Selección de Características 2D en Representaciones Tiempo Frecuencia para la Detección de Soplos Cardíacos

Juan D. Martínez-Vargas[1]
Luis D. Avendaño-Valencia[2]
Germán Castellanos-Domínguez[3]

Resumen

En el presente trabajo se propone una metodología para la reducción de dimensión en representaciones tiempo frecuencia (TFRs) enfocada a la clasificación de bioseñales no estacionarias, que trata directamente su cantidad de datos irrelevantes y redundantes, combinando una etapa de selección de características con una etapa de reducción de dimensión por medio de métodos de descomposición lineal extendidos a datos bidimensionales. La metodología se prueba sobre un conjunto de TFRs paramétricas calculadas sobre una base de datos de señales fonocardiográficas (FCG) para la detección de soplos cardiacos. Los resultados muestran una mejora comparados con otras metodologías que no tienen en cuenta la presencia de datos irrelevantes y redundantes en las representaciones, además, el uso de las metodologías de descomposición lineal bidimensionales reducen adecuadamente la redundancia de las TFRs, obteniendo un nuevo conjunto de características 2D de menor dimensión que el conjunto inicial.

1 Grupo de Control y Procesamiento Digital de señales, Universidad Nacional de Colombia Sede Manizales, jmartinezv@unal.edu.co
2 Grupo de Control y Procesamiento Digital de señales, Universidad Nacional de Colombia Sede Manizales, ldavendanov@unal.edu.co
3 Grupo de Control y Procesamiento Digital de señales, Universidad Nacional de Colombia Sede Manizales, cgcastellanosd@unal.edu.co

Fecha de recepción: 16 de Agosto de 2010
Fecha de aceptación: 16 de Enero de 2011

Palabras clave

Análisis de relevancia, selección de características, representaciones tiempo-frecuencia.

Abstract

In this paper is proposed a methodology for dimensionality reduction of time-frequency representations (TFRs) aimed to non-stationary biosignal classification that deals directly with large quantity of irrelevant and redundant data, combining a stage of feature selection with a stage of dimensionality reduction by linear decomposition methods extended to bidimensional data. The methodology is tested on a set of parametric TFRs computed from a phonocardiographic signal database (PCG) for detection of heart murmurs. Results show an improvement compared with other methodologies that do not account for irrelevant and redundant data in these representations and demonstrate that the use of bi-dimensional linear decomposition methods adequately reduce redundancy on TFRs, obtaining a new feature set of lower dimension than the original dataset.

Keywords

Relevance analysis, feature selection, time-frequency representations.

1. INTRODUCCIÓN

Las representaciones tiempo frecuencia (TFR) son mapas bidimensionales que describen como cambia el contenido en frecuencia de una señal a lo largo del tiempo, razón por la cual, son una de las técnicas más apropiadas de caracterización para señales no estacionarias. Sin embargo, a pesar de que la flexibilidad para construir el vector de características en la representación 2D se considere la ventaja principal de los métodos de clasificación basados en el dominio t-f, se encuentran aún varios problemas abiertos.

En primer lugar, los cambios sutiles en las bioseñales, que pueden ser indicadores de anormalidades, generalmente son difíciles de localizar a simple vista en las TFR. Además, la dimensionalidad intrínseca de los planos t-f es alta, por lo tanto, se hace esencial la extracción de las características más relevantes de la TFR, ya sea punto a punto o por localidades, para resolver los problemas de sobre-ajuste, sobre- entrenamiento y costo computacional, que es particularmente alto en estos casos de análisis multivariado. En consecuencia, existe una creciente necesidad de nuevos métodos de reducción de dimensión que puedan parametrizar adecuadamente la actividad de las bioseñales utilizando TFRs.

La extracción de información relevante a partir de conjuntos de características bidimensionales se ha discutido (Avendaño et al., 2010; Sanchez & Castellanos, 2010; Tzallas et al., 2009; Zhao & Zhang, 2006; Zuo et al., 2006) como una forma para mejorar el rendimiento en procesos de aprendizaje. La literatura señala que, para obtener un algoritmo efectivo de selección de características, se deben resolver principalmente dos problemas: Las distancias entre diferentes planos t-f asociadas a una medida de relevancia dada y la transformación multivariada, la cual se escoge para maximizar la medida de relevancia presente en los planos t-f por medio de su proyección a un nuevo espacio.

Así por ejemplo, una metodología simple consiste en determinar un conjunto de celdas t-f que representen, mediante una medida simple, la energía de la bioseñal en cada banda de frecuencia específica y para una ventana dada de tiempo, como puede ser la energía promedio dentro de cada celda (Tzallas et al., 2009). Sin embargo, en tal aproximación, queda un tema sin resolver asocia-

do con el análisis basado en localidades; a saber, la selección del tamaño de las regiones locales relevantes (Sejdic et al., 2009). Los métodos de descomposición lineal, los cuales hacen uso de los conceptos de distancia y transformación, también han sido considerados para la extracción de características sobre los planos TFR (Zhao & Zang, 2006; Zuo et al., 2006), aunque en este caso, no es claro cómo se debe fijar de forma previa un área conveniente de relevancia sobre el plano t-f con el fin de lograr estabilidad computacional durante el proceso de reducción de dimensión y disminuir la resolución de la TFR, como se afirmó en (Avendaño et al.. 2010).

En el presente trabajo se propone una metodología para reducción de dimensión orientada a espacios 2D, que lleva a cabo inicialmente, la selección de características relevantes, y posteriormente, la descomposición lineal sobre los planos t-f. Primero se calculan las TFRs para todo el rango espectral de la señal en análisis, luego se encuentran las bandas de frecuencia más relevantes utilizando como medida la incertidumbre simétrica, que cuantifica la dependencia estadística entre las características y las etiquetas. Después, se recalculan las TFRs únicamente para las bandas de frecuencia con mayor carga informativa. Seguidamente, la reducción de dimensión se realiza mediante métodos de descomposición lineal extendidos a datos matriciales.

