $l = c - 1$ dicriminantes, donde c es el número de clases, de modo que si se consideran los pacientes AD como una clase diferente se podrían obtener hasta $l = 2$ características finales para el entrenamiento y test, como si se tratase de un problema multiclase con $c = 3$ clases. Este "truco" permitiría el empleo de dos características en lugar de una, obteniendo así más información de clase. Una vez que las $l = 2$ características son extraídas, a los pacientes AD y MCI se les asigna la misma etiqueta como si se tratase de una sola clase. El conjunto AD + MCI se trata por tanto como una única clase durante el entrenamiento del clasificador. De

este modo, el clasificador diseñará un hiperplano de decisión que refuerza la idea de la clase MCI como patrón anormal al equipararlo con la clase AD, desplazando el hiperplano hacia el lado AD y alejándolo por tanto de la clase NORMAL. Sin embargo no estamos interesados en clasificar los pacientes AD, por lo que en la fase de test sólo se clasifican los pacientes de las clases NORMAL y MCI. La Figura 7.8 muestra las líneas de decisión definidas por un clasificador SVM con kernel polinómico cuando se considera el caso binario convencional de separación de clases y cuando se emplea el truco multiclase recién expuesto. El uso de esta técnica mejora notablemente los resultados de clasificación mediante un incremento de los valores de sensibilidad. Los valores de especificidad sin embargo demuestran un comportamiento pobre por parte de los clasificadores a la hora de distinguir los controles. Sin embargo, los valores de sensibilidad y especificidad son dependientes de la prevalencia, que en el caso del grupo 2 no es balanceada puesto que el número de pacientes de la clase MCI duplica al número de controles disponibles. Para evitar una interpretación errónea de los resultados, se calculan también los valores de PL y NL. La Tabla 13.2 muestra todos los resultados obtenidos en los experimentos y compara los resultados de una clasificación binaria convencional y la clasificación basada en la multiclase propuesta para el grupo 2.

La interpretación de los datos de precisión alcanzados para la base de datos ADNI debe tomarse con cautela. No es posible una comparación cuantitativa directa con los resultados obtenidos para las distintas bases de datos debido a las diferencias en el método de etiquetado. Las etiquetas de la base de datos ADNI son asignadas en función de los resultados que cada paciente proporciona en los tests previos sin incluir en ellas la valiosa información contenida en las imágenes. Las imágenes PET *Cartuja* por ejemplo se etiquetan tras la exploración de las imágenes, dándole mayor fiabilidad al etiquetado. Por tanto la reproducción exacta al 100 % de las etiquetas en ADNI podría no ser adecuado llevando al sistema CAD a un aprendizaje incorrecto de los estados de los pacientes. Una tasa de acierto baja sobre el etiquetado total de la base de datos podría malinterpretarse si realmente el sistema acierta sobre las etiquetas verdaderas.

Grupo 2

	m	Bayes Lineal	Bayes Cuadrático
$l = 1$	24	68.07 % (68.42/67.31) % 2.093/0.469	66.27 % (64.91/69.23) % 2.109/0.507
$l = 2$	35	80.11 % (85.08/69.23) % 2.765/0.215	78.31 % (83.33/63.46) % 2.280/0.263

	m	SVM Lineal	SVM Cuadrático	SVM Polinómico	SVM RBF
$l = 1$	18	74.1 % (86.84/46.15) % 1.613/0.285	74.1 % (86.84/46.15) % 1.613/0.285	73.49 % (85.96/46.15) % 1.596/0.304	73.49 % (85.96/46.15) % 1.596/0.304
$l = 2$	30	81.33 % (97.37/46.15) % 1.808/0.057	77.71 % (91.23/46.15) % 1.694/0.190	77.71 % (91.23/46.15) % 1.694/0.190	77.71 % (91.22/48.08) % 1.757/0.183

	m	NN k = 1	NN k = 3	NN k = 5	NN k = 7
$l = 1$	33	71.08 % (82.46/46.15) % 1.531/0.380	72.29 % (83.33/48.08) % 1.605/0.347	71.08 % (82.46/46.15) % 1.531/0.380	70.48 % (81.58/46.15) % 1.515/0.399
$l = 2$	40	79.52 % (94.74/46.15) % 1.759/0.114	78.31 % (93.86/44.23) % 1.683/0.139	79.52 % (93.86/48.08) % 1.801/0.128	79.52 % (92.10/46.15) % 1.710/0.171

Tabla 7.1: Grupo 2 de ADNI: Datos de precisión, (sensibilidad/especificidad) y PL/NL obtenidos en los experimentos de clasificación. Comparativa de los métodos aplicados con $l = 1$ y $l = 2$ características empleadas.

CAPÍTULO 8

Métodos Kernel

En los capítulos anteriores, las características PCA y LDA han demostrado contener una gran capacidad discriminativa a la hora de clasificar imágenes neurológicas. La representación de las imágenes en estos subespacios se basa en estadísticos de segundo orden del conjunto de imágenes, pero no tratan otras dependencias estadísticas de orden superior como las relaciones entre tres o más voxels. Mientras que con los *eigenbrains* se pretende encontrar las direcciones de proyección basadas en la correlación de segundo orden de las muestras, Kernel PCA y Kernel LDA proporcionan una generalización que tiene en cuenta correlaciones de orden mayor. El uso del "truco" kernel proporciona un modo eficiente para extraer de las muestras características no lineales, dando lugar por tanto una representación más rica de las mismas, o visto de otro modo, se trata de una proyección de las muestras desde el espacio de entrada a un espacio de dimensión mayor. En este capítulo se investiga la eficacia de las métodos Kernel PCA y Kernel LDA como técnicas de extracción de características para una posterior clasificación.

8.1. Funciones kernels

En los últimos años, los métodos de aprendizaje basados en kernel como por ejemplo son los SVM [Vapnik, 1995], Kernel PCA (KPCA) y Kernel Discriminant Analysis (KDA) han suscitado un interés considerable en el campo de reconocimiento de patrones y máquinas de aprendizaje [Müller et al., 2001]. KPCA fue desarrollado originalmente por Schölkopf et al. [1998], mientras que KDA fue propuesto inicialmente por Mika et al. [1999b,a], formulado para dos clases. Debido a su habilidad para extraer las características no lineales más discriminantes [Mika et al., 1999b,a; Yang, 2002; Cawley and Talbot, 2003; Lawrence and Schölkopf, 2001], KDA ha demostrado ser una técnica muy efectiva en muchas aplicaciones del mundo real [Yang, 2002; López et al., 2009c,b].

La idea básica de los métodos kernel es preprocesar los datos mediante algún mapeo no lineal Φ y después aplicar los mismos algoritmos lineales descritos por ejemplo, en los Capítulos 6 y 7, pero en el espacio de características Φ. La idea es encontrar un Φ apropiado para el cual una decisión lineal en el espacio de Φ sea suficiente para la separación de las clases. (Ver la Figura 8.1).

En cuanto a la clasificación basada en kernels, matemáticamente se puede considerar que una medida de similitud entre un dato $\mathbf{x}$ con etiqueta conocida y un nuevo patrón $\mathbf{x}'$ viene dada por la función:

$$\begin{array}{rccc} k: & \mathcal{H} \times \mathcal{H} & \longrightarrow & \mathbb{R} \\ & (\mathbf{x}, \mathbf{x}') & \longmapsto & k(\mathbf{x}, \mathbf{x}') \end{array} \tag{8.1}$$

es decir, una función que, dadas dos muestras del espacio de características $\mathbf{x}$ y $\mathbf{x}'$, devuelve un número real que caracteriza su similitud. Es natural asumir que no existe diferencia al comparar la similitud entre $\mathbf{x}$ y $\mathbf{x}'$ que entre $\mathbf{x}'$ y $\mathbf{x}$, por lo que asumiremos que esta función es simétrica, es decir, $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$, para todo $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$.

Un ejemplo sencillo de una función de este tipo es el producto escalar en un espacio vectorial, que lleva asociado una noción de distancia. La medida de la similitud en un espacio con un producto interno definido se traduce a una medida de distancia, haciendo que se pueda establecer si dos objetos son similares a través de su cercanía.

Estas ideas se pueden extrapolar a espacios más generales que $\mathbb{R}^m$, en los que también es posible definir un producto interno. Estos espacios, denom-

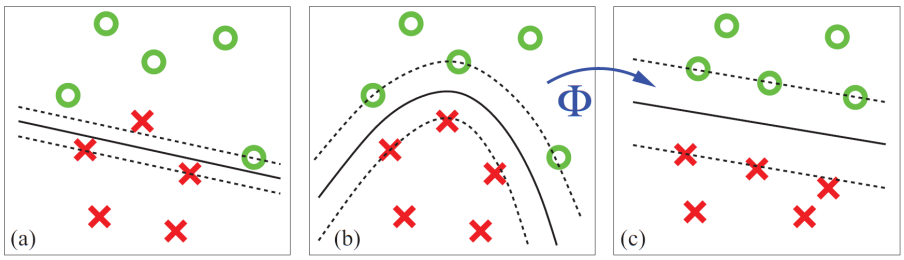


Figura 8.1: Ejemplo de clasificación basada en kernel. Tres vistas diferentes del mismo problema de separación. En (a) no es posible una separación lineal de los puntos en el espacio de entrada sin errores. En (b) la separación es posible mediante superficies no lineales en el espacio de entrada. (c) Estas superficies no lineales corresponden con superficies lineales en el espacio de características. Los puntos se mapean del espacio de entrada al espacio de características por medio de la función Φ.

inados espacios de Hilbert, son análogos al espacio Euclídeo en tanto que se puede definir en ellos los conceptos de longitud, distancia y ángulo, pero poseen características más generales. Son la generalización del espacio Euclídeo, reproduciendo el álgebra vectorial de manera abstracta e incluyendo una noción de completitud.

Es posible que la noción de similitud se defina más apropiadamente en otro espacio V diferente al espacio de características original $\mathcal{H}$. Supondremos que este espacio V es un espacio de Hilbert, con su correspondiente producto interno. Primero, necesitaremos una representación vectorial de los vectores de características en este espacio, que se construirá a través del mapeo:

$$\begin{array}{rccc} \mathbf{\Phi}: & \mathcal{H} & \longrightarrow & V \\ & \mathbf{x} & \longmapsto & \mathbf{v} = \mathbf{\Phi}(\mathbf{x}) \end{array} \tag{8.2}$$

Este mapeo definirá un nuevo espacio de características V y sus correspondientes vectores de características $\mathbf{v} \in V$, de manera que el producto escalar quedará mapeado a:

$$k(\mathbf{v}, \mathbf{v}') = \mathbf{v} \cdot \mathbf{v}' = \mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{x}) = k(\mathbf{\Phi}(\mathbf{x}), \mathbf{\Phi}(\mathbf{x})) \tag{8.3}$$

de modo que un kernel es sencillamente una función que puede ser representada como un producto interno en algún espacio de Hilbert.

Esta transformación permitirá seguir trabajando con las nociones habi-

tuales de distancia, longitud y ángulo, y las herramientas sencillas empleadas para trabajar con ellas (álgebra lineal, geometría analítica,...) pero abriendo la posibilidad de utilizar un amplio rango de diferentes medidas de similitud, gracias a la libertad a la hora de elegir el mapeo $\mathbf{\Phi}$. Uno puede preguntarse sobre las condiciones generales para la existencia de tal mapeo, cuya respuesta viene dado por el Teorema de Mercer:

Teorema. *Teorema de Mercer. Sea* $\mathbf{x} \in \mathbb{R}^l$ *y* Φ *una función de mapeo*

$$\mathbf{x} \rightarrow \Phi(\mathbf{x}) \in H$$

donde H es un espacio Euclídeo. Entonces, la operación del producto interno tiene una representación equivalente

$$\sum_r \phi_r(\mathbf{x})\phi_r(\mathbf{z}) = K(\mathbf{x},\mathbf{z})$$

donde $\phi_r(\mathbf{x})$ *es la r-ésima componente del mapeo* $\Phi(\mathbf{x})$ *de* $\mathbf{x}$, *y* $K(\mathbf{x},\mathbf{z})$ *es una función simétrica que satisface la siguiente condición*

$$\int K(\mathbf{x},\mathbf{z})g(\mathbf{x})g(\mathbf{z})d\mathbf{x}d\mathbf{z} \geq 0 \tag{8.4}$$

para cualquier $g(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^l$ *tal que*

$$\int g(\mathbf{x})^2 d\mathbf{x} < +\infty \tag{8.5}$$

El contrario también es cierto, es decir, para cualquier $K(\mathbf{x},\mathbf{z})$ que satisfaga 8.4 y 8.5 existe un espacio en el cual se define un producto interno. Tales funciones se conocen como *Kernels*. Gracias al teorema de Mercer podemos estar seguros de se puede establecer un mapeo a un nuevo espacio que permita definir una noción de similitud. Normalmente este nuevo espacio implicará un aumento de la dimensión del espacio de características, para proporcionar un marco en el que las clases se separen más fácilmente. Lo que el teorema de Mercer no revela sin embargo es cómo encontrar este espacio. Es decir, no tenemos una herramienta general para construir el mapeo $\Phi(\cdot)$ una vez que conocemos el producto interno del correspondiente espacio. Además, tampoco

podemos saber la dimensionalidad del espacio, la cual podría ser incluso infinita.

En la práctica, ha resultado que los mapeos no lineales producen estructuras suficientemente interesantes para resolver problemas complejos, siendo algunos ejemplos típicos de kernels usados en aplicaciones de reconocimiento de patrones:

- Polinómicos:

$$K(\mathbf{x}, \mathbf{x}') = [\gamma(\mathbf{x} \cdot \mathbf{x}') + c]^d. \tag{8.6}$$

- Funciones de base radial (RBF):

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2). \tag{8.7}$$

- Tangente hiperbólica:

$$K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma(\mathbf{x} \cdot \mathbf{x}') + c). \tag{8.8}$$

  Para valores apropiados de γ y c de modo que las condiciones de Mercer se cumplan.

8.2. Kernel PCA

PCA se diseñó para modelar variabilidades lineales existentes en datos de alta dimensión. Sin embargo, en muchas ocasiones estos datos tienen naturaleza no lineal. En estos casos de datos de alta dimensión pertenecientes a un subespacio no lineal PCA no puede modelar las variabilidades de los datos de forma correcta. En KPCA, mediante el uso de kernels las componentes principales se pueden calcular de forma eficiente en espacios de características de alta dimensión, que se relacionan con el espacio de entrada mediante algún tipo de mapeo no lineal. Por tanto, KPCA encuentra las componentes principales que están relacionadas de forma no lineal con el espacio de entrada llevando a cabo PCA en el espacio producido por el mapeo no lineal, donde se espera que la estructura latente de baja dimensión sea más fácil de descubrir (ver Figura 8.2)

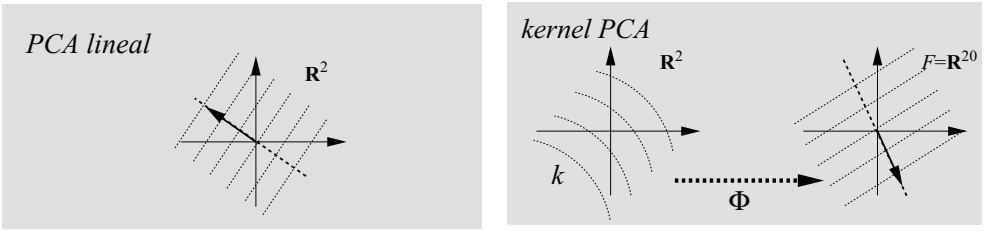


Figura 8.2: Idea básica de kernel PCA. En algún espacio de características de alta dimensión $\mathcal{F}$ (derecha) se aplica PCA lineal de igual modo que en el espacio de entrada (izquierda). Puesto que $\mathcal{F}$ se relaciona con el espacio de entrada de forma no lineal, (por medio de Φ) las líneas de contorno constantes de las proyecciones sobre los autovectores principales (dibujadas como flechas) se convierten en no lineales en el espacio de entrada.

Consideremos el mapeo $\mathbf{\Phi}$ del espacio original $\mathbb{R}$ a un nuevo espacio $\mathcal{F}$:

$$\begin{array}{rccl} \mathbf{\Phi}: & \mathbb{R} & \longrightarrow & \mathcal{F} \\ & \mathbf{x} & \longmapsto & \mathbf{y} = \mathbf{\Phi}(\mathbf{x}) \end{array} \tag{8.9}$$

Nótese que $\mathcal{F}$ podría tener una dimensión f arbitrariamente grande, incluso infinita. Se asume que se trata con datos centrados (ver Schölkopf et al. [1996]), es decir

$$\sum_{n=1}^{N} \mathbf{\Phi}(\mathbf{x}_n) = 0 \tag{8.10}$$

Usando la matriz de covarianza en $\mathcal{F}$

$$\mathbf{\Sigma}^{\Phi} = \frac{1}{N} \sum_{i=1}^{N} \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_i)^T \tag{8.11}$$

el problema de autovalores correspondiente es

$$\mathbf{\Sigma}^{\Phi} \mathbf{u}^{\Phi} = \lambda \mathbf{v}^{\Phi} \tag{8.12}$$

Todas las soluciones $\mathbf{u}^{\Phi}$ con $\lambda \neq 0$ se encuentran en el espacio expandido por $\mathbf{\Phi}(\mathbf{x}_1), \mathbf{\Phi}(\mathbf{x}_2), ..., \mathbf{\Phi}(\mathbf{x}_N)$, y por tanto existen coeficientes α_i tales que

$$\mathbf{u}^{\Phi} = \sum_{i=1}^{N} \alpha_i \mathbf{\Phi}(\mathbf{x}_i) \tag{8.13}$$

Nótese que si $\Phi(\mathbf{x})$ es f-dimensional y la dimensión del espacio de características f es grande, entonces resolver el problema de autovalores se hace poco práctico. Para reducir la dependencia sobre f, primero se asume que se tiene una matriz kernel $K(\cdot, \cdot)$ que permite calcular $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j)$. Dada tal función, se puede calcular la matriz $K = \Phi(\mathbf{x})^T \cdot \Phi(\mathbf{x})$ eficientemente sin calcular explícitamente $\Phi(\mathbf{x})$. De forma crucial, K tiene dimensión $N \times N$ y no depende de f. Por tanto se puede calcular en un tiempo de ejecución que depende sólo de N.

Para el propósito de extracción de componentes principales necesitamos calcular las proyecciones sobre los autovectores $\mathbf{u}^{\Phi}$ en $\mathcal{F}$. Sea $\mathbf{x}$ un punto de test con la imagen $\Phi(\mathbf{x})$ en $\mathcal{F}$. Entonces, la proyección de $\Phi(\mathbf{x})$ sobre los autovectores $\mathbf{u}^{\Phi}$ son las componentes principales no lineales correspondientes a $\mathcal{F}$:

$$\mathbf{u}^{\Phi} \cdot \mathbf{\Phi}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i (\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x})) = \sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}) \tag{8.14}$$

De nuevo se extraen las $m, (1 \leq m \leq N)$ componentes principales no lineales (es decir, autovectores $\mathbf{u}_i^{\Phi}, i = 1, ..., m$) usando la función kernel sin necesidad de realizar la operación que explícitamente proyecta las muestras en un espacio de alta dimensión.

Para obtener formas no lineales de PCA simplemente se elige un kernel no lineal (ver ecuaciones 8.6,8.7,8.8). Nótese que la técnica clásica de PCA es un caso especial de KPCA con orden uno ($d = 1$) para el kernel polinómico (8.6). Dicho de otro modo, KPCA es una generalización de la clásica transformación PCA puesto que los diferentes kernels pueden utilizarse para diferentes proyecciones no lineales dando lugar a espacios de características de dimensión arbitrariamente alta, incluso infinita.

8.3. Kernel LDA

Para la mayoría de los casos de aplicaciones reales, un discriminante lineal no es suficientemente complejo para separar los datos. Para aumentar la eficiencia del discriminante se podría o bien hacer uso de modelos de distribuciones más sofisticadas que las bayesianas o bien buscar direcciones no lineales (o ambas soluciones). Sin embargo, la asunción de distribuciones generales podría causar problemas a la hora de la estimación de los parámetros. Por tanto la solución se restringe a la búsqueda de direcciones no lineales mediante el mapeo no lineal de los datos a un espacio de características $\mathcal{F}$ y el subsiguiente cálculo de los discriminantes de Fisher en este espacio, que darán lugar a discriminantes no lineales en el espacio de entrada.

