EL ANÁLISIS DE COMPONENTES PRINCIPALES COMO MÉTODO DE CLASIFICACIÓN Y VISUALIZACIÓN DE TUMORES DE PARTES BLANDAS

César Vidal[1], Elías Malassidis[1], Juan Miguel García-Gómez[1], Luis Martí-Bonmatí[2,3], Montserrat Robles[1,3], José Millet[1]

[1]Bioingeniería, Electrónica y Telemedicina(BET).Universidad Politécnica de Valencia
[2]Clínica Quirón de Valencia
[3]Asociación para el Desarrollo y la Investigación de la Resonancia Magnética(ADIRM)

INTRODUCCIÓN

El Análisis de Componentes Principales (PCA) se usa en varias aplicaciones científicas como paso intermedio para análisis ulteriores, siendo un método estadístico de simplificación y reducción de la dimensionalidad de un conjunto de datos con numerosas variables. En este trabajo se describe su uso sobre una base de datos de tumores de partes blandas, agrupados en estirpes histológicas, y mediante un programa enMatlab® que usa los algoritmos más comunes para la implementación de PCA y la visualización tridimensional de los datos convertidos. Así se puede ver que PCA puede ser utilizado como un tipo de clasificador, ya que los datos transformados mantienen las características básicas de los datos iniciales y su representación en tres dimensiones da una imagen del conjunto de datos inicial sólo con una pequeña perdida de información

BASE DE DATOS

Se ha utilizado una base con 346 casos de varios hospitales: Hospital Universitario Dr. Peset (Valencia), Hospital Universitario San Juan (Alicante), y Clínica Quirón (Valencia). A partir de imágenes de Resonancia Magnética potenciadas en SE-T1 y SE-T2/STIR y datos epidemiológicos, diferentes radiólogos extrajeron independientemente 22 parámetros o variables por registro.

.

METODOLOGÍA

El Análisis de Componentes Principales forma en muchas maneras la base del análisis de datos multivariantes. PCA proporciona una aproximación de una tabla de datos, considerándola como una matriz de datos X con dimensiones N×K -donde las N filas son los casos y las K columnas, las variables- que se expresa como producto de dos matrices *t* y *p'* que mejor representan sus patrones esenciales. Así:

$$X=t\cdot p' + E$$

donde ***t*** son los *scores* (N×q) y ***p*** los *loadings* (K×q), coeficientes de las variables, y donde **q**, el número de Componentes Principales (q<=min(N,K)). En el modelo de PCA, la importancia de una variable está indicada por su varianza, mientras la matriz ***E*** contiene la parte de los datos no explicados por el modelo (*residuos*).

La varianza explicada por cada componente es el concepto algebraico del valor propio (*eigenvalue*) asociado. Una suposición básica en PCA es que los vectores de *scores* y *loadings* que corresponden a los valores propios más grandes contienen la información más útil que está relacionada con el problema, así que estos vectores propios suelen ser representados en orden decreciente.

El programa, desarrollado e implementado en Matlab®, utiliza el algoritmo de NIPALS (Non-Linear Iterative Partial Least Squares) y la función de Matlab® SVD que usa el algoritmo de la Descomposición de Valores Singulares para realizar el modelo de PCA. Ejecutando el programa, se ve una figura que representa los valores propios según el número del componente principal (figura 1). Entonces hay que elegir – o dejar que se calcule automáticamente – el número de componentes principales, según la proporción de varianza descrita por cada valor propio o basándose a varios criterios que existen para realizar esta selección.

Después de la selección, obtenemos la misma figura solamente con el número de componentes principales elegidos y la posibilidad de representar las tres primeras componentes (o cualesquiera tres de ellas) en tres dimensiones, utilizando los valores de los scores correspondientes (figura 2).