Se consideran los métodos de análisis de componentes principales (Principal Component Analysis - PCA) como transformación lineal no supervisada, siendo esta una de las metodologías de reducción de dimensión más conocidas, y el método de mínimos cuadrados parciales (Partial least Squares - PLS) como transformación lineal supervisada, que permite separar las características más discriminantes entre las clases presentes en la base de datos, tomando la información de las etiquetas (Barker & Rayens, 2003). Los métodos de reducción planteados se utilizan de forma tal que se resuelvan los problemas de dimensionalidad y redundancia en ambos ejes del plano t-f. El criterio utilizado para comparar las diferentes metodologías es la tasa de acierto de clasificación, utilizando una metodología de validación cruzada de 10 particiones y un clasificador de k vecinos más cercanos.

La metodología se prueba sobre un conjunto de TFRs paramétricas calculadas sobre una base de datos de señales fonocardio-

gráficas (FCG) para la detección de soplos cardíacos. Los resultados muestran que el cálculo de las TFRs, dando más resolución a las áreas relevantes de las representaciones ayuda a mejorar el desempeño en la clasificación, y que el uso de las metodologías de descomposición lineal bidimensionales reduce adecuadamente la redundancia de las TFRs, obteniendo un nuevo conjunto de características 2D de menor dimensión que el conjunto inicial. El artículo se organiza de la siguiente forma: Primero se presentan los métodos de selección de características basadas en TFRs asumiendo medidas de relevancia como también la extensión de los métodos de descomposición lineal para datos matriciales. Después se presentan los resultados obtenidos con la metodología propuesta y se comparan con algunos resultados obtenidos en la literatura.

2. METODOLOGÍA

2.1 Análisis de Relevancia

El objetivo en la selección de características es hallar el mínimo subconjunto de características X_r tal que (1) (Yu & Liu, 2004):

$$P(c|X) \sim P(c|X_r) \tag{1}$$

Donde $P(c|X)$ es la función de distribución de probabilidad (PDF) de las clases c dado el conjunto completo de características X y $P(c|X_r)$ es la distribución de probabilidad de las clases dado el subconjunto de características X_r. Nótese que el conjunto de características X se considera como un vector aleatorio de características con entradas X_i, con $i = 1, \dots, m$, donde m es el número de características. Sea $X_s$un conjunto de variables y $\bar{X}_s$ el conjunto complemento de X_s tal que $X_s \cup \bar{X}_s = X$. El conjunto de características X_s se dice *fuertemente relevante* si y solo si (2):

$$P\left(c\middle|X_{s,}\bar{X}_s\right) \neq P(c|\bar{X}_s) \tag{2}$$

lo cual significa que la PDF condicional de las etiquetas c cambia cuando el subconjunto X_s se extrae del conjunto de característi-

cas X. Además, el conjunto de características X_s se dice *débilmente relevante* si y solo si (3):

$$P\left(c\middle|X_s,\bar{X}_s\right) = P(c|\bar{X}_s) \tag{3}$$

y existe algún $\bar{X}_s^* \subset \bar{X}_s$ tal que (4)

$$P(c|X_s,\bar{X}_s^*) \neq P(c|\bar{X}_s^*) \tag{4}$$

Este hecho significa que la PDF condicional de las etiquetas c, dado el conjunto de características Xen (3) y (4) no presenta cambios cuando se remueve el subconjunto de características X_s, pero cambia cuando se extrae algún otro subconjunto del conjunto de características. Así, se puede concluir que el conjunto de características X tiene la misma información de X_s. De esta forma, las características débilmente relevantes se pueden relacionar con datos redundantes en el conjunto de características. Cualquier subconjunto X_s que no cumpla alguno de los enunciados previos se dice *irrelevante*.

En (1) y (2) se sugiere que se deben seleccionar los datos relevantes extrayendo aquellas variables que presenten la mayor influencia en la PDF condicional de c, dado el conjunto de características X. La dependencia entre la PDF de las etiquetas y la PDF de las características se puede medir, por ejemplo, basándose en teoría de información y el concepto de *entropía,* la cual mide la incertidumbre de una variable aleatoria. La entropía se define como (5):

$$H(X_i) = -\int_{x_i} P(X_i)\log P(X_i)\,dX_i \;, \forall i = 1,\dots,m \tag{5}$$

Además, la entropía de la característica X_i, después de observar los valores de las etiquetas de clase c, se define como (6):

$$H(X_i) = -\int_{c} P(c)\int_{x_i} P(X_i|c)\log P(X_i|c)\,dX_i dc \;, \forall i = 1,\dots,m \tag{6}$$

El valor en el cual la entropía de la característica $H(X_i)$ decrece cuando se utilizan las etiquetas de clase, refleja la información adicional que c suma a X_i, y se llama *ganancia de información*, definida como (7):

$$I_g(X_i|c) = H(X_i) - H(X_i|c), \forall i = 1, \dots, m \tag{7}$$

Los valores de la medida en (7) se normalizan si se dividen por la suma de las entropías de cada variable (8):

$$\rho_{su}(X_i, c) = 2\frac{H(X_i) - H(X_i|c)}{H(X_i) + H(c)}, \rho_{su}(X_i, c)\epsilon[0,1], \forall i = 1, \dots, m \tag{8}$$

La medida de relevancia $\rho_{su}(X_i, c)$ descrita en (8) se conoce como *Incertidumbre Simétrica*. Un valor de $\rho_{su}(X_i, c) = 1$ indica que la característica X_i predice completamente los valores de las etiquetas de clase c. Dado que para calcular la medida de (8) es necesario estimar $P(X_i)$ y $P(X_i|c)$, en este trabajo se utilizan estimaciones basadas en histogramas. Las integrales en (5) y (6) se convierten en sumas que se calculan sobre los intervalos en los cuales se estimaron los histogramas.