A continuación se desarrolla la formulación de la transformada LDA (expuesta en el Capítulo 7) en este caso basada en kernels. Sea $\mathbf{\Phi}$ una función de mapeo no lineal a un espacio $\mathcal{F}$. Para encontrar los discriminantes lineales en $\mathcal{F}$ se necesita optimizar

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b^\Phi \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^\Phi \mathbf{w}} \tag{8.15}$$

donde ahora $\mathbf{w} \in \mathcal{F}$ y $\mathbf{S}_b^\Phi$ y $\mathbf{S}_w^\Phi$ son las correspondientes matrices de dispersión inter-clase e intra-clase en $\mathcal{F}$, es decir

$$\mathbf{S}_b^\Phi = (\boldsymbol{\mu}_1^\Phi - \boldsymbol{\mu}_2^\Phi)(\boldsymbol{\mu}_1^\Phi - \boldsymbol{\mu}_2^\Phi)^T \tag{8.16}$$

$$\mathbf{S}_w^\Phi = \sum_{i=1,2} \sum_{\mathbf{x} \in \omega_i} (\Phi(\mathbf{x}) - \boldsymbol{\mu}_i^\Phi)(\Phi(\mathbf{x}) - \boldsymbol{\mu}_i^\Phi)^T \tag{8.17}$$

con $\boldsymbol{\mu}_i^\Phi = \frac{1}{N_i} \sum_{j=1}^{N_i} \Phi(\mathbf{x}_j^i)$, siendo N_i el número de muestras en la clase ω_i, $i = 1, 2$. Si $\mathcal{F}$ es un espacio de muy alta dimensión, claramente sería imposible resolver este problema directamente. Para superar esta limitación, se usa el mismo truco que en KPCA [Schölkopf et al., 1998]. En lugar de mapear los datos explícitamente, se busca una formulación del algoritmo que use sólo productos internos $(\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j))$ de los patrones de entrenamiento. Puesto que se puede calcular dichos productos internos de manera eficiente podemos resolver el problema original sin la necesidad de mapear explícitamente los datos a $\mathcal{F}$. Esto se consigue usando los kernels de Mercer (ver Sección 8.1). Para encontrar los discriminantes de Fisher en el espacio de características

$\mathcal{F}$, primero se necesita una formulación de la ecuación 14.10 en términos únicamente de productos internos de los patrones de entrada que después se sustituirán por alguna función kernel. De la teoría de kernels sabemos que cualquier solución $\mathbf{w} \in \mathcal{F}$ debe pertenecer al espacio que expanden todas las muestras de $\mathcal{F}$. Por tanto, se puede encontrar una expansión de $\mathbf{w}$ de la forma

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{\Phi}(\mathbf{x}_i) \tag{8.18}$$

Haciendo uso de 14.13 y de la definición de $\boldsymbol{\mu}_i^{\Phi}$ se puede escribir

$$\mathbf{w}^T \boldsymbol{\mu}_i^{\Phi} = \frac{1}{N_i} \sum_{j=1}^{c} \sum_{k=1}^{N_i} \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) = \boldsymbol{\alpha}^T \mathbf{M}_i \tag{8.19}$$

donde se ha definido $\mathbf{M}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} k(\mathbf{x}, \mathbf{x}_k^i)$ y se han sustituido los productos internos por funciones kernel. Consideremos ahora el numerador de la ecuación 14.10. Usando la definición de $\mathbf{S}_b^{\Phi}$ y 14.18 puede reescribirse

$$\mathbf{w}^T \mathbf{S}_b^{\Phi} \mathbf{w} = \boldsymbol{\alpha}^T M \boldsymbol{\alpha} \tag{8.20}$$

donde $M := (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T$. Considerando ahora el denominador, haciendo uso de 14.13 y de la defínción de $\mathbf{S}_w^{\Phi}$, y de forma análoga que en 14.15 se tiene

$$\mathbf{w}^T \mathbf{S}_w^{\Phi} \mathbf{w} = \boldsymbol{\alpha}^T N \boldsymbol{\alpha} \tag{8.21}$$

donde se define $N := \sum_{i=1,2} K_j (\mathbf{I} - \mathbf{1}_{l_j}) K_j^T$, K_j es una matriz de $l \times l_j$ con $(K_j)_{nm} := k(\mathbf{x}_n, \mathbf{x}_n^j)$ (es decir, la matriz kernel para la clase j), $\mathbf{I}$ es la matriz identidad y $\mathbf{1}_{l_j}$ una matriz con todos sus elementos $1/l_j$.

Combinando 14.15 y 14.16 podemos encontrar los discriminantes de Fisher lineales en $\mathcal{F}$ maximizando

$$J(\alpha) = \frac{\boldsymbol{\alpha}^T M \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T N \boldsymbol{\alpha}} \tag{8.22}$$

Este problema se puede resolver (de forma análoga que en el espacio de entrada) encontrando los autovectores de $N^{-1}M$. Esta es la aproximación KDA. La proyección de una nueva muestra $\mathbf{x}$ en $\mathbf{w}$ viene dada por

$$\mathbf{w} \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{8.23}$$

KDA resulta ser más efectivo que LDA en muchas aplicaciones. Sin embargo, KDA es menos transparente que LDA debido a la complejidad de su formulación que de algún modo cubre las características intuitivas de LDA. En Yang et al. [2004] se analiza en profundidad el algoritmo KDA, revelando que esta transformada es equivalente a KPCA + LDA.

8.4. Experimentos

La búsqueda de las componentes principales en un espacio de características transformado mediante funciones kernel es un proceso computacionalmente más complicado y menos eficiente que la técnica clásica PCA. Esto motiva la reducción del espacio de entrada mediante las técnicas de submuestreo así como el previo descarte de voxels no considerados como voxels de interés mediante la aplicación de una máscara, como se explica en la Sección 3.3 y 3.3 respectivamente. Así, todas las imágenes utilizadas para la extracción de componentes KDA han sido previamente reducidas y *filtradas* por la máscara utilizando un factor de submuestreo de $2 \times 2 \times 2$ y umbral de selección de la máscara de $t = 0{,}5$.

KDA fue evaluado sobre las bases de datos SPECT y PET *Cartuja* obteniendo en algunos casos mejoras sobre la base de datos SPECT e igualando el 100 % de precisión obtenido para PET. Los kernels aplicados para la extracción de los coeficientes KDA que mejor resultados proporcionaron son el kernel cuadrático y polinómico ($d = 2$ y $d = 3$ en la ecuación 8.6 respectivamente). La extracción de las componentes principales mediante kernel RBF produjo bajas tasas de precisión, por lo que no se muestran los resultados.

En la fase de clasificación, Bayes y SVM fueron los clasificadores más eficientes, superando a NN. Sin embargo sólo SVM consiguió mejoras sobre la técnica clásica PCA+LDA (con reordenación mediante FDR), por lo que en la próxima sección sólo se muestran los resultados obtenidos con SVM.

8.5. Resultados

La combinación de KDA empleando una única característica final ($l = 1$) con clasificadores SVM produjo resultados de precisión mostrados en la Tabla 8.1.

	Proyección PCA	m	SVM Lineal	SVM Cuadrático	SVM Polinómico	SVM RBF
SPECT	Polin. d = 2	3	87.91 %	89.01 %	87.91 %	89.01 %
	Polin. d = 3	3	90.11 %	89.01 %	86.81 %	89.01 %
PET	Polin. d = 2	34	98.33 %	98.33 %	100 %	98.33 %
	Polin. d = 3	34	98.33 %	98.33 %	100 %	98.33 %

Tabla 8.1: Precisión obtenida mediante KDA ($l = 1$) y SVM para SPECT y PET.

A priori KDA funciona correctamente para las imágenes PET, reproduciendo los mejores resultados obtenidos con las técnicas previas, mientras que para las SPECT no se produce un mejora con respecto a PCA+LDA lineal.

Llegados a este punto, se propone una discusión sobre el número final de características l empleadas en la tarea de clasificación. Según la teoría de LDA explicada en el Capítulo 7, para un problema de clasificación binaria esta transformación encuentra el eje de proyección de máxima separación entre ambas clases. Para un problema multiclase con c clases, LDA encuentra $l = c-1$ ejes de proyección [Fukunaga, 1990]. Sin embargo cabe cuestionarse si este número de características óptimo es dependiente únicamente del número de clases e independiente de la distribución de las mismas. La calidad de las características seleccionadas depende sólo de la precisión con la que el criterio, $|\mathbf{S}_w^{-1}\mathbf{S}_b|$, mide la separabilidad de las clases. Sin embargo, para clases con distribuciones multimodales con medias comunes por ejemplo, $l = c - 1$ podría no ser el número óptimos de características necesario para conseguir la mejor separación [Fukunaga, 1990].

En nuestro caso el criterio de separación viene determinado por la regla que establecen los distintos clasificadores empleados, y en concreto los clasificadores SVM para las características extraídas con KDA. Debido a diferentes posibilidades de definir superficies de decisión gracias al uso de SVM con kernels, encontramos que para nuestros datos y para superficies de separación no lineales, en algunas ocasiones la proyección de las características KPCA sobre un número de ejes $l > 1$ produce ciertas mejoras en los resultados de

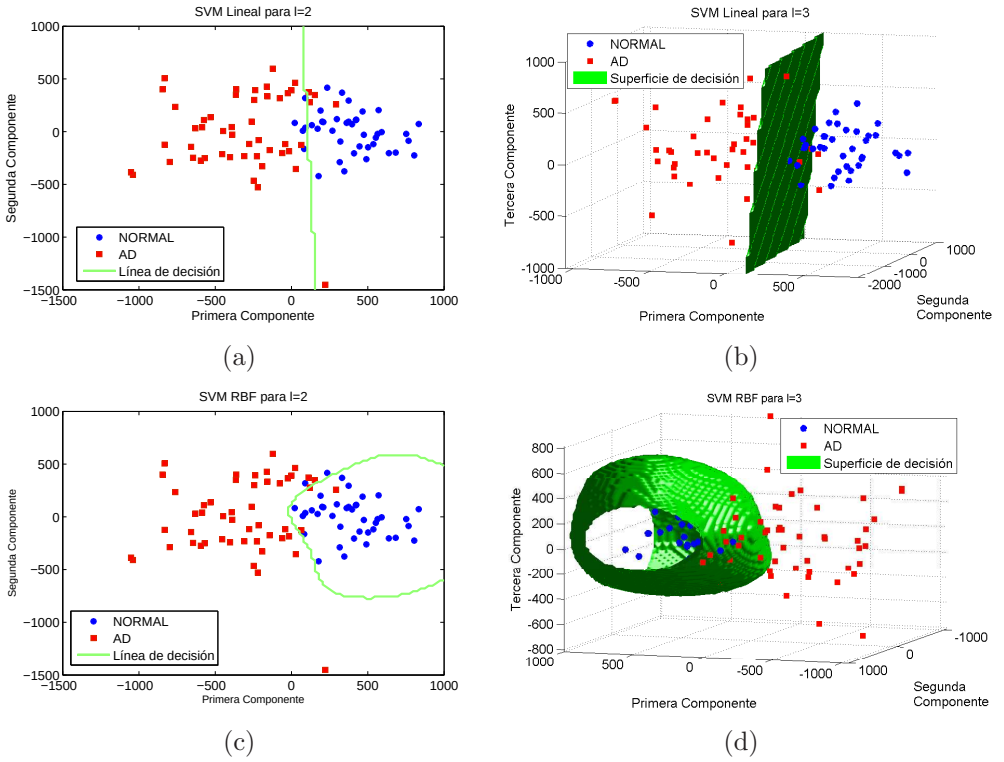


Figura 8.3: Líneas y superficies de decisión diseñadas por clasificadores SVM lineal (arriba) y SVM con kernel RBF (abajo) cuando se emplean $l = 2$ (izquierda) y $l = 3$ (derecha) características KDA. La primera característica es la más discriminante en todos los casos, sin embargo para clasificadores no lineales la segunda y tercera componentes KDA pueden contribuir a una mejora del ajuste del hiperplano de decisión.

clasificación, en concreto cuando los datos presentan dificultades para ser linealmente separados. Esto ocurre en nuestro caso para las imágenes SPECT. La proyección de las características KDA para esta base de datos sobre los $l = 2$, $l = 3$, etc, ejes de proyección no aporta mayor información para un clasificador lineal, pues dichos ejes son autovectores (ecuación 14.10) correspondientes a autovalores prácticamente nulos. Sin embargo, para superficies no lineales, y en especial para el kernel RBF, encontramos que la proyección sobre $l = 3$ ejes de proyección encontrados mediante la técnica KDA mejora los resultados.

La Figura 8.5 ilustra las superficies de decisión diseñadas por el clasificador SVM con kernel RBF para las características KDA extraídas a partir de las imágenes SPECT, usando $l = 2$ (Figuras 8.3(a) y 8.3(c)) y $l = 3$ (Fig-

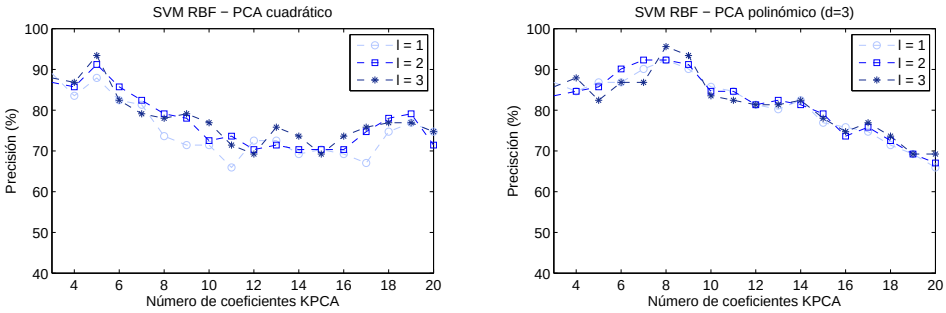


Figura 8.4: Las curvas de precisión muestran una ligera mejora cuando se emplean $l = 2$ y $l = 3$ características KDA en combinación con un clasificador SVM con kernel RBF. Los resultados se muestran en función del número de coeficientes KPCA proyectados sobre los ejes LDA.

uras 8.3(b) y 8.3(d)). En todos los casos, la primera característica KDA es sin duda la más determinante agrupando las clases en dos grupos diferenciados mayormente por el valor de esta coordenada. En el caso de un clasificador lineal, el umbral estará situado en este valor. Sin embargo, para el clasificador BRF, debido a la forma no lineal de la superficie de decisión la segunda y tercera coordenadas no son totalmente despreciables, aportando en algún sentido información de clase para el clasificador. Para SVM con kernel RBF se obtiene una mejora cuando se emplea más de una dimensión de proyección LDA. La gráfica de la Figura 8.5 muestra las curvas de precisión para los diferentes kernels cuando se emplean $l = 2$ y $l = 3$. La mejora más significante se produce para el caso de SVM con kernel RBF, aumentando la precisión de 89,01 % a 93,41 % y de 92,31 % a 95,6 % cuando se emplean las características extraídas mediante PCA-cuadrático + LDA y PCA-polinómico + LDA respectivamente. Este valor de precisión de 95,6 % representa el valor más alto obtenido para la base de datos SPECT presentado en este trabajo. En la Tabla 14.1 se recogen los valores obtenidos por la técnica KDA, incluyendo kernel lineal (técnica clásica de PCA) a modo de comparativa.

	Proyección PCA	SVM Lineal	SVM Cuadrático	SVM Polinómico	SVM RBF
SPECT	Lineal	91.21 %	89.01 %	90.11 %	89.01 %
	Polin. d = 2	87.91 %	90.11 %	90.11 %	93.41 %
	Polin. d = 3	92.31 %	92.31 %	93.41 %	95.60 %
PET	Lineal	100 %	96.67 %	98.33 %	95 %
	Polin. d = 2	98.33 %	98.33 %	100 %	98.33 %
	Polin. d = 3	98.33 %	98.33 %	100 %	98.33 %

Tabla 8.2: Mejores resultados de precisión obtenidos empleando más de una característica en la proyección KDA y SVM para SPECT y PET.

CAPÍTULO 9

Discusión y Conclusiones

En este capítulo se recogen las conclusiones que se extraen de este trabajo presentando una discusión y comparativa sobre los diversos métodos propuestos. Para ello se resumen las ventajas e inconvenientes de cada esquema de clasificación así como la eficacia en términos de precisión que cada uno de ellos alcanza. Finalmente, tras la evaluación cualitativa de los métodos desarrollados se sugieren una serie de propuestas y nuevas líneas de investigación para la continuación y mejora de este trabajo.

9.1. Discusión y conclusiones

La principal aportación de este trabajo radica en la presentación de sistemas CAD completos e independientes que han demostrado cumplir satisfactoriamente los objetivos propuestos de detección y clasificación de imágenes. Todos los esquemas presentados consiguen una discriminación con éxito entre patrones normales y patrones de hipoperfusión e hipometabolismo aportando una valiosa herramienta de ayuda al diagnóstico. Los algoritmos desarrollados en esta Tesis han sido evaluados mediante una base de datos de 91 imágenes SPECT suministrada por el "Hospital Virgen de las Nieves", (Granada) y dos conjuntos de imágenes PET: uno de ellos consiste en 219 imágenes PET obtenidas de la base de datos ADNI y el segundo se trata de una pequeña base de datos de 60 pacientes procedentes de la clínica privada "PET Cartuja" (Sevilla).

En el campo de reconocimiento de patrones, los esquemas propuestos de extracción de características y clasificación aplicados a las imágenes funcionales resuelven con alto grado de satisfacción la categorización binaria de las mismas en las clases definidas como NORMAL y AD. A modo de resumen y comparativa, a continuación se exponen las ventajas y desventajas que presenta cada sistema CAD desarrollado.

La extracción de características basada en componentes presentada en el Capítulo 5 es un método directo que no requiere transformaciones del espacio de entrada a otros subespacios. Se trata de una versión mejorada del método VAF tomado como referencia que considera todos los voxels del volumen cerebral para entrenar un clasificador SVM. Este proceso de exploración del volumen cerebral es el que más se aproxima al proceso real en el cual un experto clínico examina las imágenes visualmente, siendo además el único método de los presentados que explota la propiedad de vecindad entre los voxels de regiones afectadas por la EA. La desventaja con respecto a VAF que presenta el método de componentes es la necesidad de realizar un barrido sobre el volumen cerebral para la localización de las ROIs que seleccionarán los voxels más discriminantes. Este proceso es computacionalmente costoso, pero sólo es necesario llevarlo a cabo una vez para producir el mapa de precisión. Una vez obtenido este mapa, para un nuevo paciente de test se accede directamente a los voxels que componen las ROIs sin necesidad de realizar el barrido completo. El mapa de precisión puede además ser actualizado "off-line" cuando nuevas muestras compongan la base de datos. El entrenamiento directo con subgrupos de voxels hace que este método sea sensible al posible ruido presente en las imágenes. Sin embargo este incon-

veniente se suaviza mediante el submuestreo inicial llevado a cabo sobre los volúmenes, que además supone una reducción del coste computacional. El entrenamiento de un conjunto de clasificadores presenta ventajas con respecto a una sola evaluación del cerebro. El agregado de clasificadores añade robustez al método VAF siendo el diagnóstico final del paciente el resultado de un estudio minucioso del volumen cerebral. Los resultados obtenidos mediante esta técnica de extracción de características son satisfactorios y aumentan en gran medida, en particular para las imágenes SPECT, la precisión obtenida (97,47 % para $T = 84\,\%$) con respecto a VAF (78,5 %). El umbral de precisión T que determina las componentes seleccionadas es un parámetro libre del sistema que puede ser variado por el usuario. Un umbral menor supone una mayor cantidad de votos recuperados siendo una opción más adecuada para imágenes con posible presencia de Alzheimer, mientras que un umbral alto puede confirmar la presencia de la EA en la imagen para pacientes más avanzados en la enfermedad.

En contraposición a los métodos de aplicación directa, los métodos basados en la transformación de los datos (PCA, LDA y kernel PCA) proyectan los datos iniciales sobre subespacios donde es posible determinar otras relaciones existentes entre las muestras de la base de datos. La información relevante correspondiente a las ROIs queda recogida en las componentes principales lineales extraídas de las bases de datos mediante PCA. Un estudio en profundidad de las componentes permite seleccionar mediante el criterio FDR aquellas que optimizan la clasificación binaria de los pacientes, aportando este criterio mejoras con respecto al criterio de mayor varianza. El número de componentes óptimo se determina *a posteriori*, una vez evaluado el sistema sobre la base de datos, y es dependiente de la misma. Esto obliga por tanto a realizar una previa descripción de la base de datos por medio de componentes principales al igual que para el método de componentes. Sin embargo, la extracción de componentes principales es computacionalmente más efectiva que el barrido por componentes y alcanza precisiones de 91,21 % y 100 % para SPECT y PET respectivamente. Las características proporcionadas por las transformaciones PCA y LDA demuestran ser robustas produciendo un comportamiento muy similar para todos los clasificadores evaluados.

En un punto intermedio entre el método ROIs y las transformaciones PCA+LDA en cuanto al coste computacional se refiere se encuentran las técnicas kernel para la extracción de características. La descripción de las bases de datos en términos de componentes principales no lineales resuelve de manera satisfactoria el problema de clasificación que nos ocupa, alcanzando precisiones de 95,60 % y 100 % para las bases de datos SPECT y PET *Cartuja*

respectivamente, y haciendo uso de tan sólo tres características finales que dan como resultado un aprendizaje rápido y robusto. Las técnicas basadas en transformaciones a subespacios son además menos sensibles al ruido existente en una imagen individual al no considerar directamente el valor de los voxels sino otras relaciones subyacentes de orden mayor entre ellos.

En lo que refiere a los métodos de clasificación evaluados, en general SVM y el clasificador de Bayes muestran un mejor rendimiento frente a NN para la separación de las características seleccionadas, tanto en tiempo de ejecución como en precisión alcanzada. La regla de Bayes resulta ser la más rápida en clasificación mientras que SVM alcanza las tasas de precisión más altas. La ventaja de SVM sobre el clasificador de Bayes radica en la capacidad de establecer hiperplanos de decisión no lineales mediante la combinación con kernels. Esto permite exprimir al máximo la información contenida en las características empleadas en clasificación, incluso cuando éstas no contienen información de clase, como se discute en la Sección 8.5.

La evaluación de los métodos toma especial interés cuando éstos son testados sobre la base de datos ADNI debido a la envergadura del proyecto que recolecta estas imágenes. Sin embargo, los niveles de precisión alcanzados no deben compararse cuantitativamente con aquellos que se obtienen al aplicar los métodos a los datos PET *Cartuja* por ejemplo. La razón principal es el modo en que los pacientes han sido etiquetados. En la base de datos ADNI, el etiquetado de la imagen se basa en los tests realizados al paciente, pero no en las imágenes en sí. La información que contienen las etiquetas ADNI son por tanto menos fiables que para PET *Cartuja*, donde la etiqueta incluye información de la imagen, la cual es muy valiosa no sólo para detectar la posible presencia de una enfermedad neurológica sino también para discernir entre diferentes causas de demencia.

