REDUCCIÓN DE DIMENSIÓN PARA EL RECONOCIMIENTO AUTOMÁTICO DE PATRONES SOBRE BIOSEÑALES

Dimensionality Reduction for Automatic Pattern Recognition on Biosignals

RESUMEN
Se presentan resultados para una metodología de reducción de dimensión, mediante la comparación entre técnicas de selección y extracción de características, que tiene como aporte fundamental la identificación de las condiciones de aplicación de cada una de las técnicas. Las pruebas experimentales se realizan sobre una base de datos de voz. Los resultados reflejan que la capacidad de reducción y clasificación de las técnicas de selección es usualmente superior a las de extracción, pero la naturaleza estadística de los datos tiene gran influencia sobre los métodos de reducción.

PALABRAS CLAVES: Análisis de componentes, extracción de características, selección de características.

ABSTRACT
Results for a dimension reduction methodology based on the comparison of several feature selection and feature extraction techniques are presented. The main contribution of this work is the identification of the application conditions for each technique. Experimental tests were made employing a voice disorders database. Results for reduction and classification expose better performance for feature selection than for feature extraction; nevertheless, the statistical nature of data influences the feature reduction methods.

KEYWORDS: *Component analysis, feature extraction, feature selection.*

GENARO DAZA SANTACOLOMA
Ingeniero Electrónico, M. Sc.
Grupo de control y procesamiento digital de señales (GC&PDS)
Universidad Nacional de Colombia sede Manizales
gdazas@unal.edu.co

JOSÉ SOTO MEJÍA
Fisico, Ph. D.
Profesor titular
Ingeniería Industrial
Universidad Tecnológica de Pereira
jomejia@utp.edu.co

GERMÁN CASTELLANOS DOMÍNGUEZ
Ingeniero de Radiocomunicaciones, Ph.D.
Profesor
Universidad Nacional de Colombia sede Manizales
cgcastellanosd@unal.edu.co

1. INTRODUCCIÓN

El reconocimiento de patrones tiene por objetivo la clasificación de observaciones en un determinado número de clases [1], [2], [3]. Cuando el número de variables empleadas para medir las observaciones o *la dimensión del espacio de características* es alta, existe un gran interés en reducirla [4]. Uno de los problemas de tener conjuntos de alta dimensión es que en muchos casos no todas las variables medidas son importantes para la comprensión del fenómeno de interés [5], porque algunas son redundantes y otras no discriminan entre clases. La reducción de dimensión puede realizarse identificando las variables que no contribuyen a la tarea de clasificación, para eliminarla, obteniendo un conjunto de m características de las p disponibles. Este proceso se conoce como ***selección de características*** en el espacio original de medida o simplemente selección de características. Una segunda aproximación es hallar una *transformación* de las p características a un nuevo espacio de menor dimensión, conocido como ***extracción de características*** [6].
En este artículo se presentan los resultados de comparar las siguientes **técnicas de extracción**: análisis de componentes principales (PCA), análisis factorial, PCA probabilístico (PPCA) y análisis de componentes independientes (ICA).

En Particular, para las **técnicas de selección** se implementaron tres algoritmos subóptimos de búsqueda de características: selección secuencial hacia adelante (SFS), selección secuencial hacia atrás (SBS) y selección secuencial flotante (SFFS); éstos se estructuraron con base en seis funciones de costo: los criterios de selección con base en matrices de dispersión, $\mathcal{J}_1$, $\mathcal{J}_2$, $\mathcal{J}_3$ y $\mathcal{J}_4$, análisis de varianza multivariado (MANOVA), y *wrapper* por medio del error de entrenamiento de un clasificador Bayesiano. La sección 2, presenta una síntesis de los métodos matemáticos utilizados. La sección 3 describe el marco experimental utilizado. En la sección 4 se presentan los resultados al aplicar las diferentes técnicas multivariadas y en la sección 5 la discusión y conclusiones.

Fecha de Recepción: 23 Agosto de 2007
Fecha de Aceptación: 29 Noviembre de 2007

2. FUNDAMENTOS TEÓRICOS DE LAS TÉCNICAS DE EXTRACCIÓN Y SELECCIÓN

2.1. Extracción de características

2.1.1. Análisis de componentes principales

Dada una matriz de datos, se busca la posibilidad de representar adecuadamente la información, con un número menor de variables que son construidas como combinaciones lineales de las originales [7].
El modelo de transformación está dado por: $\mathbf{Z} = \mathbf{XW}$, donde $\mathbf{Z}$ es la matriz $(n \times p)$ de observaciones en el espacio transformado, $\mathbf{X}$ es la matriz $(n \times p)$ de observaciones en el espacio original y $\mathbf{W}$ es la matriz $(p \times m)$ de transformación que corresponde a los m vectores propios de la matriz de covarianza o correlación asociados a los m valores propios más grandes.
El Algoritmo 1 resume el procedimiento para PCA.