2.2 Métodos de Descomposición Lineal 2D

En las metodologías de descomposición lineal, el conjunto de datos original $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ se mapea a un conjunto de características con menor dimensión $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$ por medio de la transformación (9):

$$\boldsymbol{Y} = \boldsymbol{XW} \tag{9}$$

donde $\boldsymbol{W} \in \mathbb{R}^{m\times q}$ es la matriz de transformación, n es el número de muestras en la base de datos, m es el número de características en el espacio original y q es el número de características en el espacio reducido. El valor de $\boldsymbol{W}$ en (9) depende específicamente del método utilizado. En el caso de PCA, esta matriz se obtiene como los vectores singulares de la matriz de covarianza de $\boldsymbol{X}$ (Jollife, 1986), mientras que en el caso de PLS esta matriz se obtiene iterativamente como los vectores que maximizan la correlación entre $\boldsymbol{X}$

y las etiquetas de clase c, y maximicen la covarianza de $\boldsymbol{X}$ (Barker & Rayens, 2003).

Ahora, si se calculan transformaciones utilizando las filas y las columnas de la matriz como individuos independientes, cada superficie $\boldsymbol{X}_k$ se concatena en una súper matriz y se lleva a cabo PCA o PLS para obtener una matriz de transformación $\boldsymbol{V}$ con dimensión $M \times n_{rpc}$, donde n_{rpc} es el número de componentes de las filas (Yang et al., 2004). Esta transformación tendrá en cuenta la relación entre filas. Con el fin de tener en cuenta la relación entre columnas, el mismo procedimiento se aplica a una súper matriz construida con $\boldsymbol{X}_k^{\mathrm{T}}$, para obtener una matriz de transformación $\boldsymbol{W}$ con dimensiones $N \times n_{cpc}$ donde $N \times n_{cpc}$ es el número de componentes para las columnas (Zhang & Zhou, 2005). Así, cada matriz de características se transforma a una matriz reducida $\boldsymbol{Y} \in \mathbb{R}^{(n_{rpc} \mathrm{x} n_{cpc})}$, dada por (10):

$$\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{X}\boldsymbol{W}^{\mathrm{T}} \tag{10}$$

El proceso en (10) será denominado 2DPCA (Zhang & Zhou, 2005) y 2DPLS, el cual es una extensión que se realiza del método 2DPCA. Su esquema general se muestra en la Fig.1, como resultado, en el caso de las TFR, la reducción de dimensión tiene en cuenta no solo la variabilidad instante por instante de cada variable aleatoria, sino que también busca la variabilidad de información a lo largo del espectro de frecuencia.

2.3 Marco Experimental

La Fig.2 muestra la metodología considerada para ajustar el método de selección de características propuesto para la detección de soplos cardíacos. La metodología se divide en tres pasos consecutivos: 1). Estimación de las TFR paramétricas, la cual comprende la sintonización de los parámetros del estimador y la sintonización de la resolución de la TFR; 2) Selección de características, la cual presenta la selección de las bandas de frecuencia relevantes de la TFR y la transformación de los datos por medio de los métodos de descomposición lineal; y 3) Clasificación, donde se utiliza un

clasificador de k vecinos más cercanos (k-NN) y un esquema de validación cruzada de 10 particiones.

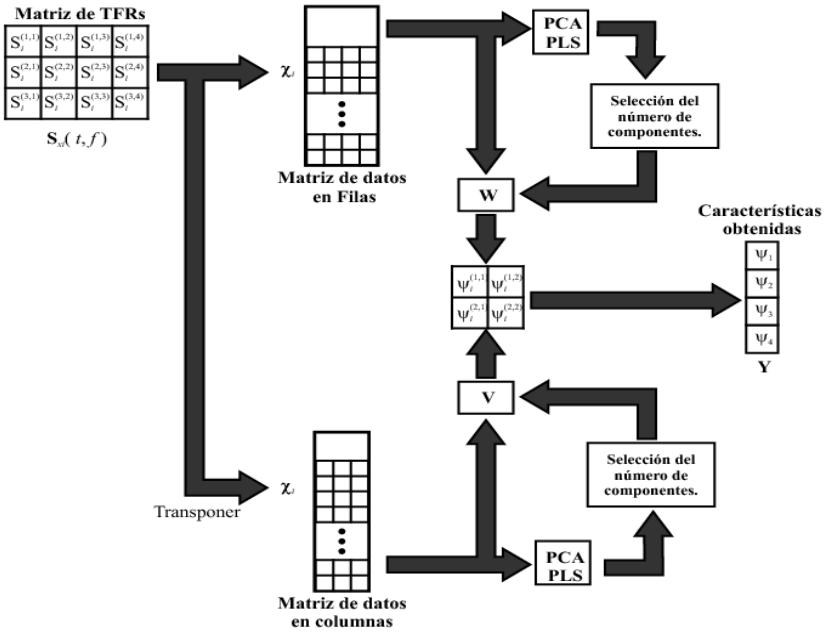

Fig. 1. Esquema general 2DPCA y 2DPLS

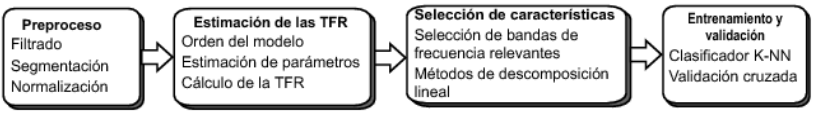

Fig. 2. Esquema general de la metodología propuesta para la detección de soplos cardíacos

2.4 Adquisición de la Base de Datos y Pre-proceso

La metodología se prueba en un conjunto de señales fonocardiográficas pertenecientes al Grupo de Control y Procesamiento Digital de Señales de la Universidad Nacional de Colombia – Sede Manizales y al grupo Telesalud de la Universidad de Caldas. La base de datos de señales FCG consta de 45 sujetos adultos, quie-

nes dieron su consentimiento informado aprobado por un comité de ética de un centro hospitalario, y se sometieron a un examen médico. El diagnóstico se llevó a cabo para cada paciente, y la severidad de la afección valvular fue evaluada por cardiólogos de acuerdo con los procedimientos rutinarios.