9.2. Trabajo futuro

A continuación se proponen nuevas líneas de investigación que se derivan del trabajo aquí presentado y que podrían contribuir a un refinamiento de las técnicas presentadas así como al desarrollo de nuevos sistemas CAD. Las propuestas son las siguientes:

- Haciendo uso de técnicas desarrolladas de procesamiento de imagen, la búsqueda de ROIs podría llevarse a cabo mediante otros métodos directos que aplican segmentación de imagen como son el crecimiento

de regiones o clustering. Una vez que los voxels quedan agrupados en regiones, se llevaría a cabo un estudio de los subconjuntos para determinar las ROIs evitando así procesar aquellos voxels con bajo nivel discriminante.

- Estudio de las imágenes en otros dominios transformados como el dominio de la frecuencia o el dominio wavelet. Es posible extraer otro tipo de información relacionada con las frecuencias que componen una imagen mediante otras transformadas como la transformada discreta de Fourier (DFT) o la transformada discreta del coseno (DCT). Estas transformadas permiten realizar un estudio de los modos que la componen así como compresión de las mismas en los coeficientes resultantes. Estos coeficientes podrían contener información de clase relevante y por tanto ser útiles para clasificación.

- Clasificación multiclase. La naturaleza progresiva de una enfermedad neurológica como la EA da lugar a un amplio rango de patrones que se extienden desde la etapa temprana de la enfermedad cuando se producen las primeras anomalías hasta el estado avanzado. A menudo los pacientes afectados por la enfermedad son etiquetados como AD-1, AD-2 y AD-3 en función del avance de la enfermedad. En la base de datos ADNI por ejemplo se consideran los pacientes MCI como una clase "intermedia" entre patrones normales y AD. Esto sugiere la posibilidad de aplicar estas técnicas de diagnóstico a un problema multiclase que sea capaz de etiquetar con mayor precisión el estado del paciente. Para resolver problemas de aprendizaje multiclase, en Bennett [1999] se proponen las SVMs en combinación con los Métodos de Programación Matemática (MPM) que dan lugar a árboles de decisión donde cada decisión es una SVM. Una clasificación multiclase también sería de gran utilidad para fines no sólo de diagnóstico de una enfermedad neurológica en concreto sino también para la diferenciación entre dos o varias anomalías, como Parkinson, la demencia frontotemporal, etc.

Part III

Summary in English

Abstract

In this Thesis new statistical models for the detection of hipoperfusion and hipometabolism patterns in SPECT (Single Photon Emission Computed Tomography) and PET (Positron Emission Tomography) images are proposed for the aided diagnosis of neurological diseases, such as the Alzheimer's disease. After the description of the spatial registration and the intensity normalization of the images, three different schemes for feature extraction are presented. First, a component-based method is described, in which an exhaustive exploration of the images is performed in order to determine the regions of interest for classification. Secondly, the Fisher Discriminant Ratio is combined with linear transformations such as PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) as image reduction and feature extraction techniques, solving the well known *small sample size* problem successfully. Finally, transformations based on kernel functions are evaluated as a generalization of the previous linear methods, making possible the extraction of non-linear dependencies from the images. These features extraction techniques are combined with supervised classification methods based on Bayesian rules, Support Vector Machine and Neural Networks, yielding up to 95.6 % and 100 % accuracy values for SPECT and PET databases and 91.43 % for the recognized ADNI database, respectively. The proposed schemes outperfom the current existing methods for the early detection of the Alzheimer's disease.

CHAPTER 10

Neurological Image for Diagnosis

Emission-computed tomography (ECT) has been widely used in biomedical research and clinical practice during the last three decades. ECT differs from many other medical imaging modalities such as magnetic resonance imaging (MRI) in producing a mapping of physiological functions instead of imaging anatomical structures. Tomographic radiopharmaceutical imaging provides in vivo three-dimensional maps of a pharmaceutical labeled with a gamma ray emitting radionuclide. The distribution of radionuclide concentrations are estimated from a set of projectional images acquired at many different angles around the patient. Brain imaging has become an important diagnostic and research tool in nuclear medicine. The ultimate value of this procedure depends on good technique in acquisition setup and proper data reconstruction [Bruyant, 2002; Ramírez et al., 2008; Vandenberghea et al., 2001].

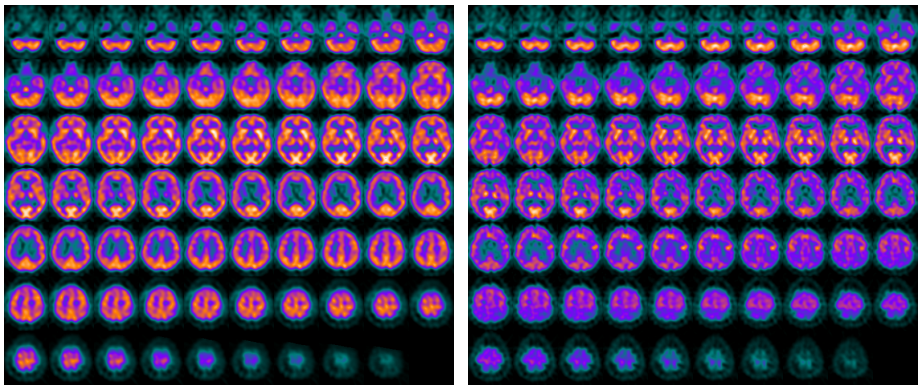
10.1. Diagnosis of Alzheimer's Disease by means of ECT

Alzheimer's disease (AD) diagnosis is usually based on the information provided by a careful clinical examination, a thorough interview of the patient and relatives, and a neuropsychological assessment [Braak and Braak, 1997; Cummings et al., 1998; Hoffman et al., 2000]. An ECT study is frequently used as a complimentary diagnostic tool in addition to the clinical findings [Higdon et al., 2004; Alexander et al., 2002]. However, in late-onset AD there are minimal perfusion in the mild stages of the disease, and age-related changes, which are frequently seen in healthy aged people, have to be discriminated from the minimal disease-specific changes. These minimal changes in the images make visual diagnosis a difficult task that requires experienced experts.

Positron emission tomography (PET) is non-invasive, nuclear medicine imaging technique which produces a three-dimensional image of functional processes in the body. The system detects pairs of gamma rays emitted indirectly by a positron-emitting radionuclide (tracer), which is introduced into the body on a biologically active molecule.When the tracer is ^{18}F-Fluorodeoxyglucose (F-FDG), its concentration give us information about tissue metabolic activity, measuring the brain's rate of glucose metabolism. Images of tracer concentration in 3-dimensional space within the brain are then reconstructed by computer analysis.

Single Photon Emission Computed Tomography (SPECT) is an ECT imaging technique that was initially developed in the 1960s, but was not widely used in clinical practice until the 1980s. It is mainly used when structural information is not enough to detect or monitor a functional disorder. Thus, SPECT is a noninvasive, three-dimensional functional imaging modality that provides clinical information regarding biochemical and physiologic processes in patients. Functional SPECT imaging has been found to be a valuable aid for the early diagnosis of the AD [Goethals et al., 2002]. Perfusion images are images that show the regional cerebral blood flow (rCBF) in the brain. Figure 10.1 shows typical brain perfusion patterns of a normal subject and a patient affected by Alzheimer's disease. Although many studies exist no final agreement has been achieved for the best regions of the brain to use when diagnosing Alzheimer's disease:

- Many studies have shown the temporo-parietal region to be practical



(a) (b)

Figure 10.1: Typical perfusion patterns of: a) a normal subject, and b) a patient affected by Alzheimer type dementia.

for the early detection of the disease in patients that are no longer characterized by specific cognitive impairment but by general cognitive decline [Claus et al., 1994]. Although bilateral temporo-parietal abnormalities, with or without other regional defects, are known as the predominant pattern for Alzheimer's disease, they appear to be neither sensitive nor specific for early Alzheimer's disease.

- Perfusion deficits in posterior cingulate gyri and precunei perfusion regions are probably more specific and more frequent in early Alzheimer's disease than temporo-parietal deficits [Kogure et al., 2000].

- Hypo-perfusion in the medial temporal lobe and hippocampus is not found in mild Alzheimer's disease due to the difficulties of imaging these deep brain structures [Braak and Braak, 1991].

10.2. Databases description and preprocessing

The ultimate value of the CAD system strongly depends on effective techniques in acquisition, proper data reconstruction and image registration. After introducing all the necessary knowledge and tools for building the diagnosis system, this section shows the image acquisition setup and preprocessing steps of the scans prior to defining the classifier.

10.2.1. Image acquisition

For the SPECT image acquisition, the patients are injected with a gamma emitting ^{99m}Tc-ECD radiopharmeceutical and the SPECT scan is acquired by a three-head gamma camera Picker Prism 3000. A total of 180 projections are taken for each patient with a 2-degree angular resolution. Finally, images of the brain cross sections are reconstructed from the projection data using the filtered backprojection (FBP) algorithm described below in combination with a Butterworth noise removal filter (see Section 10.2.2). The SPECT images used in this work were initially labeled by experienced clinicians of the "Virgen de las Nieves" Hospital (Granada, Spain). The database consists of 91 patients: 41 labeled as NORMAL and 50 as AD.

FDG PET scans were acquired according to a standardized protocol. A 30-min dynamic emission scan, consisting of 6 5-minutes frames, was acquired starting 30 minutes after the intravenous injection of 5.0-0.5 mCi of ^{1}8F-FDG, as the subjects, who were instructed to fast for at least 4 hours prior to the scan, lay quietly in a dimly lit room with their eyes open and minimal sensory stimulation. Data were corrected for radiationattenuation and scatter using transmission scans from Ge−68 rotating rod sources and reconstructed using measured attenuation correction and image reconstruction algorithms specified for each scanner. In this work, two PET databases are used. On the one hand, a 60 PET images database provided by the private clinic "PET *Cartuja*" (Seville, Spain) consisting of 18 controls and 42 AD patients. On the other hand, PET images used in the preparation of this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5-year public-private partnership. Following the scan, each image was reviewed for possible artifacts at the University of Michigan and all raw and processed study data was archived.

10.2.2. Image reconstruction

An image of the cross section of the brain can be reconstructed from projection data [Lange and Carson, 1984; Vardi et al., 1985; Hudson and Larkin, 1994; Bruyant, 2002; Chornoboy et al., 1990]. In ideal conditions, projections are a set of measurements of the integrated values of some parameter of the

object. If the object is represented by a two dimensional function $f(x, y)$ and each line integral by the (θ, t) parameters, the line integral is defined as:

$$P_\theta(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - t)dxdy, \tag{10.1}$$

where $P_\theta(t)$ is known as the Radon transform of the function $f(x, y)$.

The key to tomographic imaging is the *Fourier Slice Theorem* which relates the measured projection data to the two-dimensional Fourier transform of the object cross section. The Fourier Slice Theorem is stated as follows: "The Fourier transform $S_\theta(w)$ of a parallel projection $P_\theta(t)$ of an image $f(x, y)$ taken at angle θ and defined to be:

$$S_\theta(w) = \int_{-\infty}^{+\infty} P_\theta(t)\exp(-j2\pi wt)dt, \tag{10.2}$$

gives a slice of the two-dimensional Fourier transform:

$$F(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\exp(-j2\pi(ux + vy))dxdy, \tag{10.3}$$

subtending an angle θ with the u-axis", that is,

$$S_\theta(w) = F(u = w\cos\theta, v = w\sin\theta). \tag{10.4}$$

The above result is the essence of straight ray tomography and indicates that by having projections of an object function at angles θ_1, θ_2, ..., θ_k and taking the Fourier transform of them, the values of $F(u, u)$ can be determined on radial lines. In practice only a finite number of projections of an object can be taken. In that case it is clear that the function $F(u, v)$ is only known along a finite number of radial lines so that one must then interpolate from these radial points to the points on a square grid.

Projection data used in this study are reconstructed using the filtered backprojection (FBP) algorithm that is easily derived from the Fourier Slice Theorem. An image of the cross section $f(x, y)$ of an object can be recovered by:

$$f(x, y) = \int_0^\pi Q_\theta(x\cos\theta + y\sin\theta)d\theta, \tag{10.5}$$

where

$$Q_\theta(t) = \int_{-\infty}^{+\infty} S_\theta(w)|w|\exp(j2\pi wt)dw. \tag{10.6}$$

The FBP algorithm then consists of two steps: the filtering part, which can be visualized as a simple weighting of each projection in the frequency domain, and the backprojection part.

A major drawback of FBP is the undesired amplification of the high frequency noise and its impact on image quality. These effects are caused by the filtering operation or multiplication of $S_\theta(w)$ by $|w|$ in equation 10.6. In order to attenuate the high frequency noise amplified during FBP reconstruction, a number of window functions has been proposed. In this way, the reconstruction method described by equations 10.5 and 10.6 is normally redefined by applying a frequency window which returns to zero as the frequency tends to π. Among the most common window functions used for FBP reconstruction are: *i*) sinc (Shepp-Logan filter), *ii*) cosine, *iii*) Hamming and, *iv*) Hanning window functions. However, even when the reconstruction noise is kept low using a noise controlled FBP approach, the noise captured by the acquisition system needs to be filtered out to improve the quality of reconstructed images. Moreover, the preprocessing stage of most automatic SPECT image processing systems often incorporates prefiltering, reconstruction and postfiltering to minimize the noise acquired by the gammacamera as well as the noise amplified during FBP reconstruction.

10.2.3. Image registration

The complexity of brain structures and the differences between brains of different subjects make necessary the normalization of the images with respect to a common template. This step allows us to compare the voxel intensities of the brain images of different subjects. The images are first spatially normalized using the SPM software [Friston et al., 2007] in order to ensure that the voxels in different images refer to the same anatomical positions in the brain. The normalized method assumes a general affine model with 12 parameters [Woods, 2000] and a cost function which presents an extreme value when the template and the image are matched together. The objective function to be optimized is the mean squared difference between both images, the source and the template:

$$\text{CF} = \sum_i (f(\mathbf{M}\mathbf{x}_i) - g(\mathbf{x}_i))^2, \tag{10.7}$$

where f denotes the source image and g the template. For each voxel $\mathbf{x} = (x_1, x_2, x_3)$ in an image, the affine transformation into the coordinates $\mathbf{y} =$

(y_1, y_2, y_3) is expressed by a matrix multiplication $\mathbf{y} = \mathbf{M}\mathbf{x}$.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}. \tag{10.8}$$

After the affine normalization, the resulting image is registered using a more complex non-rigid spatial transformation model. The deformations are parameterized by a linear combination of the lowest-frequency components of the three-dimensional cosine transform bases [Ashburner and Friston, 1999]. A small-deformation approach is used and regularization is achieved by the bending energy of the displacement field.

Figure 10.2 shows an example of the operation of the normalization process on SPECT images. Left column shows arbitrary source images in the dataset, central column shows the template used for image registration, and finally the corresponding normalized images are shown in the right column. It is clearly shown that the transformed image matches the shape of the template.

After the spatial normalization, a 95×69×79 voxel representation of each subject is obtained. Each voxel represents a brain volume of 2.18×2.18×3.56 mm^3. Finally, intensity level of the images is normalized to the maximum intensity, which is computed for each volume individually by averaging over the 3 % of the highest voxel intensities following a procedure similar to Saxena et al. [1998].

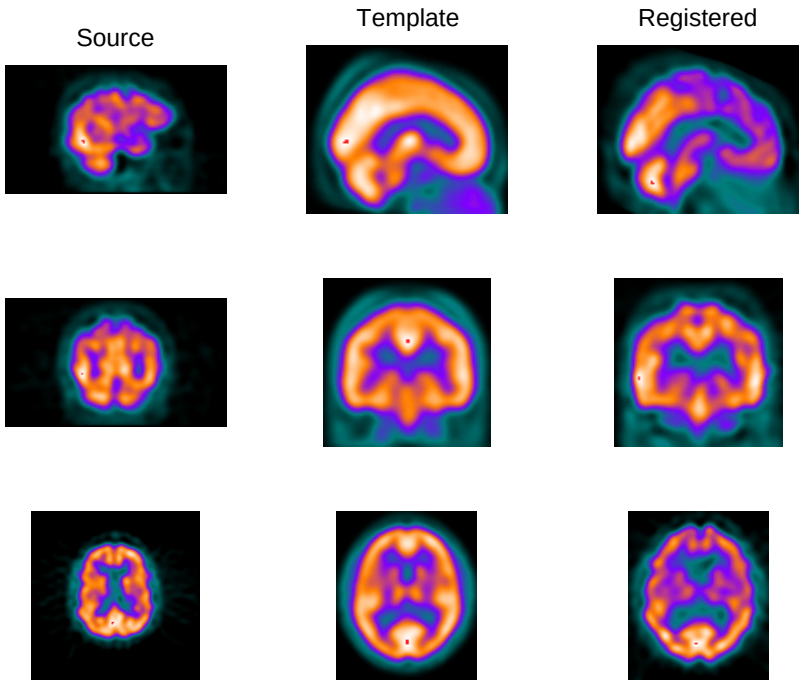


Figure 10.2: Three SPECT images. Left column: Source image. Central column: Template. Right column: Transformed image

CHAPTER 11

Classification Methods

The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories. Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems. Cases such as the digit recognition example, in which the aim is to assign each input vector to one of a finite number of discrete categories, are called classification problems. In this chapter a theoretical background of the classification methods used in this work to determine the class pattern of an individual brain image is presented.

11.1. Support Vector Machines

Support vector machines [Burges, 1998; Vapnik, 1995, 1998] are widely used for pattern recognition in a number of applications due to its ability to learn from experimental data. The reason is that SVM are much more effective than other conventional parametric classifiers. SVM separate a given set of binary labeled training data with a hyperplane that is maximally distant from the two classes (known as the maximal margin hyper-plane). The objective is to build a function $f : \mathbb{R}^n \longrightarrow \{\pm 1\}$ using training data that is, N-dimensional patterns $\mathbf{x}_i$ and class labels y_i:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_l, y_l) \in \mathbb{R}^n \times \{\pm 1\}, \tag{11.1}$$

so that f will correctly classify new examples $(\mathbf{x}, y)$.

Linear discriminant functions define decision hypersurfaces or hyperplanes in a multidimensional feature space, that is:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0, \tag{11.2}$$

where $\mathbf{w}$ is known as the weight vector and w_0 as the threshold. The weight vector $\mathbf{w}$ is orthogonal to the decision hyperplane and the optimization task consists of finding the unknown parameters w_i, i= 1,...,n, defining the decision hyperplane.

Let $\mathbf{x}_i$, i=1,2,...,N, be the feature vectors of the training set, X. These belong to either of the two classes, ω_1 or ω_2. If the classes were linearly separable the objective would be to design a hyperplane that classifies correctly all the training vectors. The hyperplane is not unique and the selection process focuses on maximizing the generalization performance of the classifier, that is, the ability of the classifier, designed using the training set, to operate satisfactorily with new data. Among the different design criteria, the maximal margin hyperplane is usually selected since it leaves the maximum margin of separation between the two classes. Since the distance from a point $\mathbf{x}$ to the hyperplane is given by $z = |g(\mathbf{x})|/||\mathbf{w}||$, scaling $\mathbf{w}$ and w_0 so that the value of $g(\mathbf{x})$ is +1 for the nearest point in ω_1 and -1 for the nearest points in ω_2, reduces the optimization problem to maximizing the margin: $2/||\mathbf{w}||$ with the constraints:

$$\begin{array}{ll} \mathbf{w}^T\mathbf{x} + w_0 \geq 1, & \forall \mathbf{x} \in \omega_1, \\ \mathbf{w}^T\mathbf{x} + w_0 \leq 1, & \forall \mathbf{x} \in \omega_2, \end{array} \tag{11.3}$$

or equivalently, minimizing the cost function $J(\mathbf{w}) = 1/2||\mathbf{w}||^2$ subject to:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, \quad i = 1, 2, ..., N. \tag{11.4}$$

Thus, designing the classifier leads to a nonlinear (quadratic) optimization task subject to a set of linear inequality constraints. By using the optimization methodology adopted by Karush-Kuhn-Tucker [Burges, 1998], the solution $\mathbf{w}$ is found to be a linear combination of $N_s \leq N$ feature vectors named support vectors and the optimum hyperplane is called the support vector machine. The support vectors are the training vectors that are closest to the linear classifier since lie on either of the two hyperplanes, i.e.: $\mathbf{w}^T\mathbf{x} + w_0 = \pm 1$. On the other hand, the optimization process with inequality constraints guarantee any local minimum is also global and unique so that the optimal maximal margin hyperplane defining the support vector machine is unique.