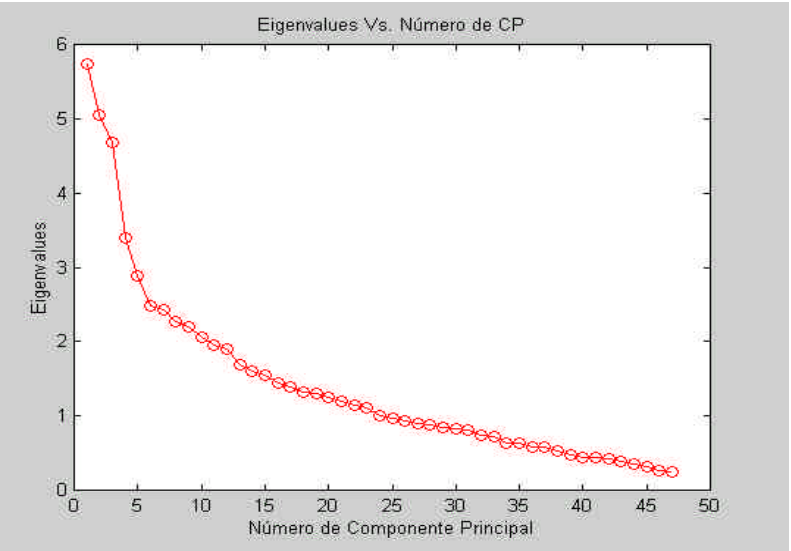

Figura 1. *Representación de los valores propios vs el número de la componente principal.*

RESULTADOS

Como se puede ver en la figura 2, representando columnas de la matriz de scores T, obtenemos una imagen de su configuración en el M-espacio. Los gráficos de unas de las combinaciones de las primeras componentes muestran los patrones más dominantes de la matriz de datos X.

En la figura tridimensional se representan los tumores de partes blandas de la base de datos con colores y formas distintas. Así se forman unas nubes de datos según el tipo histológico, con frecuencia bien separadas entre sí. Rotando la figura de Matlab®, haciendo un zoom (figura 3) o también utilizando otras combinaciones de los primeros scores (p.e. 1-2-4), se revelan distintas formas de ver los datos transformados, pudiendo algunas de clases de tumores que no se separan claramente con una combinación hacerlo con otra.

Se distinguen puntos que no pertenecen realmente a la nube donde se sitúan y que pueden ser “outliers”, datos bien clasificados pero extraños que indican casos de posible diagnóstico equivocado, que se deben por tanto revisar.

Los resultados numéricos del modelo de PCA se introdujeron en un clasificador que utiliza el método por los k vecinos más próximos para averiguar la eficacia en la separación entre las clases.

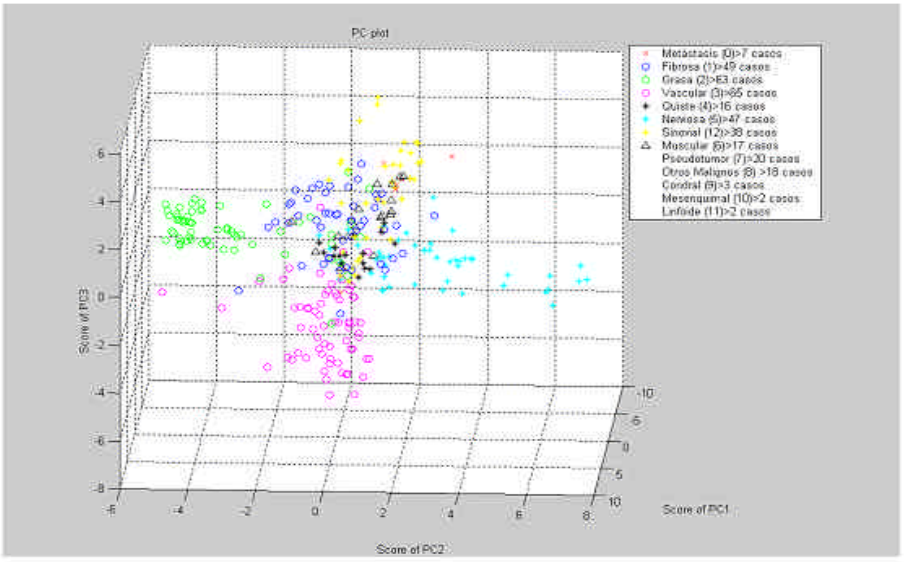

Figura 2. *Representación tridimensional de los tres primeros componentes.*

En todos los casos, PCA mejoró la clasificación de los datos en comparación con k-vecinos antes de la aplicación del modelo, variando entre 7%-18% de error mínimo, según el número de componentes principales, el número de clases elegidas para la clasificación y el método y el número de vecinos utilizados por el clasificador.