Un conjunto de 26 pacientes fue etiquetado como normal, mientras que otros 19 mostraron evidencias de soplos sistólicos o diastólicos, causados por deficiencias valvulares. Para cada paciente, se tomaron 8 registros correspondientes a los cuatro focos tradicionales de auscultación (mitral, tricúspide, aórtico y pulmonar) en las fases de apnea post-espiratoria y post-inspiratoria (Chizner, 2008). Cada registro tiene una duración aproximada de 12 s y se obtuvo con el paciente en posición de cúbito dorsal. Las señales se adquirieron con un estetoscopio electrónico (modelo WelchAllyn®Meditron), con el cual se graba simultáneamente la señal FCG y una derivación de la señal electrocardiográfica (ECG) (DII), la cual se usa como referencia de sincronización para segmentar cada uno de los latidos. Ambas señales se muestrean a una tasa de 44,1 kHz con una precisión de 16 bits.

Se realiza un preproceso que incluye un submuestreo a 4000 Hz, normalización de la amplitud y segmentación por latidos, como ha sido propuesto en (Delgado et al., 2009). Se realiza un diagnóstico por cada uno de los latidos, dado que su evaluación individual es más precisa a la que se puede obtener utilizando el registro completo, además los soplos cardíacos generalmente no aparecen en todos los focos de auscultación a menos que sean muy intensos. Finalmente se seleccionan para la base de datos 548 latidos, 274 normales y 274 patológicos. La selección de los 548 latidos utilizados en el entrenamiento y la validación del sistema fue realizada por cardiólogos expertos que escogieron los latidos más representativos de pacientes normales y patológicos (con soplos cardíacos).

2.5 Estimación de las TFR Paramétricas

Un modelo AR variante en el tiempo, TVAR(p) donde p designa el orden del modelo autorregresivo, se describe por la siguiente expresión de autorregresión lineal (Poulimenos & Fassois, 2006):

$$x[t] = \Theta^{\mathrm{T}}[t]\, \boldsymbol{h}[t-1] + e[t], e[t] \sim \mathbf{N}(0, \sigma_e^2[t]) \tag{11}$$

Donde $\boldsymbol{h}[t] = \{x[t-i]: i = 1, \ldots, p\}, \boldsymbol{h} \subset \mathbb{R}^{\boldsymbol{p}}$ es el proceso no estacionario (vector de valores reales) a ser modelado, $e[t]$ es una secuencia de innovaciones, correspondiente a la parte aleatoria que el modelo no puede predecir, la cual es no correlacionada y no observable con media cero y varianza dependiente del tiempo $\sigma_e^2[t]$, y $\Theta[t] = \{\Theta_i[t]\}, \Theta \subset \mathbb{R}^{\boldsymbol{p}}$ son los parámetros del modelo TVAR. El vector de parámetros $\Theta[t]$ y la varianza $\sigma_e^2[t]$ en (11), para una frecuencia de muestreo dada, f_s, se relacionan con el contenido espectral de $x[t]$ por medio de (12) (Tarvainen et al., 2004):

$$S_x(t,f) = \frac{\sigma_e^2[t]}{|1 + \sum_{i=1}^{p} \Theta_i[t] e^{-j\omega i t/f_s}|}, S_x(t,f) \subset \mathbb{R} \tag{12}$$

La expresión obtenida en (12) se puede asumir como la densidad espectral de potencia de la respuesta de la señal si el sistema fuera estacionario en el instante de tiempo t. Se presentan principalmente dos problemas cuando se van a estimar las TFRs paramétricas, la estimación de los parámetros $\Theta[t]$ y la obtención del orden del modelo. La estimación de $\Theta[t]$ se puede realizar por medio del método *smoothness priors* (Kitagawa & Gersch, 1985), mientras que la selección del orden del modelo se realiza siguiendo la metodología propuesta en (Poulimenos & Fassois, 2006), que se basa en la minimización del criterio de información Bayesiano (BIC) (13):

$$BIC(p) = -\sum_{t=1}^{N} \left(\ln(\sigma_e^2[t]) + \frac{e^2[t]}{\sigma_e^2[t]} \right) + p \ln N, \tag{13}$$

donde N es el número de puntos de la señal, p es el orden del modelo, $e[t]$ es la secuencia de innovaciones o los residuos de la estimación y $\sigma_e^2[t]$ es la secuencia de innovaciones de la varianza dependiente del tiempo. La Fig.3 muestra el histograma de la minimización del criterio BIC (13) para estados normales y patológicos, variando el orden del modelo p de 1 hasta 15.

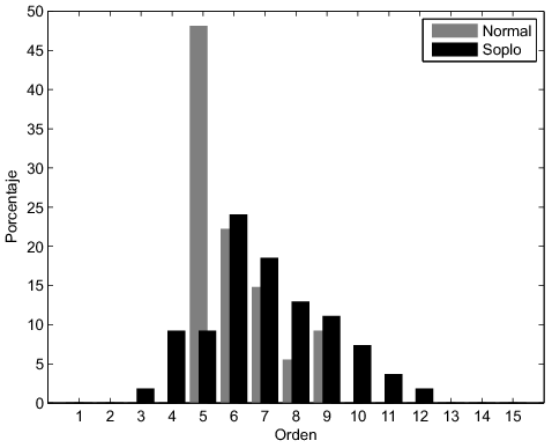

Fig. 3. Histograma de minimización del criterio BIC

Se selecciona como orden del modelo $p = 7$, buscando el mínimo orden que modele adecuadamente la mayor parte de la base de datos. Una vez se selecciona el orden, se estiman los parámetros del modelo TVAR utilizando el suavizador de Kalman (Tarvainen et al., 2004); después se calcula la TFR paramétrica por medio de (12). Los parámetros del estimador $\sigma_p^2[t], \sigma_e^2[t], y\ \sigma_\omega^2[t]$ se sintonizaron en $1e5{,}1$ y $1e-4$ respectivamente. Finalmente, se obtiene un conjunto de TFRs paramétricas, las cuales contienen $T = 600$ puntos en el tiempo, y $F = 256$ puntos en frecuencia, para un rango $t = [0\ 1.2]$ s y $f = [0\ 2000]\ Hz$.