For non-separable classes, the optimization process needs to be modified in an efficient and elegant manner. In mathematical terms, the maximal margin hyperplane for non-separable data is selected by minimizing the cost function:

$$J(\mathbf{w}, w_0, \xi) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i, \tag{11.5}$$

subject to the constraints:

$$y_i[\mathbf{w}^T x_i + w_0] \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad i = 1, 2, ..., N. \tag{11.6}$$

When no linear separation of the training data is possible, SVM can work in combination with techniques of kernels so that the hyperplane defining the SVM corresponds to a non-linear decision boundary in the input space. If the data is mapped to some other (possibly infinite dimensional) Euclidean space using a mapping $\Phi(\mathbf{x})$, the training algorithm only depends on the data through dot products in such an Euclidean space, i.e. on functions of the form $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. If a "kernel function" K is defined such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, it is not necessary to know the Φ function during the training process. In the test phase, an SVM is used by computing dot products of a given test point $\mathbf{x}$ with $\mathbf{w}$, or more specifically by computing the sign of

$$f(\mathbf{x}) = \sum_{i=1}^{N_S}\alpha_i y_i \Phi(\mathbf{s}_i) \cdot \Phi(\mathbf{x}) + w_0 = \sum_{i=1}^{N_S}\alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + w_0, \tag{11.7}$$

where $\mathbf{s}_i$ are the support vectors.

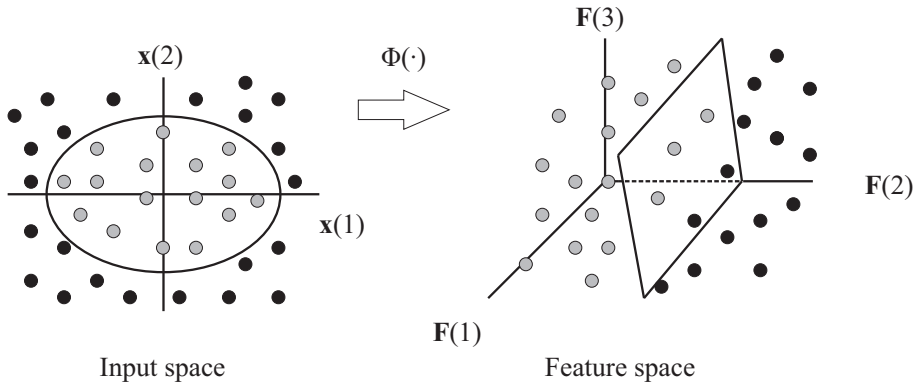


Figure 11.1: Effect of mapping the input space to the feature space where the separation boundary becomes linear

Thus, the use of kernels in SVM enables to map the data into some other dot product space (called feature space) $\mathcal{F}$ via a nonlinear transformation $\Phi : \mathbb{R}^n \longrightarrow \mathcal{F}$ and perform the above linear algorithm in $\mathcal{F}$. Figure 11.1 illustrates this process where the 2D input space is mapped to a 3D feature space where the data is linearly separable. In the input space, the hyperplane corresponds to a nonlinear decision function whose form is determined by the kernel. There are three common kernels that are used by SVM practitioners for the nonlinear feature mapping:

- Polynomial
$$K(\mathbf{x}, \mathbf{y}) = [\gamma(\mathbf{x} \cdot \mathbf{y}) + c]^d. \tag{11.8}$$
- Radial basis function (RBF)
$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2). \tag{11.9}$$
- Sigmoid
$$K(\mathbf{x}, \mathbf{y}) = \tanh(\gamma(\mathbf{x} \cdot \mathbf{y}) + c). \tag{11.10}$$

Thus, the decision function is nonlinear in the input space

$$f(\mathbf{x}) = \text{sign}\{\sum_{i=1}^{N_s} \alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + w_0\}, \tag{11.11}$$

and the parameters α_i are the solution of a quadratic optimization problem that are usually determined by Quadratic Programming (QP) or the

well known Sequential Minimal Optimization (SMO) algorithm [Platt, 1999]. Many classification problems are separable in the feature space and are able to obtain better results by using RBF kernels instead of linear and polynomial kernel functions [Clarkson and Moreno, 1999; Ganapathiraju et al., 2004].

11.2. Bayesian classifier

In a two-hypothesis test, the optimal decision rule that minimizes the error probability is the Bayes classifier. Bayesian-based classifiers have been successfully used in face recognition problems [Shakhnarovich and Moghaddam, 2004; Wang and Tang, 2003]. The Bayes classifier evaluates the *a posteriori* probability function [Fukunaga, 1990]. Let $\{\omega_1, \omega_2, ..., \omega_c\}$ denote the object classes and $\mathbf{x}$ a feature vector obtained by some feature extraction technique. The *a posteriori* probability function of $\mathbf{x}$ given ω_i is defined as

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}, \quad i = 1, 2, ..., c. \tag{11.12}$$

where $P(\omega_i)$ is *a priori* probability, $P(\omega_i|\mathbf{x})$ the conditional probability density function of $\mathbf{x}$ given ω_i and $P(\mathbf{x})$ is the mixture density. The maximum a posteriori (MAP) decision rule for the Bayes classifier is defined as

$$P(\mathbf{x}|\omega_i)P(\omega_i) = \max_j\{P(\mathbf{x}|\omega_j)P(\omega_j)\}, \quad \mathbf{x} \in \omega_i \tag{11.13}$$

The test vector data $\mathbf{x}$ is classified to ω_i of which the a posteriori probability given $\mathbf{x}$ is the largest between the classes. For the classification problem we are dealing with, the *a priori* probability $P(\omega_i)$ is initially set to 0.5, that is, AD and NORMAL classes are equally probable. Regarding the conditional probability density function, usually there are not enough samples to estimate it for each class (within-class density). A compromise, therefore, is to make an assumption of a particular density form and convert the general density estimation problem into a parametric one. The within class densities are usually modeled as normal distributions

$$P(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_i|^{1/2}} \times exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{M}_i)^t\mathbf{\Sigma}_i^{-1}(\mathbf{x} - \mathbf{M}_i)\right\} \tag{11.14}$$

where $\mathbf{M}_i$ and $\mathbf{\Sigma}_i$ are the mean and covariance matrix of class ω_i, respectively.

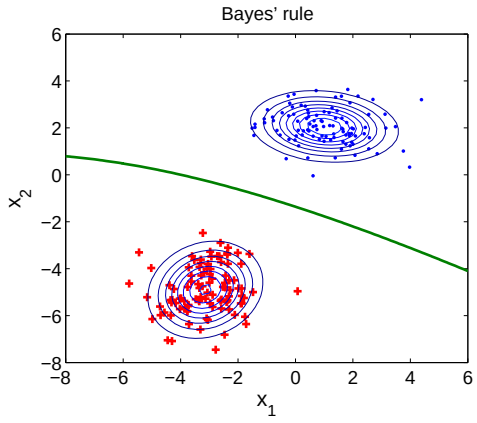


Figure 11.2: Decision boundary designed by a Bayesian classifier considering Gaussian density distributions and equal *a priori* probabilities for both classes.

Figure 11.2 shows samples of two classes and the decision boundary defined by the Bayes' rule. Samples density distributions are assumed to be Gaussians. Ellipses around the clusters show lines of equal probability density of the Gaussian. Considering the *a priori* probabilities equal for both class 1 (represented by blue points) and class 2 (red crosses), the green line divides up the 2D plane in two regions: a sample placed above or below this line is more likely to belong to class 1 or class 2, respectively.

11.3. Neural Networks

An Artificial Neural Network (ANN) [McCulloch and Pitts, 1943] is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, processes information. ANNs can be viewed as weighted directed graphs in which artificial neurons are nodes and directed edges (with weights) are connections between neuron outputs and neuron inputs. Based on the connection pattern (architecture), ANNs can be grouped into two categories: *i*) feed-forward networks, in which graphs have no loops, and *ii*) recurrent (or feedback) networks, in which loops occur because of feedback connections. Different connectivities yield different network behaviors. Generally speaking, feed-forward networks are static, that is, they produce only one set of output values rather than a sequence of values from a given input. Feed-forward networks are memory-less in the sense that their response to an input is independent of the previous network state. Recurrent

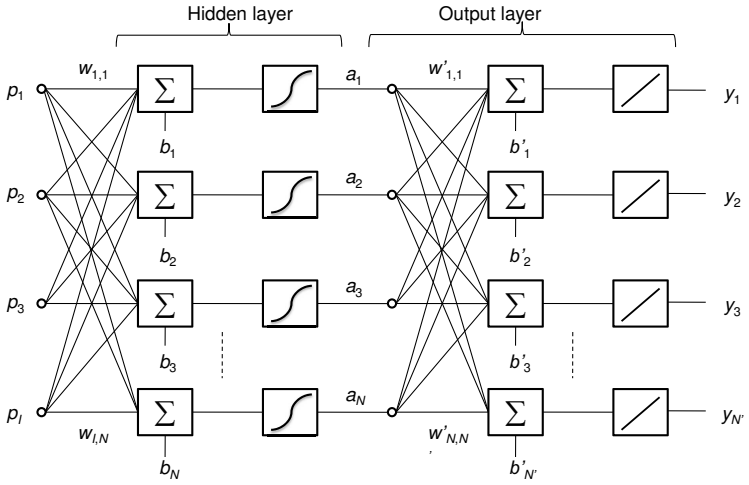


Figure 11.3: Feed-forward neural network architecture with hidden layer of neurons plus linear output layer.

or feedback networks, on the other hand, are dynamic systems. When a new input pattern is presented the neuron outputs are computed. Because of the feedback paths, the inputs to each neuron are then modified, which leads the network to enter a new state.

Feed-forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons as shown in Figure 11.3. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors.

Learning process in the ANN context can be viewed as the problem of updating network architecture and connection weights so that a network can efficiently perform a specific task. The ability of ANNs to automatically learn from examples makes them attractive and exciting. The development of the back-propagation learning algorithm for determining weights in a multilayer perceptron has made these networks the most popular among ANN researchers.

For the experiments presented in this work a feed-forward neural network with the following configuration was used:

- One hidden layer and increasing number of neurons and a linear output layer.
- Hyperbolic tangent sigmoid transfer function: $f(n) = 2/(1 + exp(-2 * n)) - 1$, for input layers.
- Linear transfer function: $f(n) = n$, for output layer.
- Weight and bias values are updated according to Levenberg-Marquardt optimization.
- Gradient descent with momentum weight and bias is used as learning function.

CHAPTER 12

Regions of Interest

In the AD diagnosis process by means of visual exploration of the images, clinicians focus on those brain regions where the AD produces a decrease of the brain activity. That means that those regions that are not affected by the AD do not provide any useful information for the AD diagnosis and therefore, they can be discarded. In this chapter, the automatic detection of the regions of interest (ROIs) is sought by means of an exhaustive exploration carried out throughout the whole brain volume. This exploration is performed by extracting smaller subsets of voxels called *components*, which are used to train a unique SVM classifier and by some pasting-vote method a final diagnostic on the patient will be given. An accuracy map can be made up in order to define a criterion to select only the most discriminant components and locate the ROIs for the AD detection.

12.1. Component-based feature extraction

Component-based feature extraction process has actually been applied to the face detection problem [Heisele et al., 2001b; Schneiderman and Kanade, 2000; Heisele et al., 2001a; Brunelli and Poggio, 1993; Nefian and Hayes, 1999; Wiskott, 1995], centering the components on the eyes, nose and mouth. Our approach does not search for any particular zone, but computes a systematic scan of the whole brain image. For the analysis, let N be the number of patients, s the number of components the brain is divided into, and C_i $i = 1, 2, ..., s$ denote subset of voxels that make up the i-th component. The whole brain volume $I(V)$ of a given patient can be expressed as:

$$I(V) = \bigcup_{m=1}^{s} \mathbf{I}(C_m) - \bigcap_{m=1}^{s} \mathbf{I}(C_m) \tag{12.1}$$

where the second part of the right term eliminates the redundance due to the possible overlapping of the components. The j-th volume I_j will have an associated label y_j, which will be $+1$ in case the patient is AD, and -1 in case the patient is NORMAL. Thus, this label y_j is shared with all the components of the image I_j $(C_{1j}, C_{2j}, ..., C_{sj})$. Each component C_{ij} is used as the feature vector input to train and test a single SVM classifier by means of a Leave-One-Out cross-validation strategy, that is, the classifier is trained on all but one component $\{C_{i1}, C_{i2}, ..., C_{is-1}\}$, categorizing the remaining component C_{is} by a label $\ell_{ij} \in \{\pm 1\}$. This process will provide as many SVM classifiers as components the brain is broken down into, being each classifier trained only on its associated region of the brain volume. If the information of a particular region is important with regard to the AD diagnosis, the associated SVM classifier will have a good performance in the classification task. In other words, after the training and testing steps, it is possible to assign an accuracy value $A^{(i)}$, $i = 1, 2, ..., s$ to each component according to the number of correctly classified patients it provided. The accuracy map for SPECT images are represented in Figure 12.1. The ROIs correspond to the most discriminant components, near red colors, meaning dark red a 100 % accuracy in classification.

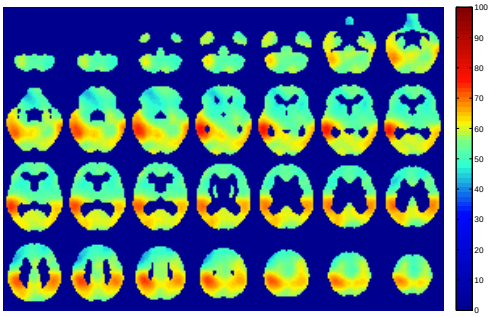


Figure 12.1: Map of the ROIs for SPECT images determined by the values $A^{(i)}$.

12.2. SVM ensemble

The brain image of each patient is divided into components with several shapes and sizes. These components are obtained after a feature extraction process and then, an individual SVM is applied on each component. This yields to an SVM ensemble for each brain image that can be used to grow a global diagnostic of the patient.

In SVM ensemble, individual SVMs are aggregated to make a collective decision in several ways such as the majority voting, least-squares estimation-based weighting, and the double layer hierarchical combing [Hyun-Chul et al., 2003]. The training SVM ensemble can be conducted in the way of bagging or boosting. In bagging, each individual SVM is trained independently using the randomly chosen training samples via a bootstrap technique. In boosting, each individual SVM is trained using the training samples chosen according to the samples probability distribution that is updated in proportion to the error in the sample. When memory limitations exist, voting many classifiers built on small subsets of data ("pasting small votes") is an approach for learning from massive datasets [Breiman, 1999]. SVM ensemble is essentially a type of cross-validation optimization of single SVM, having a more stable classification performance than other models.

12.2.1. Pasting-votes methods

The SVM ensemble will serve us to define a new decision function based on the pasting-votes technique Breiman [1999]. The function defined as

$$\mathcal{F}(I_j) = \sum_{i=1}^{s} \ell_{ij} \tag{12.2}$$

is an non-weighted sum of votes that each component casts and will classify the patient j as NORMAL if $\mathcal{F}(I_j) < 0$, and as AD if $\mathcal{F}(I_j) > 0$. This method is therefore called *majority voting*. An improvement is easily introduced by assigning a *relevance* weight to each vote based on the individual accuracy $A^{(i)}$, which defines a new function:

$$\mathcal{R}(I_j) = \sum_{i=1}^{s} \ell_{ij} A^{(i)} \tag{12.3}$$

Furthermore, the considered components can be limited to only those that provide higher accuracy rates. The definition of a criterion function $\mathcal{T}$ based on the weights will pick only the s' most discriminant components, where s' is chosen by setting a threshold T in the accuracy values: only those components $C' \subset I$ whose values $A^{(i)}$ are higher than T are allowed to vote:

$$\mathcal{T}(I_j) = \sum_{i \subset S} \ell_{ij}, \quad S = \{i \setminus A^{(i)} > T\} \tag{12.4}$$

12.3. Results

The component-based approach for the detection of the AD was evaluated on three different datasets:

- 79 SPECT images: 41 labeled as NORMAL and 38 labeled as AD.
- 60 PET images (PET *Cartuja*): 18 labeled as NORMAL and 42 labeled as AD.
- 192 ADNI images: 97 labeled as NORMAL and 95 as AD.

	v	15 Elongated	20 Elongated	25 Elongated	Cubic
SPECT	4	79.75 %	82.28 %	81.01 %	75.95 %
	5	79.75 %	78.48 %	83.54 %	83.54 %
	6	83.54 %	82.28 %	84.81 %	84.81 %
	7	83.54 %	86.08 %	**87.34 %**	82.28 %
PET *Cartuja*	4	96.67 %	96.67 %	96.67 %	98.33 %
	5	96.67 %	96.67 %	96.67 %	96.67 %
	6	96.67 %	96.67 %	96.67 %	95 %
	7	96.67 %	96.67 %	96.67 %	88.33 %

Table 12.1: Results obtained by the majority voting recount for different types of components and subsampling factors v

The dimensionality reduction of the original $79 \times 95 \times 69$ voxel sized brain volume was performed by subsampling the volumes by a factor $v \times v \times v$ where v ranges from 4 to 7. After that, the brain volumes were divided into 15, 20 or 25 elongated as well as cubic components of size $4 \times 4 \times 4$. Elongated components are independently studied along each axis direction, and each one is shifted 30 times in such a way the shifted component overlaps. Therefore, the total number of elongated components is $15 \times 3 \times 30 = 1350$, $25 \times 3 \times 30 = 2250$ or $25 \times 3 \times 30 = 2250$. Regarding cubic components, they are made up of $4 \times 4 \times 4$ voxel-sized subsets, and the number of components depends on the reduction factor v.

When evaluating the majority voting $\mathcal{F}$ function (see Table 12.1), best accuracy values for SPECT images were found when a $7 \times 7 \times 7$ initial averaging was performed for 25 elongated components. For PET *Cartuja* images, best results were achieved when a $4 \times 4 \times 4$ initial dimension reduction is applied for cubic components. For these images and making use of the $\mathcal{F}$ function, only one patient, which was initially labeled as NORMAL, is misclassified by the classifier. This misclassification is corrected by using the $\mathcal{R}$ function instead. Clinicians have detected that, although it is a normal patient in the sense that he does not present any sign of AD, there is a peculiarity in his thalamus metabolism.

Figure 12.2 shows the accuracy results obtained when the function $\mathcal{R}$ is evaluated for different values of the threshold T. The number of components considered when the threshold is applied is also plotted and as expected, it decreased when the T increases, making more restrictive the components selection criterion. Setting $T = 84\,\%$ and $T = 92\,\%$ yields to 97,47 % and 98,33 % accuracy rates for SPECT and PET *Cartuja* images respectively. This can be considered a fair trade-off between the strength of the threshold

and the number of considered components, since for these T values more than 200 votes are considered. Component-based technique outperforms the baseline VAF approach, which reaches 78.5 % and 96.6 % accuracy values for SPECT and PET *Cartuja* images respectively.

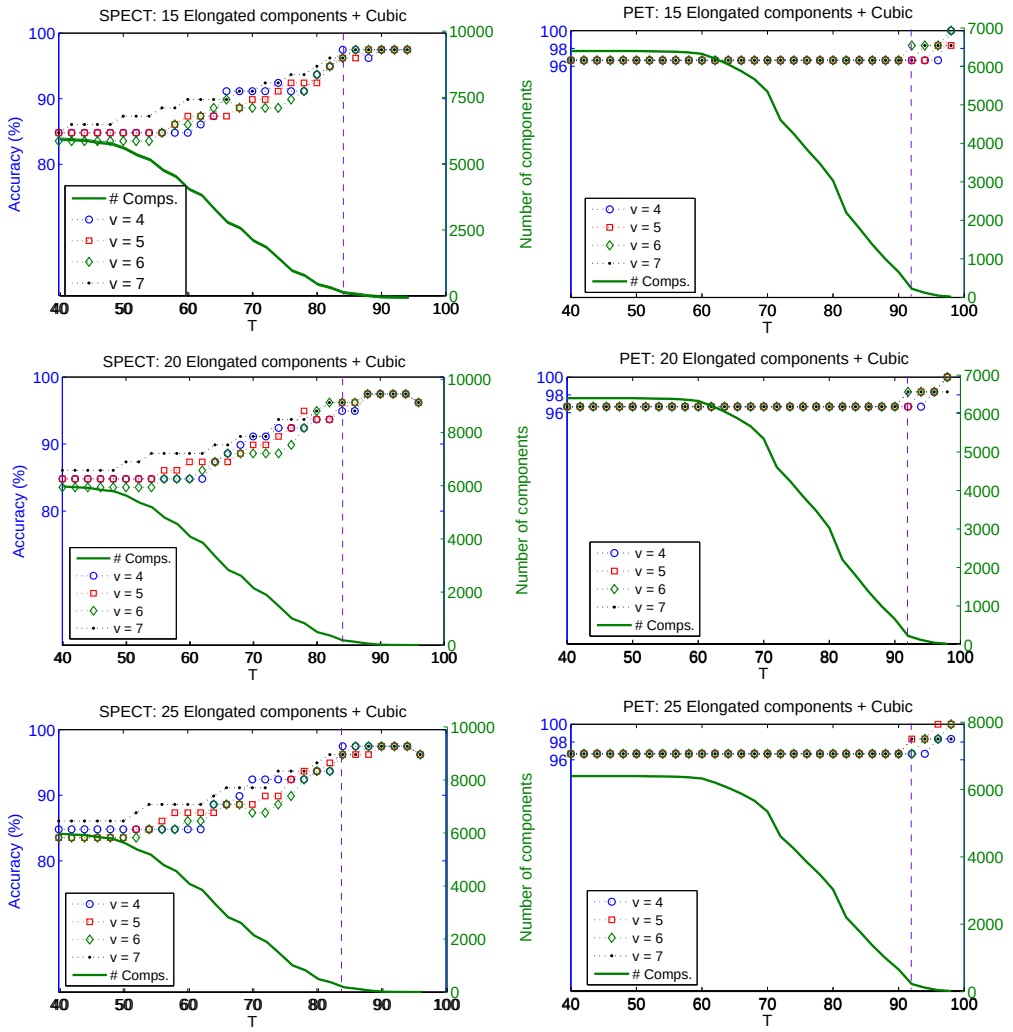


Figure 12.2: Results obtained by evaluating the relevance voting function $\mathcal{R}$ for different values of the threshold T for SPECT (left) and PET *Cartuja* (right). The higher the threshold is, the lower the number of components are considered.