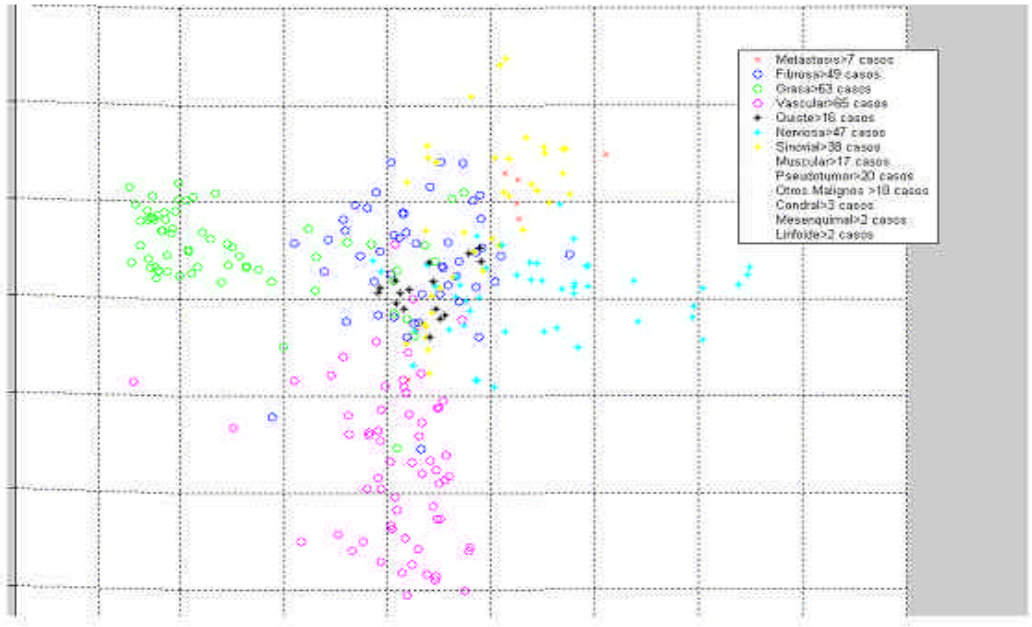

Figura 3. Representación tridimensional de los tres primeros componentes (zoom).

CONCLUSIONES

PCA reduce la dimensionalidad de un conjunto de datos, es decir, ofrece una transformación con un numero de variables importantemente reducido, a pesar de lo cual sigue describiendo los datos en su mayor parte.

El analisis de Componentes Principales se puede utilizar como clasificador y detector de 'outliers', ya que la representación de los primeros vectores de scores o componentes ofrece una imagen con las clases de los casos separados, donde un punto extraño puede ser un caso posiblemente mal diagnosticado que ha de ser revisado.

Los resultados de aplicar la técnica de k vecinos más cercanos a los datos obtenidos tras aplicar PCA con respecto a la aplicación sobre los datos originales mejoran en todos los casos.

BIBLIOGRAFÍA

1. Paul Geladi – Hans Grahn, *Multivariate Image Analysis*, John Wiley and Sons, 1996.
2. Tomás Aluja Banet – Alain Morineau, *Aprender de los datos: El análisis de componentes principales – Una aproximación desde el Data Mining*, EUB, 1999.
3. S.Wold, K.Esbensen & P.Geladi, *Principal Component Analysis*, Chemometrics and Intelligent Laboratory Systems, Elsevier Science Publishers B.V. (1987).
4. S.Raychaudhuri, J.M.Stuart and R.B.Altman, *Principal Components Analysis to summarize microarray experiments: Application to sporulation time series*, Stanford Medical Informatics.
5. A.Elbergali, J.Nygren, M.Kubista, An automated procedure to predict the number of components in spectroscopic data, Elsevier Science B.V. Analitica Quimica. Acta 379 (1999) 143-158.
6. www.nr.com: *Numerical Recipes in C: The art of scientific computing*, Cambridge University Press, 1988-1992.
7. R. Duda, P. Hart, D. Stork. *Pattern Classification*. John Wiley & Sons, Inc., New York, 2001.