2.6 Análisis de Relevancia

La selección de características se lleva a cabo obteniendo los valores de relevancia del rango de frecuencias estudiado, utilizando la medida de Incertidumbre Simétrica descrita en la sección anterior. La Fig. 4 muestra los valores de relevancia ρ obtenidos para cada una de las frecuencias. Las bandas de frecuencia son aquellas que superen cierto umbral. La sintonización de este umbral se describe más adelante. Se puede observar que las variables con mayor información se concentran en una zona claramente definida, mientras que las demás pueden ser consideradas como poco relevantes.

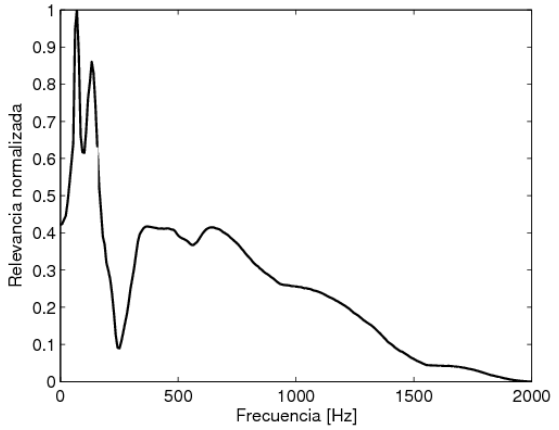

Fig. 4. Relevancia normalizada para la frecuencia utilizando Incertidumbre Simétrica (SU)

Después de sintonizar el valor de ρ, y aprovechando que las TFR paramétricas se pueden calcular sobre los rangos de frecuencia deseados, se construye un nuevo conjunto de TFRs con el mismo número de puntos en tiempo y en frecuencia que el conjunto inicial, pero con mayor resolución en la zona deseada, con el fin de resaltar las características con mayor información.

2.7 Sintonización de los Métodos de Descomposición Lineal

Después de seleccionar las variables más relevantes, el conjunto de datos obtenido se reduce utilizando los métodos de descomposición lineal 2D descritos en la sección anterior. Una de las dificultades consiste en cómo hallar el número de componentes necesarias tanto en filas como en columnas para obtener la mejor generalización para la predicción de nuevas observaciones. Un estimado preliminar se escoge basándose en el número de componentes que superen el 90% de la variabilidad tanto en el eje del tiempo como en el eje de la frecuencia. Los valores obtenidos se utilizan como base para la sintonización propuesta en la sección *Resultados*.

2.8 Clasificación

El enfoque de validación cruzada o *cross-validación* utilizado para evaluar el rendimiento de la metodología, consiste en dividir la base de datos en 10 particiones o *folds* que contengan diferentes registros entre ellas y una cantidad uniforme de registros de cada clase. Nueve de estos *folds* se utilizan para entrenamiento y el restante se utiliza para la validación. La metodología propuesta se aplica a los *folds* de entrenamiento, y el espacio de características resultante se utiliza para entrenar un clasificador de k-vecinos más cercanos *k-NN*. Después, las zonas relevantes, las matrices de transformación y el clasificador entrenado, se utilizan para clasificar los registros del *fold* de validación. Este procedimiento se repite cambiando los *folds* de entrenamiento y validación hasta que los 10 *folds* se hayan utilizado para validar el clasificador. El rendimiento de clasificación se obtiene por medio de algunas medidas estándar definidas por (14):

$$\begin{aligned} Tasa\ de\ acierto\ (\%) &= \frac{N_C}{N_T} \times 100 \\ Sensibilidad\ (\%) &= \frac{N_{TP}}{N_{TP}\ +\ N_{FN}} \times 100 \\ Especificidad(\%) &= \frac{N_{TN}}{N_{TN}\ +\ N_{FP}} \times 100 \end{aligned} \tag{14}$$

Donde N_C es el número de registros clasificados correctamente, N_T es el número total de registros ingresados al clasificador, N_{TP} es el número de verdaderos positivos (clase objetivo clasificada correctamente), N_{FN} es el número de falsos negativos (clase objetivo clasificada como clase de referencia), N_{TN} es el número de verdaderos negativos (clase de referencia clasificada correctamente) y N_{FP} es el número de falsos positivos (clase de referencia clasificada como clase objetivo). La clase patológica corresponde a la clase objetivo mientras que la clase normal corresponde a la clase de referencia. La tasa de acierto, la sensibilidad y la especificidad se miden para cada *fold*, de éste modo los valores de la media y la desviación estándar se utilizan para obtener los índices de desempeño.

3. RESULTADOS Y DISCUSIÓN

En esta sección se muestra la evaluación de la sintonización de los parámetros de la metodología: El número de vecinos del clasificador *k-NN*, el número de componentes tanto en filas como en columnas utilizados por los métodos de descomposición lineal (2DPCA y 2DPLS), y el umbral de relevancia ρ para seleccionar las bandas de frecuencia más relevantes. Con el fin de comparar los resultados obtenidos con los diferentes métodos, se calculan la media y la desviación estándar de las medidas de rendimiento utilizadas.

3.1 Sintonización del Umbral de Relevancia

Con el fin de evaluar la efectividad de la medida de relevancia utilizada, se incrementa el umbral de relevancia para la selección de las bandas de frecuencia que más información aportan al proceso. Dado que los posibles valores de ρ se encuentran entre 0 y 1, este valor se incrementa desde 0,1 hasta 0,8 con pasos de 0,1. La prueba se realiza para ambos métodos de descomposición lineal (2DPCA y 2DPLS). La Fig. 5 muestra los resultados obtenidos.

Se puede observar que ambos métodos de descomposición lineal presentan un comportamiento similar. Se selecciona un valor de ρ=0.4, dado que presenta una tasa de clasificación mayor con una menor desviación estándar.

3.2 Sintonización del Clasificador k-NN

Se incrementa paso a paso el número de vecinos del clasificador con el fin de determinar cuál es el valor de k que optimiza la tasa de clasificación. Se varía k entre 1 y 9 tomando únicamente los valores impares para evitar empates en la regla de decisión. La Fig. 6 muestra los resultados de la prueba para 2DPCA (arriba) y 2DPLS (abajo) respectivamente.