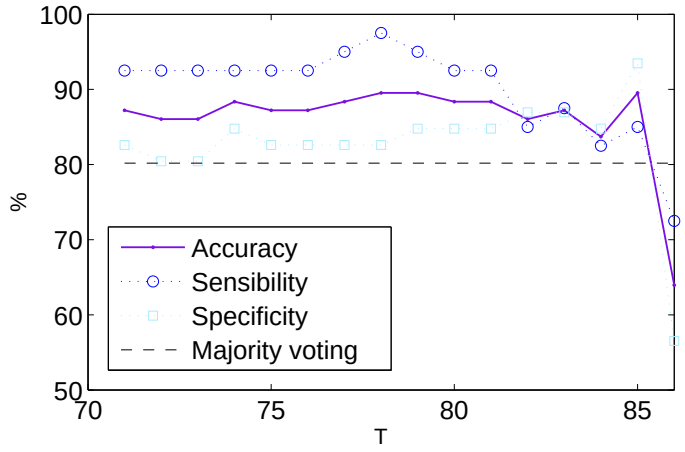


Figure 12.3: ADNI: Accuracy, sensibility and specificity values obtained by means of the component-based technique when the relevance voting function $\mathcal{R}$ for different values of the threshold T is evaluated. Comparison to majority voting.

Regarding the ADNI database, results of accuracy, sensitivity and specificity are shown in Figure 12.3. Due to the high computational cost required, the validation of the ROIs technique on the ADNI database was performed by the k-fold strategy instead of Leave-One-Out. The samples are initially divided into 2 random subgroups containing the same number of samples. The first group is used to compose the precision map of the components and the second group is used to validate the system. It can be seen from the curves that sensibility is in most cases higher than specificity. This is due to the labeling criteria used for the categorization of the ADNI images, for which only clinical tests are used without including the information contained in the images. Clinical tests have shown to provide high specificity and low sensibility values [Jobst et al., 1998], and this is therefore matched by the CAD system.

CHAPTER 13

Principal Component and Linear Discriminant Analyses

Principal component analysis (PCA) [Jolliffe, 1986], also known as Karhunen-Loeve (KL) transform, has been called one of the most valuable results from applied linear algebra. PCA is used abundantly in all forms of analysis - from neuroscience to computer graphics - because it is a simple, non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it. Linear Discriminant Analysis (LDA) usually provides better features for classification since it is a class specific projection. This technique and other PCA-based methods have been successfully applied for different image classification purposes [Kirby and Sirovich, 1990; Spetsieris et al., 2009; Turk and Pentland, 1991], and specifically for neuroimage classification problems [López et al., 2009c,b]. In this chapter, we present several classification schemes based on these linear projections.

13.1. Principal Component Analysis

Principal Component Analysis (PCA) corresponds to multivariate approaches and has been already applied to functional brain images in a descriptive fashion, where there exists the impossibility of using this transformation to make any statistical inference [Friston et al., 2007]. However, in this work, a new PCA-based approach is used in combination with supervised learning methods, which in turn solves the *small sample size* problem since the dimension of the feature space undergoes a significant reduction.

PCA generates an orthonormal basis vector that maximizes the scatter of all the projected samples. After the preprocessing steps, the n remaining voxels for each subject are rearranged into a vector form. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ be the sample set of these vectors, where N is the number of patients. After normalizing the vectors to unity norm and subtracting the grand mean, a new vector set $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$ is obtained, where each $\mathbf{y}_i$ represents an $n-$dimensional normalized vector, $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in})^T, i = 1, 2, ..., N$. The covariance matrix of the normalized vectors set is defined as

$$\mathbf{\Sigma}_Y = \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i\mathbf{y}_i^T = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T \tag{13.1}$$

and the eigenvector and eigenvalue matrices $\mathbf{\Phi}$, $\mathbf{\Lambda}$ are computed as

$$\mathbf{\Sigma}_Y\mathbf{\Phi} = \mathbf{\Phi}\mathbf{\Lambda} \tag{13.2}$$

Note that $\mathbf{Y}\mathbf{Y}^T$ is an $n \times n$ matrix while $\mathbf{Y}^T\mathbf{Y}$ is an $N \times N$ matrix. If the sample size N is much smaller than the dimensionality n, then diagonalizing $\mathbf{Y}^T\mathbf{Y}$ instead of $\mathbf{Y}\mathbf{Y}^T$ reduces the computational complexity [Turk and Pentland, 1991]

$$(\mathbf{Y}^T\mathbf{Y})\mathbf{\Psi} = \mathbf{\Psi}\mathbf{\Lambda}_1 \tag{13.3}$$

$$\mathbf{T} = \mathbf{Y}\mathbf{\Psi} \tag{13.4}$$

where $\mathbf{\Lambda}_1 = diag\{\lambda_1, \lambda_2, ..., \lambda_N\}$ and $\mathbf{T} = [\mathbf{\Phi}_1, \mathbf{\Phi}_2, ..., \mathbf{\Phi}_N]$. Derived from the *eigenface* concept [Turk and Pentland, 1991], and due to its still brain-like appearance, the eigenvectors or principal components (PCs) $\mathbf{\Phi}_i, i = 1, ..., N$

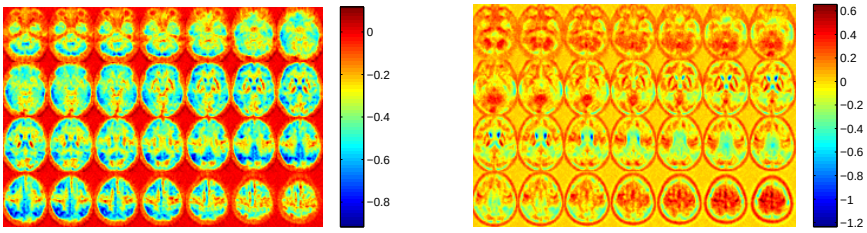


Figure 13.1: First (left) and second (right) *eigenbrains* extracted from the ADNI database. They represent the principal components where original images will be projected onto to obtain a dimension reduction.

of the covariance matrix are called *eigenbrains* [Álvarez et al., 2009a]. Figure 13.1 shows the first and second eigenbrains obtained from the ADNI database when AD and NORMAL subjects are considered in the solution of the eigenvalue problem.

13.1.1. Criterion for selecting the eigenbrains

PCA provides the best description of a complete dataset in terms of variability. However, in the PCs extraction process class labels are not taken into account. This fact motivates us to sort out the obtained PCA coefficients by using a more useful measure for classification purposes instead of their associated eigenvalues. We propose the Fisher Discriminant Ratio (FDR) as criteria for rearranging the obtained coefficients. This makes sense since the FDR takes into account the information of the labels preassigned to the data. It is defined as follows:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{13.5}$$

where μ_i and σ_i^2 denote the i-th class within class mean value and variance, respectively. Thus, in the classification stage the training set is used to compute the PCs and the FDR values of the resulting PCA projection coefficients are computed. The test sample is projected onto the computed PCs, and the obtained test coefficients are rearranged in decreasing order according to the previous FDR values obtained from the train set. In some cases, this rearrangement will improve the classification results as shown in Section 13.4. Other metrics to determine the robustness of the PCs used for classification

purposes are proposed in Markiewicz et al. [2009].

13.1.2. Slices of Interest (SOIs)

PCA can be applied to the entire brain volumes in order to obtain a better description of the whole database. Theoretically the obtained PCs will hold the main differences among the subjects that make the database up, among which hopefully the differences between the two classes samples can be found. However, when dealing with such high dimensional data, PCA is sensitive to any other differences, maybe derived from the normalization process or due to the different origins of the images, in such a way that using the main eigenbrains as projection axes might contain "noisy" differences – that is, not useful information for distinguishing the AD.

This fact motivates us to apply PCA to the brain volume areas of interest. We can then search for the slices of interest (SOI) and discard the rest for the classification task. We explore the slices along the three directions and apply PCA to determine each one's discriminatory capacity. This allows us to pre-locate the most useful voxels for the classification task we are dealing with.

13.2. Linear Discriminant Analysis

Since the learning set is labeled, it makes sense to use this information to build a more reliable method for reducing the dimensionality of the feature space. Here we argue that using class specific linear methods for dimensionality reduction and simple classifiers in the reduced feature space, one may get better recognition rates than with other multivariate approaches. LDA [Fisher, 1936] is an example of a class specific method, in the sense that it tries to 'shape' the scatter in order to make it more reliable for classification. Let the between-class scatter matrix be defined as

$$\mathbf{S}_b = \sum_{i=1}^{c} N_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{13.6}$$

and the within-class scatter matrix be defined as

$$\mathbf{S}_w = \sum_{i=1}^{c} \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T \tag{13.7}$$

where $\boldsymbol{\mu}_i$ is the mean image of class ω_i, N_i is the number of samples in class ω_i and c is the number of classes. If $\mathbf{S}_W$ is nonsingular, LDA chooses the optimal projection $\mathbf{W}_{opt}$ as the matrix with orthonormal columns which maximizes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples, i.e.,

$$\mathbf{W}_{opt} = \arg \max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} = [\mathbf{w}_1 \mathbf{w}_2 ... \mathbf{w}_l] \tag{13.8}$$

where $\{\mathbf{w}_i | i = 1, 2, ..., l\}$ is the set of generalized eigenvectors of $|\mathbf{S}_w^{-1} \mathbf{S}_b|$ and corresponding to the l largest generalized eigenvalues $\{\lambda_i | i = 1, 2, ..., l\}$, i.e.,

$$\mathbf{S}_b \mathbf{w}_i = \lambda_i \mathbf{S}_w \mathbf{w}_i \tag{13.9}$$

Note that there are at most $c - 1$ nonzero generalized eigenvalues, and so an upper bound on l is $c - 1$, where c is the number of classes [Duda and Hart, 1973].

In the image classification problem, one is confronted with the difficulty that the within-class scatter matrix $\mathbf{S}_w \in \mathbb{R}^{n \times n}$ is always singular. This is a consequence of the fact that the rank of $\mathbf{S}_w$ is at most $N - c$, and, in general, the number of images in the learning set N is much smaller than the number of selected features in each image n. This means that it is possible to choose the matrix $\mathbf{W_{opt}}$ such that the within-class scatter of the projected samples can be made exactly zero. In order to overcome the drawback of a singular $\mathbf{S}_w$, LDA is usually applied after the PCA transform. Thus, PCA reduces the dimension of the feature space to $m \leq N - c$, and then, the standard LDA transform is applied to reduce the dimension to $l = c - 1$ [Belhumeour et al., 1997].

13.3. Experiments

In this section, different combinations of the proposed feature extraction and classification schemes are evaluated on SPECT and PET databases. The obtained results are compared with the reference VAF method, which uses all the voxels in the brain images directly as features to train and test a linear SVM [Stoeckel et al., 2001]. All the experiments are carried out by the Leave-One-Out cross-validation strategy, that is, the complete classification system is trained by taking into account all the samples but one, which is used as test sample. This procedure is repeated as many times as samples in the database, leaving each sample out in each iteration. Finally, an average accuracy rate is computed. Leave-One-Out has been used to assess the discriminative accuracy of different multivariate analysis methods applied to the discrimination of frontotemporal dementia from AD [Higdon et al., 2004] and in classifying atrophy patterns based on magnetic resonance imaging (MRI) images [Fan et al., 2008].

In all experiments, the images were subsampled by a $2 \times 2 \times 2$ factor. For the SPECT database, a study on the SOIs is performed in order to detect the most discriminant slices.

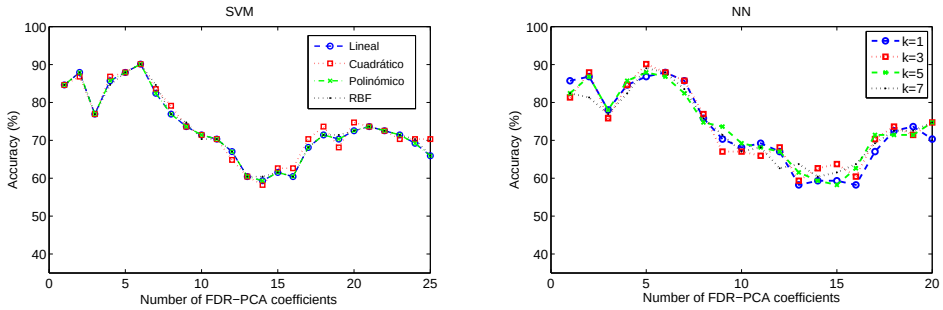
The datasets on which the experiments were carried out consist of:

- SPECT: 91 subjects; 41 labeled as NORMAL and 50 as AD.
- PET *Cartuja*: 60 subjects; 18 labeled as NORMAL and 42 as AD.
- ADNI:
  - Group 1: 105 subject; 52 labeled as NORMAL and 53 as AD.
  - Group 2: 166 subject; 52 labeled as NORMAL and 114 as MCI.

13.4. Results

13.4.1. Results on SPECT database

Classical PCA and LDA techniques have been evaluated on the SPECT database as feature extraction techniques in combination with the supervised classifiers SVM and NN. Results of evaluating the complete CAD system are



(a) (b)

Figure 13.2: SPECT: Accuracy results when the number of PCA coefficients projected onto the LDA axis increases for (a) SVM and (b) NN classifiers.

shown in Figure 13.2. The number of PCA coefficients used in the classification step has been varied from $m = 1$ to $m = N - 1$ being $N = 91$, the number of samples in the database. However, as expected, the best results are achieved when only a few of them are used. The application of LDA to these coefficients improves the accuracy rates in all cases in spite of the reduction of the feature space to $l = 1$. This final feature has shown to be more useful for separating NORMAL and AD classes. The rearrangement of the PCs by the FDR criterion improves the accuracy results with respect to the variance criterion. Both SVM and NN reach a peak accuracy of 90,11 % when this final feature is used to train them. In any case, all the experiments increased the sensitivity – and therefore, the final accuracy rate obtained by the baseline VAF approach, which was 85,71 % (83.67/87.8 sensitivity/specificity).

The search of SOIs in order to perform PCA only on the most interesting brain regions was also carried out on the SPECT database. When applying PCA slice by slice, we find that the classification results improve slightly with respect to using the whole volume. Figure 13.3 shows the accuracy rates obtained by using the PCA coefficients extracted from each slice as features in combination with SVM classifiers. As expected, the SOIs correspond to those regions of the brain mainly affected by the AD, i. e., the posterior cingulate gyri and precunei, as well as the temporo-parietal region. A complete set of results is shown in Table 13.1, where the number of PCs m used for the data projection and the SOI are specified. Again we find that no more than $m = 3$ or $m = 4$ PCA coefficients are needed to classify satisfactorily the samples when the SOI is found and selected for classification. SVM with quadratic kernel reaches 96,7 % accuaracy when PCA is applied on the 10th slice along the axial axis, which was the most discriminant one. The combination of several PCA coefficients corresponding to SOIs along the three axes was

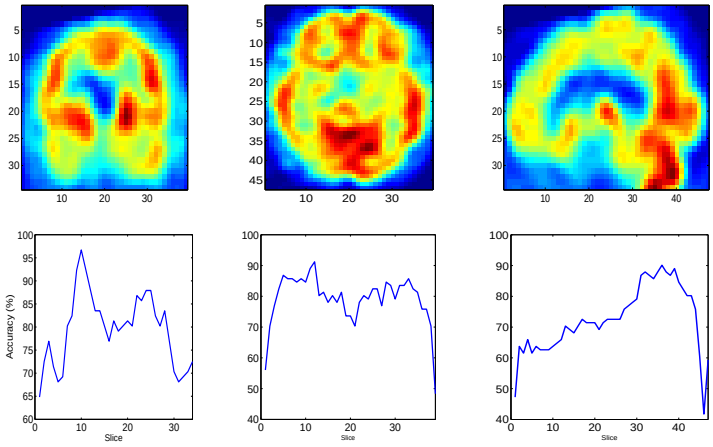


Figure 13.3: Slices of interest (SOIs) found when exploring the brain volume along the three directions and accuracy rates reached by each one by means of PCA transformation and SVM classifiers.

evaluated as well. However, it did not provide significant improvements.

SPECT

Axis Direction	m/ SOI	SVM Linear	SVM Quadratic	SVM Polynomial	SVM RBF
Axial	3/10	87.91 %	96.7 %	90.11 %	93.41 %
	3/11	87.91 %	92.31 %	90.11 %	93.41 %
	4/10	86.81 %	90.11 %	85.71 %	92.31 %
	4/11	87.91 %	90.11 %	89.01 %	89.01 %
Sagittal	1/35	89.01 %	86.81 %	86.81 %	87.91 %
	1/36	87.91 %	89.01 %	90.11 %	90.11 %
Coronal	3/11	89.01 %	91.21 %	84.62 %	86.81 %
	3/12	91.21 %	89.01 %	85.71 %	84.62 %
	4/11	89.01 %	87.91 %	81.32 %	85.71 %
	4/12	91.21 %	89.01 %	78.08 %	85.71 %

Table 13.1: Results obtained from the evaluation of SVM and slice-by-slice PCA application.

13.4.2. Results on PET *Cartuja* database

When PCA+LDA is applied on the PET *Cartuja* database we find very high accuracy results reaching for some particular number of PCs up to 100 % accuracy. This means that the *small sample size* problem is successfully solved yielding the maximum accuracy rate by using the minimun number

of features. Figure 13.4 shows the best accuracy curves obtained for some classifiers.

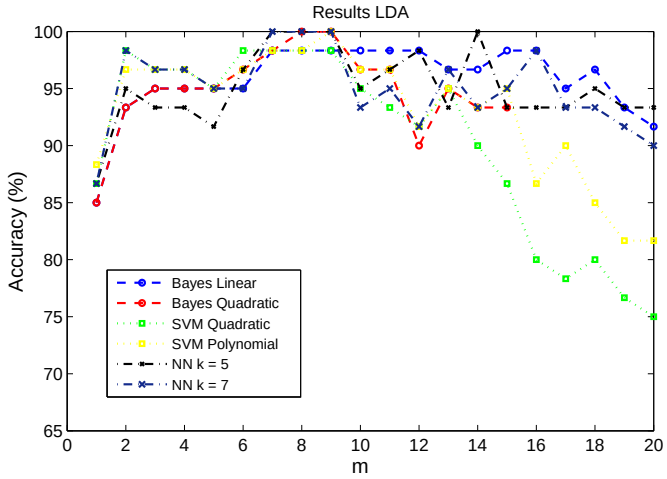


Figure 13.4: PET *Cartuja*: Accuracy results obtained when the number of PCA coefficients selected to be projected onto the LDA axis increases.

SVM and Bayes classifiers reach a peak accuracy of 100 % by selecting the first $m = 6, 7, 8$ PCA coefficients rearraanged by the FDR criterion. Note that for the other set of experiments, a significant number of times occurs that two samples are misclassified. The process of classifying a test sample using PCA+LDA involves two linear projections that yield to the final feature, where the subsets of vectors onto which the test is projected are computed without taking into account the test sample. Therefore, sometimes the resulting feature lies out of the range of values spanned by the samples of the same class.

13.4.3. Results on ADNI database

The variety of samples coming from different origins makes the ADNI database appealing for testing algorithms aiming at detecting AD. A large number of MCI patients are included in the database which entails having a wide variety of perfusion patterns that range from NORMAL to AD patients. Obviously, the classification task becomes more difficult since the transition between NORMAL and AD patients is less abrupt. The 219 selected patients are represented in terms of PCA plus LDA coefficients in Figure 13.6. The best description of this database in terms of PCA coefficients is found when

around $m = 30$ PCs are used to project the images, especially when MCI subjects are included. When only AD and NORMAL patients are considered, $m = 8$ coefficients is enough to describe the existing variabilities between these two classes. Results obtained from this database are presented in the next sections depending on which classes are desired to be separated.

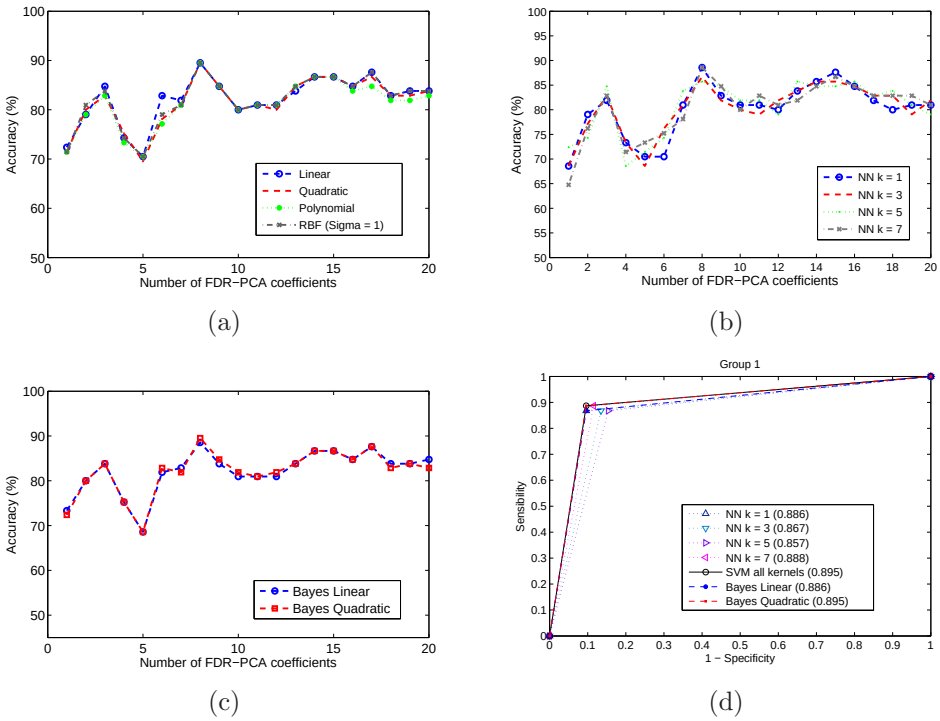


Figure 13.5: ADNI Group 1: Accuracy results obtained by evaluating the FDR-PCA+LDA feature extraction technique in combination with (a) SVM, (b) NN y (c) Bayesian classifiers. ROC curves for all the classifiers are represented in (d).