Se puede observar que la tasa de clasificación disminuye a medida que el número de vecinos aumenta. Se escoge como valor óptimo $k = 1$ dado que es valor de k que maximiza la tasa de acierto en la clasificación. Por el resultado obtenido, se puede inferir que,

la estructura general del espacio de características puede presentar pequeñas agrupaciones (*clusters*), o una frontera de decisión poco suave.

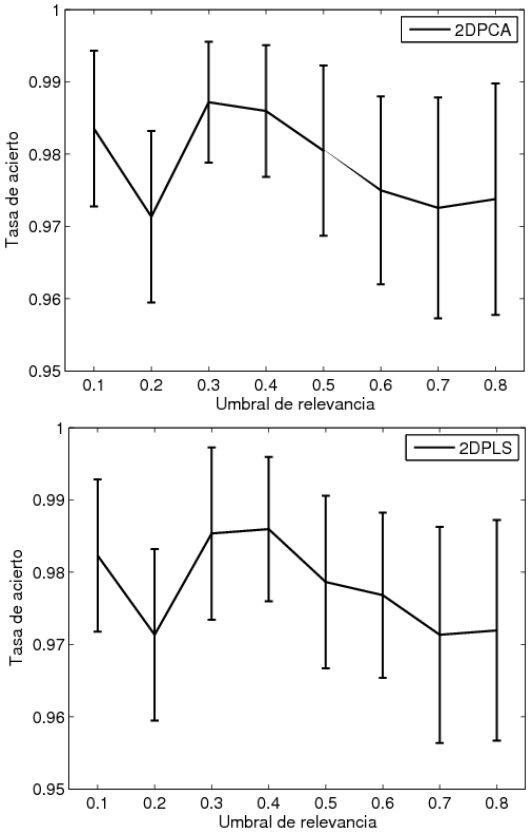

Fig. 5. Sintonización del umbral de relevancia

3.3 Sintonización del Número de Componentes

La sintonización del número de componentes se realiza tanto en filas como en columnas. Primero, se incrementa paso a paso el número de componentes tomando valores iguales tanto en filas como en columnas ($n_{rpc=}n_{cpc}$). La Fig. 6 muestra que la tasa de clasificación se estabiliza en ambos casos (2DPCA y 2DPLS) en 20

componentes aproximadamente. Ahora, para obtener una mejor aproximación del número de componentes requeridos, se fija el número de componentes en columnas $n_{cpc} = 21$ y se varía el número de componentes en filas n_{cpc} entre 1 y 21 (Fig. 7 arriba para 2DPCA y Fig. 8 arriba para 2DPLS), se realiza también el procedimiento inverso, es decir, se fija el número de componentes en filas mientras se modifica el número de componentes en columnas (Fig. 7 abajo para 2DPCA y Fig. 8 abajo para 2DPLS)

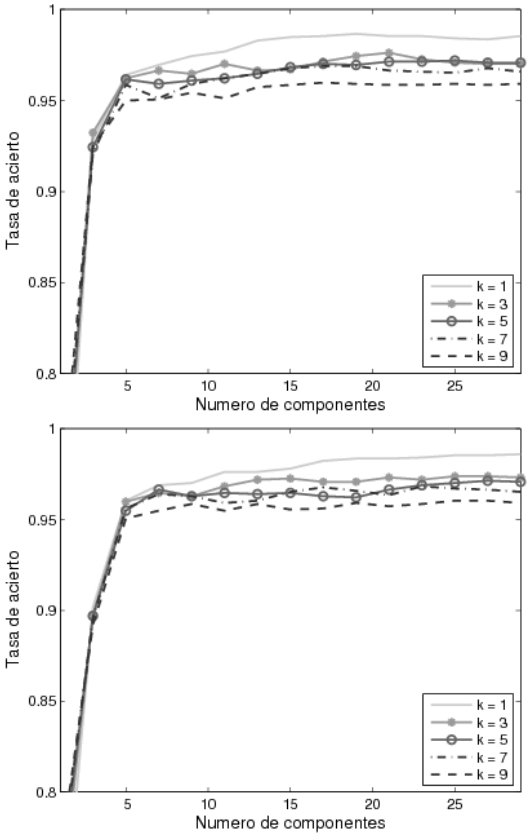

Fig. 6. Sintonización del número de vecinos del clasificador y el número de componentes en los métodos de descomposición lineal

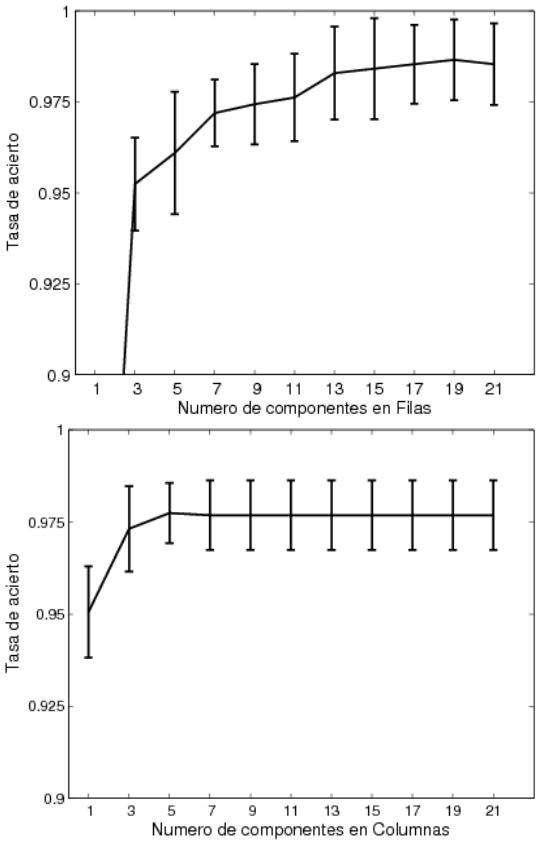

Fig. 7. Sintonización del número de componentes en filas y en columnas para 2DPCA

Aunque los resultados obtenidos variando el número de componentes en columnas se estabilizan en pocas componentes, la tasa de clasificación de los resultados variando el número de componentes en filas es mayor, por esta razón se seleccionan los valores de la sintonización en filas. La Tabla 1 muestra el número de componentes tanto en filas como en columnas seleccionados para ambas metodologías y el tamaño del espacio de características antes y después de realizar la transformación.