Group 1

Best results when classifying NORMAL controls versus AD patients are found when no voxel selection is performed, that is, the whole volumes are used to extract the PCA coefficients and the subsequent LDA features. An accuracy peak value is reached by all the tested classifiers when $m = 8$ PCA coefficients are used to project the images before applying LDA. The application of the FDR criterion on these coefficients improved the results, as shown in Figure 13.5 for all the classifiers, reaching an accuracy peak of 89,52 %.

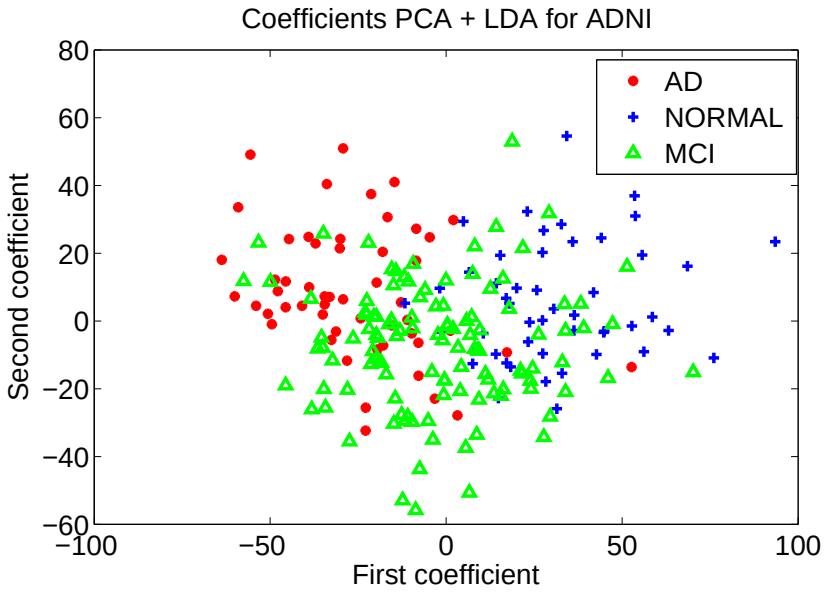
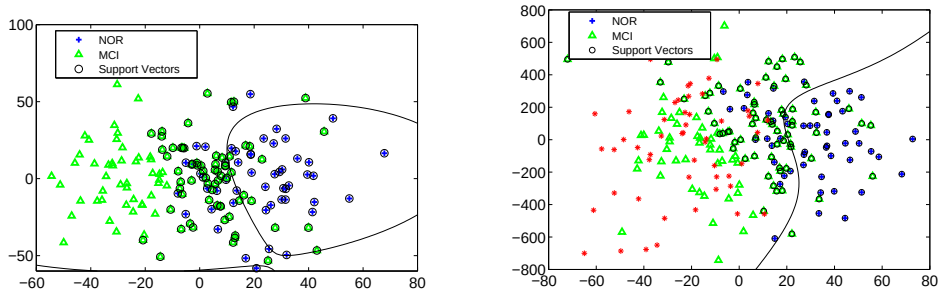


Figure 13.6: Descriptions of the ADNI database in terms of PCA+LDA coefficients. The original images were projected onto $m = 30$ PCs and subsequently onto $l = 2$ LDA axes projection.

Note that the obtained curves have similar forms for all the classifiers, which can be interpreted as a quality of robutness of the extracted feature. By using this feeature, the dependence on the classifier behaviour is removed. Best values of sensitivity and specificity obtained by SVM are represented as ROCs in Figure 13.5(d).

Group 2

The most difficult classification task concerning ADNI database is to distinguish between NORMAL and MCI patients, due to the wide range spanned by the features extracted from MCI patients (See Figure 13.6). When a conventional binary classification process is performed on these two datasets, accuracy rates do not exceed 74.1 % by using PCA+LDA features (i. e., one final feature) combined with SVM with linear or quadratic kernel. Since MCI can be considered a previous stage of AD [Minoshima et al., 1997; Silverman et al., 2001], we make profit of counting with an AD set of images to improve the classification results. Recall that LDA finds $l = c - 1$ axis projection, where c is the number of classes, so if we consider AD patients as a third different class, we obtain $l = 2$ features instead of one, as if we were deal-



(a) (b)

Figure 13.7: ADNI Group 2: decision line designed by an SVM classifier with polynomial kernel when (a) only MCI and NORMAL subjects are used in the training step and (b) when AD patients (depicted as red crosses) are included to design the classification rule.

ing with a multiclass problem. Once the $l = 2$ features are extracted, MCI and AD patients are labeled together as an unique class in the training step. Thus, the classifier will design a classification rule that gives strength to MCI patients, bringing them towards the AD vicinity. Once the classification rule has been established, only NORMAL and MCI samples are used to test the CAD system. Figure 13.7 shows the differences on the decision line designed by a polynomial kernel SVM classifier when solving the conventional binary classification problem and when the multiclass trick is used. This method improves classification results by increasing sensitivity values. Specificity values prove a poor performance of classifiers in recognizing NORMAL subjects, but recall group 2 is an imbalanced dataset since the number of MCI samples doubles the number of NORMAL subjects. To avoid a misinterpretation of the results, positive likelihood (PL) and negative likelihood (NL) ratios are computed as well. Table 13.2 shows the results obtained for all the experiments by using Bayesian, SVM and NN classifiers and compares the results of the conventional binary classification method with the multiclass proposal for group 2.

Group 2

	m	Bayes Linear	Bayes Quadratic
$l = 1$	24	68.07 % (68.42/67.31) % 2.093/0.469	66.27 % (64.91/69.23) % 2.109/0.507
$l = 2$	35	80.11 % (85.08/69.23) % 2.765/0.215	78.31 % (83.33/63.46) % 2.280/0.263

	m	SVM Linear	SVM Quadratic	SVM Polynomial	SVM RBF
$l = 1$	18	74.1 % (86.84/46.15) % 1.613/0.285	74.1 % (86.84/46.15) % 1.613/0.285	73.49 % (85.96/46.15) % 1.596/0.304	73.49 % (85.96/46.15) % 1.596/0.304
$l = 2$	30	81.33 % (97.37/46.15) % 1.808/0.057	77.71 % (91.23/46.15) % 1.694/0.190	77.71 % (91.23/46.15) % 1.694/0.190	77.71 % (91.22/48.08) % 1.757/0.183

	m	NN k = 1	NN k = 3	NN k = 5	NN k = 7
$l = 1$	33	71.08 % (82.46/46.15) % 1.531/0.380	72.29 % (83.33/48.08) % 1.605/0.347	71.08 % (82.46/46.15) % 1.531/0.380	70.48 % (81.58/46.15) % 1.515/0.399
$l = 2$	40	79.52 % (94.74/46.15) % 1.759/0.114	78.31 % (93.86/44.23) % 1.683/0.139	79.52 % (93.86/48.08) % 1.801/0.128	79.52 % (92.10/46.15) % 1.710/0.171

Table 13.2: Group 2: Accuracy, (sensitivity/specificity) and PL/NL values obtained by evaluating Bayes, SVM and NN in the classification task.

CHAPTER 14

Kernel Methods

Principal Component Analysis and Linear Discriminant Analysis methods have demonstrated their success in image classification and pattern recognition. The representations in these subspace methods are based on second order statistics of the image set, and do not address higher order statistical dependencies such as the relationships among three or more pixels. In this chapter, we investigate the use of Kernel Principal Component Analysis and Kernel Discriminant Analysis (KDA) for learning low dimensional representations for neurological images classification. While PCA and LDA methods aim to find projection directions based on second order correlation of samples, Kernel PCA and KDA methods provide generalizations which take higher order correlations into account. We compare the performance of kernel methods with classical algorithms.

14.1. Kernel PCA

The basic idea of the so called kernel-methods is to first preprocess the data by some non-linear mapping Φ and then to apply the same linear algorithm as before, but in the image space of Φ. The hope is that for a sufficiently nonlinear and appropriate Φ a linear decision in the image space of Φ will be enough (see Figure 2.8 for an illustration). More formally we apply the mapping Φ,

$$\begin{array}{rccc} \mathbf{\Phi}: & \mathbb{R} & \longrightarrow & \mathcal{F} \\ & \mathbf{x} & \longmapsto & \mathbf{y} = \mathbf{\Phi}(\mathbf{x}) \end{array} \tag{14.1}$$

to the data $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \in \mathcal{X}$ and now we consider our algorithm in $\mathcal{F}$ instead of $\mathcal{X}$.

Let us consider a set of N sample images $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ taking values in an n-dimensional image space, and assume that each image belongs to one of c classes $\{\omega_1, \omega_2, ..., \omega_c\}$. In kernel PCA, each vector $\mathbf{x}$ is projected from the input space $\mathbb{R}^n$, to a high dimensional feature space, $\mathbb{R}^f$, by a non-linear mapping function: $\mathbf{\Phi} : \mathbb{R}^n \rightarrow \mathbb{R}^f$. Note that the dimensionality of the feature space can be arbitrarily large. In $\mathbb{R}^f$, the eigenvalue problem is

$$\mathbf{C}^{\Phi}\mathbf{w}^{\Phi} = \lambda\mathbf{w}^{\Phi} \tag{14.2}$$

where $\mathbf{C}^{\Phi}$ is a covariance matrix. All solutions $\mathbf{w}^{\Phi}$ with $\lambda \neq 0$ lie in the space spanned by $\mathbf{\Phi}(\mathbf{x}_1), \mathbf{\Phi}(\mathbf{x}_2), ..., \mathbf{\Phi}(\mathbf{x}_N)$, and there exist coefficients α_i such that

$$\mathbf{w}^{\Phi} = \sum_{i=1}^{N} \lambda_i \mathbf{\Phi}(\mathbf{x}_i) \tag{14.3}$$

Denoting an $N \times N$ matrix K by

$$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x}_j) \tag{14.4}$$

the kernel PCA problem becomes [Schölkopf et al., 1998]

$$N\lambda K\boldsymbol{\alpha} = K^2\boldsymbol{\alpha} \equiv N\lambda\boldsymbol{\alpha} = K\boldsymbol{\alpha} \tag{14.5}$$

where $\boldsymbol{\alpha}$ denotes a column vector with entries $\alpha_1, ..., \alpha_N$. The above derivation assumes that all the projected samples $\mathbf{\Phi}(\mathbf{x})$ are centered in $\mathbb{R}^f$. See Schölkopf et al. [1998] for a method to center the vectors $\mathbf{\Phi}(\mathbf{x})$ in $\mathbb{R}^f$.

To get non-linear forms of PCA we simply choose a non-linear kernel. The two commonly used families of kernels are polynomial kernels and radial basis functions (RBF), defined as follows:

- Polynomial:
$$K(\mathbf{x}_i, \mathbf{x}_j) = [\gamma(\mathbf{x}_i \cdot \mathbf{x}_j) + c]^d \tag{14.6}$$
- Radial Basis Functions:
$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2) \tag{14.7}$$

Note that classical PCA is a special case of kernel PCA with first order polynomial kernel. The case of $d = 2$ gives a quadratic kernel. Moreover kernel PCA is a generalization of PCA in the respect that it is performing PCA in feature spaces of arbitrarily large (possibly infinite) dimension.

We can now project the vectors in $\mathbb{R}^f$ to a lower dimensional space spanned by the eigenvectors $\mathbf{w}^\Phi$. Let $\mathbf{x}$ be a test sample whose projection is $\mathbf{\Phi}(\mathbf{x})$ in $\mathbb{R}^f$, then the projection of $\mathbf{\Phi}(\mathbf{x})$ onto the eigenvectors $\mathbf{w}^\Phi$ is the non-linear principal components corresponding to $\mathbf{\Phi}$

$$\mathbf{w}^\Phi \cdot \mathbf{\Phi}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i (\mathbf{\Phi}(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x})) = \sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}) \tag{14.8}$$

In other words, we can extract the first $m, (1 \leq m \leq N)$ non-linear principal components (i.e., eigenvectors $\mathbf{w}_i^\Phi, i = 1, ..., m$) using the kernel function without the expensive operation that explicitly projects samples to a high dimensional space $\mathbb{R}^f$. Therefore, the initial image vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ become reduced to m-dimensional kernel PCA coefficients vectors $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$ by means of the projection

$$\mathbf{Y} = \mathbf{W}_{kPCA}^T \mathbf{X} \tag{14.9}$$

where $\mathbf{W}_{kPCA} = [\mathbf{w}_1^\Phi \mathbf{w}_2^\Phi ... \mathbf{w}_m^\Phi]$. The first m kernel components correspond to the first m non-increasing eigenvalues of equation 14.5.

14.2. Kernel Discriminant Analysis

Let $\mathbf{\Phi}$ be a non-linear mapping to some feature space $\mathcal{F}$. To find the linear discriminant in $\mathcal{F}$ we need to maximize

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b^\Phi \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^\Phi \mathbf{w}} \tag{14.10}$$

where now $\mathbf{w} \in \mathcal{F}$ and $\mathbf{S}_b^\Phi$ y $\mathbf{S}_w^\Phi$ are the corresponding within-class and between-class scatter matrices in $\mathcal{F}$, i. e.

$$\mathbf{S}_b^\Phi = (\boldsymbol{\mu}_1^\Phi - \boldsymbol{\mu}_2^\Phi)(\boldsymbol{\mu}_1^\Phi - \boldsymbol{\mu}_2^\Phi)^T \tag{14.11}$$

$$\mathbf{S}_w^\Phi = \sum_{i=1,2} \sum_{\mathbf{x} \in \omega_i} (\Phi(\mathbf{x}) - \boldsymbol{\mu}_i^\Phi)(\Phi(\mathbf{x}) - \boldsymbol{\mu}_i^\Phi)^T \tag{14.12}$$

with $\boldsymbol{\mu}_i^\Phi = \frac{1}{N_i} \sum_{j=1}^{N_i} \Phi(\mathbf{x}_j^i)$, being N_i the number of samples in class ω_i, $i = 1, 2$. If $\mathcal{F}$ is very high – or even infinitely dimensional this will be impossible to solve directly. To overcome this limitation we use the same trick as in Kernel PCA [Schölkopf et al., 1998] or Support Vector Machines. Instead of mapping the data explicitly we seek a formulation of the algorithm which uses only dot-products $(\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j))$ of the training patterns. This can be achieved using Mercer kernels (e.g. [Saitoh, 1988]). To find Fisher's discriminant in the feature space $\mathcal{F}$, we first need a formulation of 14.10 in terms of only dot products of input patterns which we then replace by some kernel function. From the theory of reproducing kernels we know that any solution $\mathbf{w} \in \mathcal{F}$ must lie in the span of all training samples in $\mathcal{F}$. Therefore we can find an expansion for $\mathbf{w}$ of the form

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{\Phi}(\mathbf{x}_i) \tag{14.13}$$

Using the expansion 14.13 and the definition of $\boldsymbol{\mu}_i^\Phi$ we write

$$\mathbf{w}^T \boldsymbol{\mu}_i^\Phi = \frac{1}{N_i} \sum_{j=1}^{c} \sum_{k=1}^{N_i} \alpha_j k(\mathbf{x}_j, \mathbf{x}_i) = \boldsymbol{\alpha}^T \mathbf{M}_i \tag{14.14}$$

where we defined $(\mathbf{M}_i)_j = \frac{1}{N_i}\sum_{k=1}^{N_i} k(\mathbf{x}_j, \mathbf{x}_k^i)$ and replaced the dot products by the kernel function. Now consider the numerator of 14.10. By using the definition of $\mathbf{S}_b^\Phi$ and 14.18 it can be rewritten as

$$\mathbf{w}^T \mathbf{S}_b^\Phi \mathbf{w} = \boldsymbol{\alpha}^T M \boldsymbol{\alpha} \tag{14.15}$$

where $M := (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T$. Considering the denominator, using 14.13, the definition of $\mathbf{S}_w^\Phi$ and a similar transformation as in 14.15 we find:

$$\mathbf{w}^T \mathbf{S}_w^\Phi \mathbf{w} = \boldsymbol{\alpha}^T N \boldsymbol{\alpha} \tag{14.16}$$

where we set $N := \sum_{i=1,2} K_j(\mathbf{I} - \mathbf{1}_{l_j})K_j^T$, K_j is a $l \times l_j$ matrix with $(K_j)_{nm} := k(\mathbf{x}_n, \mathbf{x}_n^j)$ (this is the kernel matrix for class j), $\mathbf{I}$ is the identity and $\mathbf{1}_{l_j}$ the matrix with all entries $1/l_j$.

Combining 14.15 and 14.16 we can find Fisher's linear discriminant in $\mathcal{F}$ by maximizing

$$J(\alpha) = \frac{\boldsymbol{\alpha}^T M \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T N \boldsymbol{\alpha}} \tag{14.17}$$

This problem can be solved (analogously to the algorithm in the input space) by finding the leading eigenvector of $N^{-1}M$. We will call this approach (non-linear) Kernel Discriminant Analysis (KDA). The projection of a new pattern $\mathbf{x}$ onto $\mathbf{w}$ is given by

$$\mathbf{w} \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{14.18}$$

14.3. Experiments

One disadvantage of kernel methods is that they are computationally more costly than linear techniques. This motivate us to reduce the input space by simple techniques. The dimension of the volume representing each subject brain was reduced to $34 \times 47 \times 39$ by decimating the original 3D volume by a $2 \times 2 \times 2$ factor. After that, as proposed in Górriz et al. [2008], a mask is applied so that voxels whose mean intensity value averaged over

all images is lower than the half of the maximum mean intensity value are rejected. This will make kernel algorithms to be significantly more efficient since the preprocessing steps retain the main amount of useful information in the images.

KDA feature extraction technique was evaluated on the SPECT and PET *Cartuja* images. The resulting features yielded improvements in the final results with respect to linear projections for the SPECT images and matched the 100 % accuracy peak for PET *Cartuja* images. Best performances were found when polynomial kernels were applied to map the data ($d = 2$ and $d = 3$ in equation 8.6, whereas RBF kernel did not provide a good description of the data, and therefore results obtained with this kernel trasnformation are not shown. Regarding the classifiers, Bayes and SVM outperformed NN but only SVM yielded an improvement over the classical linear techniques PCA+LDA. In order to make easier a comparative with linear techniques, only the results obtained by means of SVM classifiers are presented.

14.4. Results

The number of kernel PCA coefficients selected to be projected onto the LDA axes has been evaluated from $m = 1$ to $m = N-1$, being N the number of samples in the database. These coefficients undergo the LDA transformation onto the first l projecting axis obtained by solving the maximization of Eq. . Regarding the values of l, recall from equation 13.8 that $\mathbf{S}_w^{-1}\mathbf{S}_b$ has $c-1$ non generalized autovectors, being c the number of classes. Under a bayesian framework, the use of the corresponding $l = c-1$ LDA coefficients is enough to classify a test sample in one of c classes, since the Bayes error in the $(c-1)$-dimensional feature space is identical to the Bayes error in the original n-dimensional space [Fukunaga, 1990]. In our binary classification problem, the first KDA coefficient is clearly the most discriminant one. However, when the second, third and fourth KDA coefficients are added to the feature vector used to train an SVM classifier, the accuracy rates improve slightly [Huang et al., 2002]. This fact proves that SVM is able to exploit the class information contained in these coefficients and makes use of it to design the separation hyperplane. This fact can be observed in Figure 14.1. The designed hyperplanes are predominantly vertical, i.e., there exists a strong influence of the first KDA coefficient. However, second and third components also contribute giving different shapes to the separation hyperplanes.