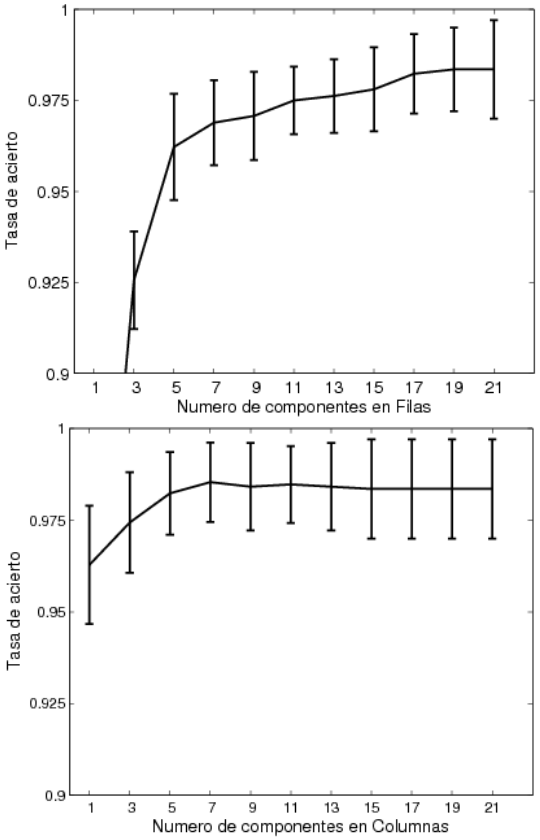

Fig. 8. Sintonización del número de componentes en filas y en columnas para 2DPLS

Tabla 1. Resultados de la reducción de dimensión

Método	$\mathbf{n_{rpc}}$	$\mathbf{n_{cpc}}$	Dimensión inicial	Dimensión final
2DPCA	21	21	256x600	21x21
2DPLS	21	19	256x600	21x19

Los resultados indican que tanto la actividad temporal como la actividad en frecuencia presentan gran variabilidad, ya que no se puede agrupar en pocas componentes.

3.4 Sumario de Resultados

La Tabla 2 resume los resultados obtenidos aplicando la metodología propuesta con los diferentes métodos de descomposición lineal (2DPCA y 2DPLS). Además presenta algunos resultados obtenidos sin realizar el análisis de relevancia previo, y con los métodos de descomposición lineal en una dimensión como se propone en Avendaño et al. (2010). Se puede observar que los resultados obtenidos con la metodología propuesta están aproximadamente en el 98% de acierto en clasificación.

Tabla 2. Sumario de los resultados obtenidos con las metodologías propuestas

	Tasa de acierto (%)	Sensibilidad (%)	Especificidad (%)
PCA con TFR vectorizadas	92,46 ± 3,29	93,50 ± 4,19	95,22 ± 2,22
PLS con TFR vectorizadas	95,83 ± 1,31	98,66 ± 1,31	96,55 ± 1,77
2DPCA	97,74 ± 1,22	97,56 ± 1,29	97,92 ± 0,82
2DPCA + Rel	98,53 ± 1,12	98,04 ± 1,92	99,02 ± 0,77
2DPLS	97,86 ± 1,08	97,56 ± 2,07	98,17 ± 1,03
2DPLS + Rel	98,65 ± 1,10	98,29 ± 1,83	99,02 ± 0,77

Finalmente, la Tabla 3 muestra los mejores resultados obtenidos con la metodología propuesta, comparados con otras metodologías en el estado del arte para la detección de soplos cardíacos. Se compara la metodología propuesta con la publicada en (Wang et al., 2006) utilizando una base de datos similar. Además, se compara con los resultados obtenidos en (Delgado et al., 2009) y (Quiceno et al., 2010), utilizando un subconjunto de la base de datos aplicada en este trabajo, que no contiene soplos diastólicos, y que por ende constituye un problema de clasificación de menor dificultad y generalidad.

3.5 Discusión

Se llevaron a cabo diferentes pruebas para evaluar el comportamiento de la metodología propuesta. Se utilizó un conjunto de señales fonocardiográficas para la detección de soplos cardíacos. La medida de relevancia reflejó que, para las señales FCG, la información se encuentra localizada en unas bandas de frecuencia

específicas (Fig 4), y que según los valores de incertidumbre simétrica obtenidos, el resto de la información puede considerarse irrelevante; además, la tasa de acierto de clasificación cuando se realizó el análisis de relevancia previo supera los resultados sin desarrollar esta etapa. Por otro lado, ambos métodos de descomposición lineal presentaron resultados similares, aunque 2DPLS alcanzó un mejor desempeño con una menor cantidad de componentes, la diferencia en clasificación fue únicamente 1 punto. Los resultados obtenidos con las otras medidas de desempeño (sensibilidad y especificidad) concuerdan con lo mencionado anteriormente. Los resultados obtenidos con la metodología propuesta superan por lo menos en 3 puntos de clasificación a los obtenidos con las metodologías de descomposición lineal convencionales.