Table 14.1 shows the results obtained by KDA and compares them with

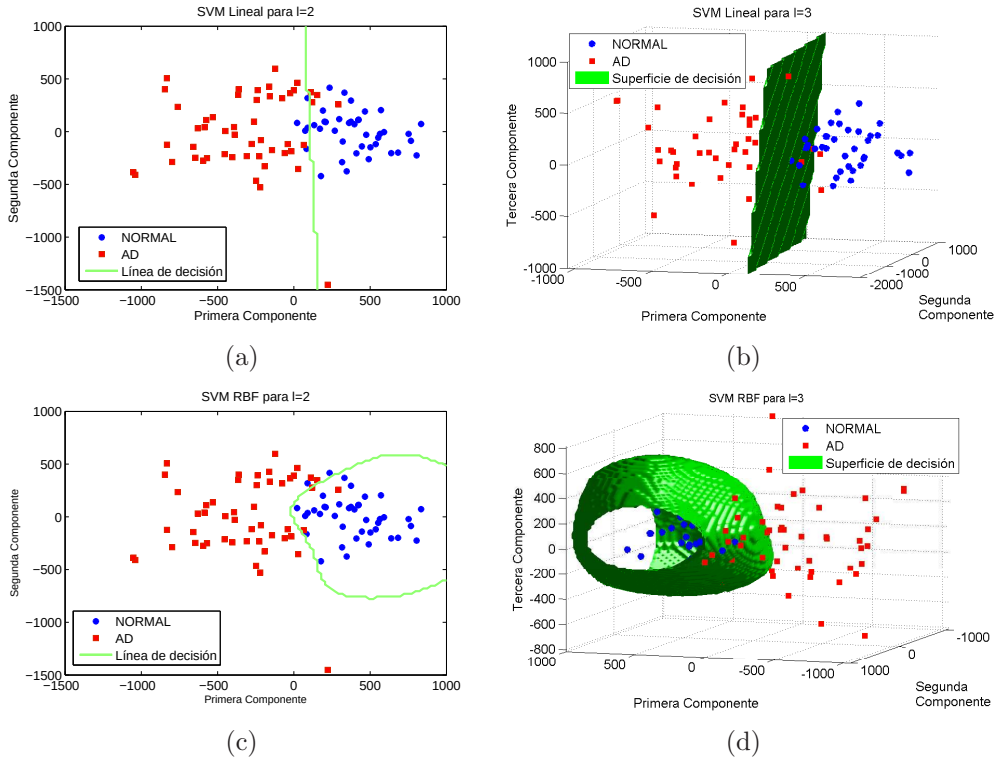


Figure 14.1: Decision lines and surfaces designed by a linear SVM (up) and REB (down) when $l = 2$ (left) and $l = 3$ (right) LDA features are used. The first LDA features is the most discriminant one for all cases. However, for non-linear classifiers, the second and third features are not totally useless, and they can contribute to the surface to fit the decision surfaces better to the class distributions.

classical (lineal) PCA+LDA. Moreover, it can be seen that using $l = 3$ coefficients improves the accuracy result with respect to $l = 1$. Particularly, for SPECT images, the use of RBF SVM increases the accuracy from 89,01 % to 93,41 % and from 92,31 % to 95,6 % for quadratic and polynomial ($d = 3$) kernels used in the feature extraction step, respectively. These are the highest results for the SPECT database obtained throughout all this work.

	Projection PCA	l	SVM Linear	SVM Quadratic	SVM Polynomial	SVM RBF
SPECT	Linear	1	91.21 %	89.01 %	90.11 %	89.01 %
	Polyn. $d = 2$		87.91 %	89.01 %	87.91 %	89.01 %
	Polyn. $d = 3$		90.11 %	89.01 %	86.81 %	89.01 %
	Polyn. $d = 2$	3	87.91 %	90.11 %	90.11 %	93.41 %
	Polyn. $d = 3$		**92.31** %	**92.31** %	**93.41** %	**95.60 %**
PET	Linear	1	100 %	96.67 %	98.33 %	95 %
	Polyn. $d = 2$		98.33 %	98.33 %	100 %	98.33 %
	Polyn. $d = 3$		98.33 %	98.33 %	100 %	98.33 %
	Polyn. $d = 2$	3	98.33 %	98.33 %	100 %	98.33 %
	Polyn. $d = 3$		98.33 %	98.33 %	100 %	98.33 %

Table 14.1: Acuracy results obtained by means of kernel methods for SPECT and PET. Comparative with classical techniques.

CHAPTER 15

Discussion and Conclusions

This chapter first gathers the conclusions of the presented work and a discussion on the different methods, and summarizes briefly the main advantages and disadvantages of each proposed CAD system. Secondly, some ideas for the continuation of this work are proposed as future research lines.

15.1. Discussion and conclusions

In this work, a set of complete and independent CAD systems has been developed and implemented. They all achieve a satisfactory classification performance, being capable of distinguishing between normal and AD patients successfully.

Feature extraction technique based on ROIs extraction is presented as an improvement of the baseline VAF approach. A selection of the most interesting voxels is performed by means of an exhaustive exploration of the brain and learning machine based techniques. This method is the only one of the presented methods that considers the spatial proximity property of the voxels affected by a disease, which is the natural way of image exploration when this is performed by clinicians. On the other hand, the classifiers aggregation provides high robustness to the final decision process. The main drawback of this approach lies in the high computational cost and memory requirements.

In comparison to the direct application methods, data transformation-based techniques search for discriminant information underlying in other subspaces. PCA+LDA in combination with the FDR selection criterion has proved to be a suitable feature extraction technique for neurological image classification reaching up to 100 % and 91,21 % accuracy rates for PET and SPECT data respectively, by using only one final feature, which solves the *small sample size* problem in the best possible way. Non-linear feature extraction techniques like kernel PCA and kernel LDA improve these results yielding up to 95,6 % for SPECT and matching the 100 % for PET, showing to be robust features for classification.

Regarding the classification methods, SVM and Bayes' classifier show in most cases better performance than NN for the same evaluated features. SVM requires higher computational resources but in contrast is able to design non-linear decision surfaces that have proved to be more suitable in some cases for detecting AD patterns. In all cases,

15.2. Future work

As future lines of the research developed throughout this work, we present the following proposals:

- The use of image segmentation or clustering methods to find the ROIs.

Once the voxels are put into groups that share some common properties of gray levels and proximity, the discriminant power of each group may be analyzed in order to hold only the most discriminant voxels.

- Analysis of the images in other transformed spaces like frequency domain, where other relationships between AD and normal patients might be found out by exploring the modes that compose the images.
- Multiclass classification that allows to categorize the images in more than two classes and therefore to distinguish different stages of the progressive diseases.

Referencias

Adler, R., 1981. The Geometry of random fields. Wiley, New York.

Alexander, G. E., Pietrini, P., Rapoport, S. I., Reiman, E. M., 2002. Longitudinal PET evaluation of cerebral metabolic decline in dementia: A potential outcome measure in Alzheimer's disease treatment studies. The American Journal of Psychiatry 159 (5), 783–745.

Álvarez, I., Górriz, J. M., Ramírez, J., Salas-Gonzalez, D., López, M., Puntonet, C. G., Segovia, F., 2009a. Alzheimer's diagnosis using eigenbrains and support vector machines. IET Electronics Letters 45(7), 342–343.

Álvarez, I., Górriz, J. M., Ramírez, J., Salas-Gonzalez, D., López, M., Puntonet, C. G., Segovia, F., May 2009b. Independent component analysis of SPECT images to assist the Alzheimer's disease diagnosis. In: Advances in Neural Networks Research, ISNN 2009. Advances in Intelligent and Soft Computing. Wuhan (China).

Álvarez, I., Górriz, J., Ramírez, J., Salas-Gonzalez, D., López, M., Segovia, F., Padilla, P., Puntonet, C., 2010. Projecting independent components of SPECT images for computer aided diagnosis of Alzheimer's disease. (Accepted in) Pattern Recognition Letters.

Álvarez, I., López, M., Górriz, J. M., Ramírez, J., Puntonet, C. G., Salas-Gonzalez, D., November 2008. Automatic classification system for the diag-

nosis of Alzheimer disease using component-based SVM aggregations. In: ICONIP 2008 Proceedings, Lecture Notes in Computer Science. Auckland (New Zealand).

Alzheimer, A., 1907. Über eine eigenartige erkrankung der hinrinde. Allgemeine Zeitschrift für Psychiatrie und Psychischgerichtliche Medizin LXIV, 146–148.

Americana de Psiquiatría, A., 1994. Diagnostic and Statistical manual of mental disorders. American Psychiatris Association.

Ashburner, J., Friston, K. J., 1999. Nonlinear spatial normalization using basis functions. Human Brain Mapping 7 (4), 254–66.

Belhumeour, P. N., Hespanha, J. P., Kriegman, D. J., 1997. Eigenfaces vs. Fisherfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 711–720.

Bennett, K. P., 1999. Advances in Kernel Methods - Support Vector Learning. MIT Press, Ch. Combining Support Vector and Mathematical Programming Methods for Classification, pp. 307–326.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Bobinski, M., de Leon, M., Convit, A., Santi, S. D., Weigel, J., Tarshish, C., Louis, L. S., Wisniewski, H., 1999. MRI of entorhinal cortex in mild Alzheimer's disease. Lancet 353 (9146), 38–40.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathologica 82 (4), 239–259.

Braak, H., Braak, E., 1997. Diagnostic criteria for neuropathologic assessment of Alzheimer's disease. Neurobiology and Aging 18 (4), S85–S88.

Breiman, L., 1999. Pasting small votes for classification in large databases and on-line. Machine Learning 36, 85–103.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and regression trees, 1st Edition. Chapman & Hall.

Brunelli, R., Poggio, T., 1993. Face recognition: Features versus templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (10), 1042–1052.

Bruyant, P. P., 2002. Analytic and iterative reconstruction algorithms in SPECT. The Journal of Nuclear Medicine 43 (10), 1343–1358.

Burges, C. J. C., 1998. A tutorial on Support Vector Machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Carr, D. B., Goate, A., Phil, D., Morris, J. C., Sep. 1997. Current concepts in the pathogenesis of Alzheimer's disease. The American Journal of Medicine 103 (3A), 3S–10S, PMID: 9344401.

Cawley, G., Talbot, N., 2003. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition 36 (11), 2585–2592.

Chase, T. N., Foster, N. L., Fedio, P., Brooks, R., Mansi, L., Chiro, G. D., 1984. Regional cortical dysfunction in Alzheimer's disease as determined by positron emission tomography. Annals of Neurology 15 Suppl, S170–4, PMID: 6611118.

Chaves, R., Ramírez, J., Górriz, J., López, M., Salas-Gonzalez, D., Álvarez, I., Segovia, F., 2009. SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. Neuroscience Letters 461, 293–297.

Chen, K., Reiman, E. M., Huan, Z., Caselli, R. J., Bandy, D., Ayutyanont, N., Alexander, G. E., 2009. Linking functional and structural brain images with multivariate network analyses: A novel application of the partial least square method. NeuroImage 47 (2), 602–610.

Chornoboy, E. S., Chen, C. J., Miller, M. I., Miller, T. R., Snyder, D. L., 1990. An evaluation of maximum likelihood reconstruction for SPECT. IEEE Transactions on Medical Imaging 9 (1), 99–110.

Clarkson, P., Moreno, P. J., 1999. On the use of support vector machines for phonetic classification. In: Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing. Vol. 2. pp. 585–588.

Claus, J. J., van Harskamp, F., Breteler, M. M. B., Krenning, E. P., de Koning abd J. M. van der Cammen, I., Hofman, A., Hasan, D., 1994. The diagnostic value of SPECT with TC 99m HMPAO in Alzheimer's disease. a population-based study. Neurology 44 (3), 454–461.

Cover, T. M., 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Transactions on Electronic Computers 14 (3), 326–334.

Cummings, J. L., Vinters, H. V., Cole, G. M., Khachaturian, Z. S., 1998. Alzheimer's disease: etiologies, pathophysiology, cognitive reserve, and treatment opportunities. Neurology 51 (suppl. 1), S2–S17.

Devijver, P. A., Kittler, J., 1982. Pattern Recognition: A Statistical Approach, first edition Edition. Prentice Hall.

Drzezga, A., Lautenschlager, N., Siebner, H., Riemenschneider, M., Willoch, F., Minoshima, S., Schwaiger, M., Kurz, A., 2003. Cerebral metabolic changes accompanying conversion of Mild Cognitive Impairment into Alzheimer's disease: a PET follow-up study. European Journal of Nuclear Medicine and Molecular Imaging 30 (8), 1104–1113.

Duara, R., Grady, C., Haxby, J., Sundaram, M., Cutler, N. R., Heston, L., Moore, A., Schlageter, N., Larson, S., Rapoport, S. I., Jul. 1986. Positron emission tomography in Alzheimer's disease. Neurology 36 (7), 879–887.

Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Duin, R. P. W., 2000. Classifiers in almost empty spaces. In: Proceedings 15th International Conference on Pattern Recognition. Vol. 2. IEEE, pp. 1–7.

Enqing, D., Guizhong, L., Yatong, Z., Xiaodi, Z., 2002a. Applying support vector machines to voice activity detection. In: 6th International Conference on Signal Processing. Vol. 2. pp. 1124–1127.

Enqing, D., Heming, Z., Yongli, L., 2002b. Low bit and variable rate speech coding using local cosine transform. In: Proc. of the 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCON '02). Vol. 1. pp. 423–426.

Fan, Y., Batmanghelich, N., Clark, C., Davatzikos, C., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39, 1731–1743.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Fisher, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

Foster, N. L., Chase, T. N., Fedio, P., Patronas, N. J., Brooks, R. A., Chiro, G. D., Aug. 1983. Alzheimer's disease: Focal cortical changes shown by positron emission tomography. Neurology 33 (8), 961.

Foster, N. L., Chase, T. N., Mansi, L., Brooks, R., Fedio, P., Patronas, N. J., Chiro, G. D., Dec. 1984. Cortical abnormalities in Alzheimer's disease. Annals of Neurology 16 (6), 649–54, PMID: 6335378.

Frackowiak, R. S. J., Ashburner, J. T., Penny, W. D., Zeki, S., December 2003. Human Brain Function, Second Edition. Academic Press.

Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E., Penny, W. D., 2007. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, New York.

Fung, G., Stoeckel, J., 2007. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. Knowledge and Information Systems 11 (2), 243–258.

Ganapathiraju, A., Hamaker, J. E., Picone, J., 2004. Applications of support vector machines to speech recognition. IEEE Transactions on Signal Processing 52 (8), 2348–2355.

Goethals, I., van deWiele, C., Slosman, D., Dierckx, R., 2002. Brain SPECT perfusion in early Alzheimer disease: where to look? European Journal of Nuclear Medicine 29 (8), 975–978.

Gool, W. A. V., Walstra, G. J., Teunisse, S., der Zant, F. M. V., Weintein, H. C., Royen, E. A. V., 1995. Diagnosing Alzheimer's disease in elderly, mildly demented patients: the impact of routine single photon emission computed tomography. Journal of Neurology 242 (6), 401–405.

Górriz, J. M., Ramírez, J., Lassl, A., Salas-Gonzalez, D., Lang, E. W., Puntonet, C. G., Álvarez, I., López, M., Gómez-Río, M., October 2008. Automatic computer aided diagnosis tool using component-based SVM. In: IEEE Nuclear Science Symposium Conference Record, Medical Imaging Conference. pp. 4392–4395, Dresden (Germany).

Górriz, J. M., Puntonet, C. G., Salmerón, M., de la Rosa, J. J. G., 2004. A new model for time-series forecasting using radial basis functions and exogenous data. Neural Computing & Applications 13 (2), 101–111.

Górriz, J. M., Ramírez, J., Lassl, A., Álvarez, I., Segovia, F., Salas, D., López, M., June 2009. Classification of SPECT images using clustering techniques revisited. In: IWINAC 2009 Proceedings. Lecture Notes in Computer Science. pp. 168–178, santiago de Compostela (Spain).

Haynes, J., Rees, G., 2006. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7, 523–534.

Heisele, B., Ho, P., Poggio, T., 2001a. Face recognition with support vector machines: global versus component-based approach. In: Proc. 8th International Conference on Computer Vision. Vol. 2. pp. 688–694.

Heisele, B., Poggio, T., Pontil, M., 2000. Face detection in still gray images. In: A.I. memo 1687, Center for Biological and Computational Learning. Massachusetts Institute of Technology, Cambridge, MA.

Heisele, B., Serre, T., Pontil, M., Poggio, T., 2001b. Component-based face detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. Vol. 1. Hawaii, pp. 657–662.

Hellman, R. S., Tikofsky, R. S., Collier, B. D., Hoffmann, R. G., Palmer, D. W., Glatt, S., Antuono, P. G., Isitman, A. T., Papke, R. A., 1989. Alzheimer disease: quantitative analysis of I-123-iodoamphetamine SPECT brain imaging. Radiology 172, 183–188.

Herholz, K., Schopphoff, H., Schmidt, M., Mielke, R., Eschner, W., Scheidhauer, K., Schicha, H., Heiss, W., Ebmeier, K., 2002. Direct comparison of spatially normalized PET and SPECT scans in Alzheimer's disease. Journal of Nuclear Medicine 43 (1), 21–26.

Higdon, R., Foster, N. L., Koeppe, R. A., DeCarli, C. S., Jagust, W. J., Clark, C. M., Barbas, N. R., Arnold, S. E., Turner, R. S., Heidebrink, J. L., Minoshima, S., 2004. A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging. Statistics in Medicine 23, 315–326.

Hoffman, E. J., Phelps, M. E., 1986. Positron Emission Tomography and Autoradiography: Principles and Applications for the Brain and Heart. Ch. Positron emission tomography: principles and quantitation, pp. 237–286.

Hoffman, J. M., Welsh-Bohmer, K. A., Hanson, M., 2000. FDG PET imaging in patients with pathologically verified dementia. Journal of Nuclear Medicine 41 (11), 1920–1928.

Holman, B. L., Johnson, K. A., Gerada, B., Carvalho, P. A., Satlin, A., 1992. The scintigraphic appearance of Alzheimer's disease: A prospective study using Technetium-99m-HMPAO SPECT. Journal of Nuclear Medicine 33 (2), 181–185.

Huang, R., Liu, Q., Lu, H., Ma, S., 2002. Solving the *small sample size* problem of LDA. In: Proceedings of the 16th International Conference on Pattern Recognition. Vol. 3. pp. 29–32.

Huber, P. J., 1981. Robust Statistics. Wiley.

Hudson, H. M., Larkin, R. S., 1994. Accelerated image reconstruction using ordered subsets of projection data. IEEE Transactions on Medical Imaging 13 (4), 601–609.

Hyun-Chul, K., Shaoning, P., Hong-Mo, J., Kim, D., Bang, S. Y., 2003. Constructing support vector machine ensemble. Pattern Recognition 36 (12), 2757–2767.

Ibañez, V., Pietrini, P., Alexander, G. E., Furey, M. L., Teichberg, D., Rajapakse, J. C., Rapoport, S. I., Schapiro, M. B., Horwitz, B., Jun. 1998. Regional glucose metabolic abnormalities are not the result of atrophy in Alzheimer's disease. Neurology 50 (6), 1585–93, PMID: 9633698.

Ishii, K., Kono, A. K., Sasaki, H., Miyamoto, N., Fukuda, T., Sakamoto, S., Mori, E., 2006. Fully automatic diagnostic system for early- and late-onset mild Alzheimer's disease using FDG PET and 3D-SSP. European Journal of Nuclear Medicine and Molecular Imaging 33 (5), 575–583.

Ishii, K., Sasaki, M., Yamaji, S., Sakamoto, S., Hitagaki, H., Mori, E., 1997. Demonstration of decrease posterior cingulate perfusion in midl Alzheimer's disease by means of H215O positron emission tomography. European Journal of Nuclear Medicine and Molecular Imaging 26 (6), 670–673.

Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In: Lecture Notes in Computer Science. Vol. 1398. pp. 137–142.

Jobst, K. A., Barneston, L. P., Shepstone, B. J., 1998. Accurate prediction of histologically confirmed Alzheimer's disease and the differential diagnosis of dementia: the use of NINCDS-ADRDA and DSM-III-R criteria,

SPECT, X-ray CT, and APO E4 medial temporal lobe dementias. the Oxford Project to Investigate Memory and Aging. International Psychogeriatrics 10 (3), 271–302.

Jobst, K. A., Smith, A. D., Barker, C. S., Wear, A., King, E. M., Smith, A., Anslow, P. A., Molyneux, A. J., Shepstone, B. J., Soper, N., 1992. Association of atrophy of the medial temporal lobe with reduced blood flow in the posterior parietotemporal cortex in patients with a clinical and pathological diagnosis of Alzheimer's disease. Journal of Neurology Neurosurgery and Psychiatry 55 (3), 190–194.

John, G. H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In: International Conference on Machine Learning. pp. 121–129.

Johnson, K. A., Kijewski, M. F., Becker, J. A., Garada, B., Satlin, A., Holman, B. L., 1993. Quantitative brain SPECT in Alzheimer's disease and normal aging. Journal of Nuclear Medicine 34 (11), 2044–2048.

Jolliffe, I., 1986. Principal Component Analysis. Springer Verlag, New York.

Kalatzis, I., Pappas, D., Piliouras, N., Cavouras, D., 2003. Support vector machines based analysis of brain SPECT images for determining cerebral abnormalities in asymptomatic diabetic patients. Medical Informatics and the Internet in Medicine 28 (3), 221–230.

Kim, K. I., Jung, K., Park, S. H., Kim, H. J., 2002. Support vector machines for texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (11), 1542–1550.

Kirby, M., Sirovich, L., 1990. Application of the KL procedure for the characterization of human faces. IEEE Trans. on Pattern Analysis and Machine Intelligence 12 (1), 103–108.

Kogure, D., Matsuda, H., Ohnishi, T., Asada, T., Uno, M., Kunihiro, T., Nakano, S., Takasaki, M., 2000. Longitudinal evaluation of early Alzheimer disease using brain perfusion SPECT. The Journal of Nuclear Medicine 41 (7), 1155–1162.

Kohavi, R., John, G. H., 1995. Wrappers for feature subset selection. AIJ special issue on relevance, 273–324.

Langbaum, J. B., Vhen, K., Lee, W., Reschke, C., Bandy, D., Fleisher, A. S., Alexander, G. E., Foster, N. L., Weiner, M. W., Koeppe, R. A., Jagust,

W. J., Reiman, E. M., 2009. Categorical and correlation analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's disease neuroimaging initiative (ADNI). Neuroimage 45 (4), 1107–1116.

Lange, K., Carson, R., 1984. EM reconstruction for emission and transmission tomography. Journal of Computer Assisted Tomography 8, 306–312.

Lawrence, N., Schölkopf, B., 2001. Estimating a kernel fisher discriminant in the presence of label noise. In: Proceedings of the 18th International Conference on Machine Learning. pp. 306–313.

Leon, M. J. D., Convit, A., Wolf, O. T., Tarshish, C. Y., DeSanti, S., Rusinek, H., Tsui, W., Kandil, E., Scherer, A. J., Roche, A., Imossi, A., Thorn, E., Bobinski, M., Caraos, C., Lesbre, P., Schlyer, D., Poirier, J., Reisberg, B., Fowler, J., Sep. 2001. Prediction of cognitive decline in normal elderly subjects with 2-(18)Ffluoro-2-deoxy-D-glucose/poitron-emission tomography (FDG/PET). Proceedings of the National Academy of Sciences of the United States of America 98 (19), 10966–10971, PMID: 11526211.

Leon, M. J. D., Ferris, S. H., George, A. E., Reisberg, B., Christman, D. R., Kricheff, I. I., Wolf, A. P., Sep. 1983. Computed tomography and positron emission transaxial tomography evaluations of normal aging and Alzheimer's disease. Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism 3 (3), 391–4, PMID: 6603463.

Leung, T. K., Burl, M. C., Perona., P., 1995. Finding faces in cluttered scenes using random labeled graph matching. In: International Conference on Computer Vision. pp. 637–644.

López, M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Gómez-Río, M., June 2009a. Support vector machines and neural networks for the Alzheimer's disease diagnosis using PCA. In: IWINAC 2009 Proceedings. Lecture Notes in Computer Science. Santiago de Compostela (Spain).

López, M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Puntonet, C. G., June 2009b. Automatic system for Alzheimer's disease diagnosis using eigenbrains and Bayesian classification rules. In: IWANN 2009 Proceedings, Lecture Notes in Computer Science. Salamanca (Spain).

López, M., Ramírez, J., Górriz, J. M., Salas-Gonzalez, D., Álvarez, I., Segovia, F., Puntonet, C. G., 2009c. Automatic tool for the Alzheimer's disease diagnosis using PCA and Bayesian classification rules. IET Electronics Letters 45(8), 389–391.

López, M. M., Ramírez, J., Górriz, J. M., Álvarez, I., Salas-Gonzalez, D., Chaves, R., 2009d. SVM-based CAD system for early detection of the Alzheimer's disease using kernel PCA and LDA. Neuroscience Letters 464, 233–238.

López, M., Ramirez, J., Gorriz, J., Salas-Gonzalez, D., Alvarez, I., Segovia, F., Chaves, R., October 2009a. Multivariate approaches for Alzheimer's disease diagnosis using bayesian classifiers. In: Proceedings of the 2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC). Orlando (Fl.).

López, M., Ramirez, J., Gorriz, J., Salas-Gonzalez, D., Alvarez, I., Segovia, F., Chaves, R., October 2009b. Neurological image classification for the Alzheimer's disease diagnosis using kernel PCA and support vector machines. In: Proceedings of the 2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC). Orlando (Fl.).

Markiewicz, P., Matthews, J., Declerck, J., Herholz, K., 2009. Robustness of multivariate image analysis assessed by resampling techniques and applied to FDG-PET scans of patients with Alzheimer's disease. NeuroImage 46, 472–485.

McCulloch, W. S., Pitts, W., 1943. A logical calculus of ideas immanent in nervous activity. Bull. Mathematical Biophysics 5, 115–133.

McGeer, E. G., Peppard, R. P., McGeer, P. L., Tuokko, H., Crockett, D., Parks, R., Akiyama, H., Calne, D. B., Beattie, B. L., Harrop, R., Feb. 1990. 18Fluorodeoxyglucose positron emission tomography studies in presumed Alzheimer cases, including 13 serial scans. The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques 17 (1), 1–11, PMID: 2311010.

McKhann, E., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E., 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. Neurology 34, 939–944.

McMurdo, M. E., Grant, D. J., Kennedy, N. S., Gilchrist, J., Findlay, D., McLennan, J. M., 1994. The value of HMPAO SPECT scanning in the

diagnosis of early Alzheimer's Disease in patients attending a memory clinic. Nuclear Medicine Communications 15 (6), 405–409.

Messa, C., Perani, D., Lucignani, G., Zenorini, A., Zito, F., Rizzo, G., Grassi, F., Del Sole, A., Franceschi, M., Gilardi, M. C., Fazio, F., 1994. High-resolution Technetium-99m-HMPAO SPECT in patients with probable Alzheimer's disease: Comparison with Fluorine-18-FDG PET. Journal of Nuclear Medicine 35 (2), 210–216.

Mika, S., Rätsch, G., Schölkopf, B., Smola, A., Weston, J., Müller, K.-R., 1999a. Invariant Feature Extraction and Classification in Kernel Spaces. MIT Press.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.-R., 1999b. Fisher discriminant analysis with kernels. Proceedings of IEEE International Workshop on Neural Networks for Signal Processing IX, 41–48.

Minoshima, S., Foster, N., Kuhl, D., Sep. 1994. Posterior cingulate cortex in Alzheimer's disease. The Lancet 344 (8926), 895.

Minoshima, S., Frey, K. A., Koeppe, R. A., Foster, N. L., Kuhl, D. E., Jul. 1995. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of fluorine-18-FDG PET. Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine 36 (7), 1238–48, PMID: 7790950.

Minoshima, S., Goirdani, B., Berent, S., Frey, K., Foster, N., Khul, D., 1997. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. Annals of Neurology 42 (1), 85–94.

Miranda, A., Borgne, Y. L., Bontempi, G., 2008. New routes from minimal approximation error to principal components. Neural Processing Letters 27 (3), 197–207.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12 (2), 181–201.

Morris, J., 1993. Clinical dementia rating. Neurology 43, 2412–2414.

Morris, R. G., Kopelman, M. D., November 1986. The memory deficits in Alzheimer-type dementia: a review. The Quarterly Journal of Experimental Psychology 38 (A), 939–944.

Mosconi, L., Tsui, W. H., Herholz, K., Pupi, A., Drzezga, A., Lucignani, G., Reiman, E. M., Holthoff, V., Kalbe, E., Sorbi, S., Diehl-Schmid, J., Perneczky, R., Clerici, F., Caselli, R., Beuthiem-Baumann, B., Kurz, A., Minoshima, S., de Leon, M. J., 2008. Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease and other dementias. Journal of Nuclear Medicine 49 (3), 390–398.

Nefian, A. V., Hayes, M. H., 1999. An embedded hmm-based approach for face detection and recognition. In: In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 6. pp. 3553–3556.

Nestor, P. J., Fryer, T. D., Smielewski, P., Hodges, J. R., 2003. Limbic hypometabolism in Alzheimer's disease and mild cognitive impairment. Annals of Neurology 54 (3), 343–351.

Nitrini, R., Buchpiguel, C. A., Caramelli, P., Bahia, V. S., Mathias, S. C., Nascimento, C. M. R., Degenszajn, J., Caixeta, L., March 2000. SPECT in Alzheimer's disease: features associated with bilateral parietotemporal hypoperfusion. Acta Neurologica Scandinavica 101 (3), 172–176.

Nobili, F., Salmaso, D., Morbelli, S., Girtler, N., Piccardo, A., Brugnolo, A., Dessi, B., Larsson, S. A., Rodriguez, G., Pagani, M., 2008. Principal component analysis of FDG PET in amnestic MCI. European Journal of Nuclear Medicine and Molecular Imaging 35 (12), 2191–2202.

Norman, K., Polyn, S. M., Detre, G., Haxby, J. V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences.

Platt, J. C., 1999. Advances in Kernel Methods - Support Vector Learning. MIT Press, Ch. Fast Training of Support Vector Machines using Sequential Minimal Optimization, pp. 185–208.

Qi, F., Bao, C., Liu, Y., 2004. A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech. In: International Symposium on Chinese Spoken Language Processing. pp. 77–80.

Ramírez, J., Chávez, R., Górriz, J., Álvarez, I., López, M., Salas-Gonzalez, D., Segovia, F., June 2009. Functional brain image classification techniques for early Alzheimer disease diagnosis. In: IWINAC 2009 Proceedings. Lecture Notes in Computer Science. Santiago de Compostela (Spain).

Ramírez, J., Górriz, J., Romero, A., Lassl, A., Salas-Gonzalez, D., López, M., Río, M. G., 2008. Computer aided diagnosis of Alzheimer type dementia combining support vector machines and discriminant set of features. (Accepted in) Information Sciences.

Ramírez, J., Górriz, J. M., Gómez-Río, M., Romero, A., Chaves, R., Lassl, A., Rodríguez, A., Puntonet, C. G., Theis, F., Lang, E., 2008. Effective emission tomography image reconstruction algorithms for SPECT data. In: ICCS 2008, Part I, LNCS. Vol. 5101/2008. Springer-Verlag Berlin Heidelberg, pp. 741–748.

Ramírez, J., Yélamos, P., Górriz, J. M., Puntonet, C. G., Segura, J. C., 2006a. SVM-enabled voice activity detection. In: Lecture Notes in Computer Science. Vol. 3972. pp. 676–681.

Ramírez, J., Yélamos, P., Górriz, J. M., Segura, J. C., 2006b. SVM-based speech endpoint detection using contextual speech features. Electronics Letters 42 (7), 877–879.

Ramírez, J., Górriz, J., Segovia, F., Chaves, R., Salas-Gonzalez, D., López, M., Álvarez, I., Padilla, P., 2010. Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. Neuroscience Letters 472 (2), 99–103.

Ramírez, J., Górriz, J. M., Chaves, R., Salas-Gonzalez, D., Álvarez, I., López, M., Segovia, F., 2009. SPECT image classification using random forests. IET Electronics Letters 45 (12), 604–605.

Raudys, S. J., Jain, A. K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (3), 252–264.

Richards, J., 1995. Remote Sensing Digital Image Analysis. Springer-Verlag.

Rodríguez, G., Vitali, P., Calvini, P., Bordoni, C., Girtler, N., Taddei, G., Mariani, G., Nobili, F., 2000. Hippocampal perfusion in mild Alzheimer's disease. Psychiatry Research 100 (2), 39–62.

Roth, M., 1981. The diagnosis of dementia in late and middle life. The epidemiology of dementia. Oxford University Press.

Saitoh, S., 1988. Theory of Reproducing Kernels and its Applications. Longman Scientific & Technical, Harlow, England.

Salas-Gonzalez, D., Gorriz, J., Ramirez, J., Alvarez, I., Lopez, M., Segovia, F., Gomez-Rio, M., November 2009a. Skewness as feature for the diagnosis of Alzheimer's disease using SPECT images. In: Proceedings of the 16th IEEE International Conference on Image Processing (ICIP). Cairo (Egypt).

Salas-Gonzalez, D., Górriz, J. M., Ramírez, J., Lassl, A., Puntonet, C. G., October 2008a. Improved Gauss-Newton optimization methods in affine registration of SPECT brain images. IET Electronics Letters 44 (22), 1291–1292.

Salas-Gonzalez, D., Gorriz, J. M., Ramírez, J., López, M., Álvarez, I., Segovia, F., Puntonet, C. G., November 2008b. Computer aided diagnosis of alzheimer disease using support vector machines and classification trees. In: ICONIP 2008 Proceedings, Lecture Notes in Computer Science. Auckland (New Zealand).

Salas-Gonzalez, D., Górriz, J. M., Ramírez, J., López, M., Álvarez, I., Segovia, F., Puntonet, C. G., June 2009b. Analysis of brain SPECT images for the diagnosis of Alzheimer using first and second order moments. In: IWINAC 2009 Proceedings. Lecture Notes in Computer Science. Santiago de Compostela (Spain).

Salas-Gonzalez, D., Gorriz, J. M., Ramírez, J., López, M., Álvarez, I., Segovia, F., Puntonet, C. G., June 2009c. Selecting regions of interest for the diagnosis of Alzheimer's disease in brain SPECT images using Weltch's t-test. In: IWANN 2009 Proceedings, Lecture Notes in Computer Science. Salamanca (Spain).

Salas-Gonzalez, D., Górriz, J., Ramírez, J., López, M., Illán, I. A., Puntonet, C., Gómez-Río, M., 2009d. Analysis of SPECT brain images for the diagnosis of Alzheimer's disease using moments and support vector machines. Neuroscience Letters 461 (1), 60–64.

Salas-Gonzalez, D., Górriz, J. M., Ramírez, J., López, M., Álvarez, I., Segovia, F., Chaves, R., Puntonet, C., 2010. Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees. Physics in Medicine and Biology 55 (10), 2807–2817.

Salmon, E., Kerrouche, N., Perani, D., Lekeu, F., Holthoff, V., Beuthien-Baumann, B., Sorbi, S., Lemaire, C., Collette, F., Herholz, K., 2009. On the multivariate nature of brain metabolic impairment in alzheimer's disease. Neurobiology of Aging 30 (2), 186 – 197.

Santi, S. D., de Leon, M. J., Rusinek, H., Convit, A., Tarshish, C. Y., Roche, A., Tsui, W. H., Kandil, E., Boppana, M., Daisley, K., Wang, G. J., Schlyer, D., Fowler, J., Aug. 2001. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiology of Aging 22 (4), 529–539, PMID: 11445252.

Saxena, P., Pavel, D. G., Quintana, J. C., Horwitz, B., 1998. An automatic thresholdbased scaling method for enhancing the usefulness of Tc-HMPAO SPECT in the diagnosis of Alzheimers disease. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI, Lecture Notes in Computer Science. Vol. 1496. pp. 623–630.

Scarmeas, N., Habeck, C. G., Zarahn, E., Anderson, K. E., Park, A., Hilton, J., Pelton, G. H., Tabert, M. H., Honig, L. S., Moeller, J. R., Devanand, D. P., Stern, Y., 2004. Covariance PET patterns in early Alzheimer's disease and subjects with cognitive impairment but no dementia: utility in group discrimination and correlations with functional performance. NeuroImage 23 (1), 35 – 45.

Schölkopf, B., Smola, A., Müller, K. R., 1996. Nonlinear component analysis as a kernel eigenvalue problem. Tech. Rep. 44, Max-Planck-Institut für biologische Kybernetik.

Schölkopf, B., Smola, A., Müller, K. R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural computation 10 (5), 1299–1319.

Schneiderman, H., Kanade, T., 2000. A statistical method for 3d object detection applied to faces and cars. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 746–751.

Segovia, F., Gorriz, J., Ramirez, J., Salas-Gonzalez, D., Illan, I., Lopez, M., Chaves, R., Padilla, P., Puntonet, C., October 2009a. Automatic selection of ROIs using a model-based clustering approach. In: Proceedings of the 2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC). Orlando (Fl.).

Segovia, F., Gorriz, J., Ramirez, J., Salas-Gonzalez, D., Illan, I., Lopez, M., Chaves, R., Puntonet, C., Lang, E., Keck, I., October 2009b. fMRI data analysis using a novel clustering technique. In: Proceedings of the 2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC). Orlando (Fl.).

Segovia, F., Górriz, J., Ramírez, J., Salas-González, D., Álvarez, I., López, M., Chaves, R., Padilla, P., 2010. Classification of functional brain images

using a GMM-based multi-variate approach. Neuroscience Letters 474 (1), 58–62.

Shakhnarovich, G., Moghaddam, B., 2004. Face Recognition in Subspaces. Springer Verlag.

Sheikh, J., Yesavage, J., 1986. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. NY: The Haworth Press.

Silverman, D. H., Apr. 2004. Brain 18F-FDG PET in the diagnosis of neurodegenerative dementias: Comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging. Journal of Nuclear Medicine 45 (4), 594–607.

Silverman, D. H., Small, G. W., Chang, C. Y., 2001. Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome. Journal of the American Medical Association 286 (17), 2120–2127.

Silverman, D. H. S., Truong, C. T., Kim, S. K., Chang, C. Y., Chen, W., Kowell, A. P., Cummings, J. L., Czernin, J., Small, G. W., Phelps, M. E., Nov. 2003. Prognostic value of regional cerebral metabolism in patients undergoing dementia evaluation: comparison to a quantifying parameter of subsequent cognitive performance and to prognostic assessment without PET. Molecular Genetics and Metabolism 80 (3), 350–355, PMID: 14680983.

Spetsieris, P. G., Ma, Y., Dhawan, V., Eidelberg, D., 2009. Differential diagnosis of parkinsonian syndromes using functional PCA-based imaging features. NeuroImage 45 (4), 1241–1252.

Stoeckel, J., Ayache, N., Malandain, G., Koulibaly, P. M., Ebmeier, K. P., Darcourt, J., 2004. Automatic classification of SPECT images of Alzheimer's disease patients and control subjects. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI. Vol. 3217 of Lecture Notes in Computer Science. Springer, pp. 654–662.

Stoeckel, J., Malandain, G., Migneco, O., Koulibaly, P. M., Robert, P., Ayache, N., Darcourt, J., 2001. Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI. Vol. 2208 of Lecture Notes in Computer Science. Springer, pp. 666–674.

Swain, P. H., King, R. C., 1973. Two effective feature selection criteria for multispectral remote sensing. In: Proceedings of the 1st International Conference on Pattern Recognition. pp. 536–540.

Talairach, J., Tournoux, P., 1988. A Co-planar Stereotatic Atlas of the Human Brain. Stuttgart: Thieme.

Talbot, P. R., Lloyd, J. J., Snowden, J. S., Neary, D., Testa, H. J., 1998. A clinical role for 99mTc-HMPAO SPECT in the investigation of dementia? Journal of Neurology, Neurosurgery, and Psychiatry 64 (3), 306–313.

Tao, D., Tang, X., Li, X., Wu, X., 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7), 1088–1099.

Teipel, S. J., Stahl, R., Dietrich, O., Schoenberg, S. O., Perneczky, R., Bokde, A. L., Reiser, M. F., Möller, H.-J., Hampel, H., Feb. 2007. Multivariate network analysis of fiber tract integrity in Alzheimer's disease. NeuroImage 34 (3), 985–995.

Theodoridis, S., Koutroumbas, K., 2003. Pattern Recognition, Second Edtion. Academic Press.

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. Journal of congnitive neuroscience 3 (1), 71–86.

Vaillant, R., Monroc, C., LeCun, Y., 1993. An original approach for the localization of objects in images. In: International Conference on Artificial Neural Networks. pp. 26–30.

Vandenberghea, S., D'Asselera, Y., de Wallea, R. V., Kauppinenb, T., Koolea, M., Bouwensa, L., Laerec, K. V., Lemahieua, I., Dierckx, R., 2001. Iterative reconstruction algorithms in nuclear medicine. Computerized Medical Imaging and Graphics 25, 105–111.

Vapnik, V. N., 1982. Estimation of Dependences Based on Empirical Data. Springer-Verlag, New York.

Vapnik, V. N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin.

Vapnik, V. N., 1998. Statistical Learning Theory. John Wiley and Sons, Inc., New York.

Vardi, Y., Shepp, L. A., Kaufman, L., 1985. A statistical model for positron emission tomography. Journal of the American Statistical Association 80 (389), 8–20.

Viola, P., 1996. Complex feature recognition: A bayesian approach for learning to recognize objects. Tech. Rep. AIM-1591, MIT.

Wang, X., Tang, X., 2003. Unified subspace analysis for face recognition. In: Proceedings of IEEE International conference on Computer Vision. pp. 679–686.

Wiskott, L., 1995. Labeled graphs and dynamic link matching for face recognition and scene analysis. Ph.D. thesis, Ruhr-Universität Bochum, Bochum.

Woods, R. P., 2000. Spatial transformation models. In: Bankman, I. N. (Ed.), Handbook of Medical Imaging. Academic Press, San Diego, Ch. 29, pp. 465–490.

Yang, J., Jin, Z., yu Yang, J., Zhang, D., Frangi, A. F., 2004. Essence of kernel Fisher discriminant: KPCA plus LDA. Pattern Recognition 37, 2097–2100.

Yang, J., Yang, J. Y., 2003. Why can LDA be performed in PCA transformed space? Pattern Recognition 36, 563–566.

Yang, M. H., 2002. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. Proceedings of the fifth IEEE Interantional Conference on Automatic Face and Gesture Recognition, 215–220.

Yélamos, P., Ramírez, J., Górriz, J. M., Puntonet, C. G., Segura, J. C., 2006. Speech event detection using support vector machines. In: Lecture Notes in Computer Science. Vol. 3991. pp. 356–363.

Yu, H., Yang, J., 2000. A direct LDA algorithm for high-dimensional data – with application to face recognition. Pattern Recognition 34, 2067–2070.

Abreviaturas

AD: Alzheimer's Disease
ADNI: Alzheimer Disease Neuroimaging Initiative
CAD: Computer Aided Diagnosis
CDR: Clinical Dementia Rating
DGV: Degeneración Granulovacuolar
DNF: Degeneración Neurofribilar
EA: Enfermedad de Alzheimer
ECT: Emission Computed Tomography
FDR: Fisher Discriminant Ratio
fMRI: functional Magnetic Resonance Imaging
FWHM: Full Width at Half Maximum
GDS: Global Deterioration Scale
ICA: Independent Component Analysis
KDA: Kernel Discriminant Analysis
KPCA: Kernel Principal Component Analysis
LDA: Linear Discriminant Analysis
MCI: Mild Cognitive Impairment
MMSE: Mini-mental State Exam
MRI: Magnetic Resonance Imaging
PCA: Principal Component Analysis
PET: Positron Emission Tomography
PS: Placas Seniles
rCBF: regional Cerebral Blood Flow
ROI: Region of Interest
SPECT: Single Photon Emission Computed Tomography
SPM: Statistical Parametric Mapping
SVM: Support Vector Machine
VAF: Voxels-as-Features