Tabla 3. Comparación con otras metodologías propuestas en la literatura

Método	Tasa de acierto (%)	Sensibilidad (%)	Especificidad (%)
Este estudio	98,65	98,29	99,02
Características Perceptuales (Wang et al, 2006)	88,90	93,03	85,81
Características fractales, perceptuales y tiempo frecuencia (Delgado et al, 2009)	96,39	95,40	95,00
Contornos dinámicos (Quiceno et al, 2010)	98,00	96,90	97,20

4. CONCLUSIONES

En este artículo se prueba la capacidad del análisis de relevancia para encontrar las bandas de frecuencia más informativas de las TFRs paramétricas, así como la capacidad de los métodos de descomposición lineal bidimensionales tanto supervisados (2DPLS) como no supervisados (2DPCA) para la reducción de dimensión de estas superficies. Para esto se planteó como problema la detección de soplos cardiacos a partir de un conjunto de TFRs calculadas sobre una base de datos de señales FCG. El rendimiento del sistema se probó de acuerdo con la tasa de clasificación obtenida con un clasificador de k-vecinos más cercanos y una meto-

dología de validación cruzada de 10 particiones. El análisis de relevancia utilizado resultó ser efectivo para detectar y eliminar las zonas de las superficies con poca información, lo que permitió recalcular estas representaciones dándole mayor resolución a las zonas relevantes. De esta forma, se permite resaltar las características discriminantes en cada una de las clases (normal y patológica).

Por otro lado, los métodos de descomposición propuestos permiten obtener una representación de cada superficie con una cantidad de puntos notablemente menor tanto en el eje del tiempo como en el eje de la frecuencia, convirtiéndose así en herramientas útiles para tratar bases de datos 2D de gran dimensión. Como se pudo observar, la metodología supervisada presenta mejores resultados tanto en la clasificación como en la reducción de dimensión, debidos a la información adicional que brindan las etiquetas de clase para el entrenamiento.

Como trabajo futuro se plantea la utilización de otras medidas de relevancia propuestas en la literatura, además del uso de diferentes técnicas de análisis de redundancia para mejorar el desempeño de la metodología propuesta. Además, extender esta metodología a otras bases de datos de bioseñales, a otro tipo de representaciones en el plano t-f como las no paramétricas, y a imágenes, con el fin de probar la robustez de la metodología.

5. AGRADECIMIENTOS

Este trabajo se enmarca en el proyecto “Servicio de monitoreo remoto de actividad cardíaca para el tamizaje clínico en la red de telemedicina del Departamento de Caldas", financiado por “Proyectos de investigación e innovación conjuntos entre grupos de trabajo académico, Universidad de Caldas y Universidad Nacional sede Manizales, hacia tercer milenio”. Además los autores quieren agradecer a la Convocatoria de Apoyo a Tesis de Posgrados - DIMA 2010 de la Universidad Nacional de Colombia - Sede Manizales.

6. REFERENCIAS

Ahlstrom, C., Hult, P., Rask, P., Karlsson, J., Nylander, E., Dahlstrom, U., Ask, P., (2006); Feature Extraction for systolic heart murmur classification, Annals of Biomedical Engineering, 34, 1666-1677.

Avendaño-Valencia, L.D., Godino-Llorente, J.I., Blanco-Velasco, M., Castellanos-Dominguez G., (2010); Feature extraction from parametric time-frequency representations for heart murmur detection, Annals of Biomedical Engineering, 38(8), 2716-2732.

Barker, M., Rayens, W., (2003); Partial least squares for discrimination, Journal of Chemometrics, 17(3), 166-173.

Cassidy, M., Penny, W., (2002); Bayesian nonstationary autoregressive models for biomedical signal analysis, IEEE Transactions on Biomedical Engineering, 49(10), 1142-1152.

Chizner, M.A., (2008); Cardiac auscultation: Rediscovering the lost art, Curr. Probl. Cardiol, 33(7), 326–408.

Delgado-Trejos, E., Quiceno-Manrique, A., Godino-Llorente, J., Blanco-Velasco, M., Castellanos-Dominguez, G., (2009); Digital auscultation analysis for heart murmur detection, Annals of Biomedical Engineering, 37(2), 337-353.

Jollife, I.T., (1986); Principal Component Analysis, Springer Verlag.

Kitagawa, G., Gersch, W., (1985); A smoothness priors long AR model method for spectral estimation, IEEE J AC, 30, 57-65.

Poulimenos, A., Fassois, S., (2006); Parametric time-domain methods for non-stationary random vibration modeling and analysis – A critical survey and comparision, Mechanical System and Signal Proccesing, 20(4), 763-816.

Quiceno-Manrique, A., Godino-Llorente, J., Blanco-Velasco, M., Castellanos-Dominguez, G., (2009); Selection of dynamic features based on time frequency representations for heart murmur detection from phonocardiographyc signals, Annals of Biomedical Engineering. 38(1), 118.137.

Sanchez-Giraldo, L., Castellanos-Dominguez, G., (2010); Weighted feature extraction with a functional data extension, Neurocomputing, 73, 1760-1773.

Sejdic, E., Djurovic, I., Jiang, J., (2009); Time-frequency feature representation using energy concentration: An overview of recent advances, Digital signal Processing, 19(1), 153-183.

Tarvainen, M.P., Hiltunen, J.K., Ranta-aho, P.O., Karjalainen, P.A., (2004); Estimation of nonstationary EEG with Kalman smoother approach: An application to event-related synchronization, IEEE Transactions On Biomedical Enginnering, 51(3), 516-524.

Tzallas, A., Tsipouras, M., Fotiadis, D., (2009); Epileptic seizure detection in electroencephalograms using time-frequency analysis, IEEE Transactions on Information Technology in Biomedicine, 13, to be appeared.

Wang, P., Lim, C., Chauhan S., Yong, J., Foo, A., Anantharaman V., (2006); Phonocardiographic signals analysis method using modified hidden Markov model, Annals of Biomedical Engineering, 35(3), 367-374.

Yang, J., Zhang, D., Frangi, A.F., Yu Yang, J., (2004); Two dimensional PCA: A new approach to appearance-based face representation and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(1), 131-137.

Yu, L., Liu, H., (2004); Efficient Feature Selection via Analysis of Relevance and Redundance, Journal of Machine Learning Research, 5, 1205-1224.

Zhang, D., Zhou, Z-H., (2005); Two directional two-dimensional PCA for efficient face representation and recognition, Neurocomputing, 69(1-3), 224-231.

Zuo, W., Zhang, D., Wang, K., (2006); An assembled matrix distance metric for 2DPCA-based image recognition, Pattern Recognition Letters, 21, 210-216.