Algoritmo 1. Cálculo de PCA

1: Centralizar la matriz de datos $\mathbf{X}$.
2: Obtener la matriz de covarianza $\mathbf{S} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$
3: Calcular los valores propios de la matriz $\mathbf{S}$ y sus respectivos vectores propios.
4: Ordenar de forma descendente los valores propios.
5: Proyectar los datos sobre las direcciones principales (vectores propios ordenados).

2.1.2. Análisis factorial

El objetivo principal del análisis factorial es explicar un conjunto de variables observadas a partir de combinaciones lineales de un conjunto menor de factores que son aleatorios. Los factores son construcciones subyacentes, no medibles, que generan a las variables observadas. Si estas variables observadas presentan algún grado de correlación, es posible inferir que existe un conjunto menor de variables que explican el fenómeno con menos redundancia [8], [7], [9].
La matriz de variables aleatorias observadas puede ser expresada como la combinación lineal de factores comunes independientes, más un termino de error que representa la parte de la dispersión que varía de forma única para cada variable. El modelo para un vector observación es el siguiente:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon} \tag{1}$$

donde, $\mathbf{f}$ es un vector $(m \times 1)$ de variables latentes, que se asume sigue una distribución $N_m(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\Lambda}$ es una matriz $(p \times m)$ de constantes desconocidas $(m < p)$, denominada matriz de carga y describe cómo los factores afectan a las variables originales, $\boldsymbol{\varepsilon}$ es un vector $(p \times 1)$ de perturbaciones no observadas, se asume que tiene distribución $N_p(0, \boldsymbol{\psi})$ donde $\boldsymbol{\psi}$ es diagonal y las perturbaciones no están relacionadas con los factores. Por tanto, $\boldsymbol{\mu}$ es la media de las variables $\mathbf{x}$ que tienen distribución normal, entonces $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $\boldsymbol{\Sigma}$ es la matriz de covarianzas de $\mathbf{x}$,

$$\begin{aligned} \boldsymbol{\Sigma} &= \operatorname{cov}(\mathbf{x}) = \operatorname{cov}(\boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon}) \\ \boldsymbol{\Sigma} &= \operatorname{cov}(\boldsymbol{\Lambda}\mathbf{f}) + \operatorname{cov}(\boldsymbol{\varepsilon}) \\ \boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\operatorname{cov}(\mathbf{f})\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\psi} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\psi} \end{aligned} \tag{2}$$

Esto se puede interpretar como: la varianza observada es la suma de las variabilidades común y específica.
Además, la ecuación (1) se reescribe unificando el modelo para un conjunto de observaciones como:

$$\mathbf{X} = \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{F}\boldsymbol{\Lambda}^{\mathrm{T}} + \mathbf{E} \tag{3}$$

El Algoritmo 2 presenta el método del factor principal.

Algoritmo 2. Análisis factorial – Método del factor principal

1: Partir de una estimación inicial de $\hat{\boldsymbol{\Lambda}}_{\hat{k}}$ o de $\hat{\boldsymbol{\psi}}$ (para la iteración $\hat{k} = 1$)¨, a través de,

$$\hat{\boldsymbol{\psi}}_{\hat{k}} = \operatorname{diag}\left\{\mathbf{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^{\mathrm{T}}\right\} \tag{4}$$

2: Calcular la matriz cuadrada y simétrica $\mathbf{Q}_{\hat{k}} = \mathbf{S} - \hat{\boldsymbol{\psi}}_{\hat{k}}$
3: Obtener la descomposición espectral de $\mathbf{Q}_{\hat{k}}$,

$$\mathbf{Q}_{\hat{k}} = \mathbf{H}_{1\hat{k}}\mathbf{G}_{1\hat{k}}\mathbf{H}_{1\hat{k}}^{\mathrm{T}} + \mathbf{H}_{2\hat{k}}\mathbf{G}_{2\hat{k}}\mathbf{H}_{2\hat{k}}^{\mathrm{T}} \tag{5}$$

4: Tomar $\hat{\boldsymbol{\Lambda}}_{\hat{k}+1} = \mathbf{H}_{1\hat{k}}\mathbf{G}_{1\hat{k}}^{1/2}$ y retornar al paso 1. Iterar hasta que $\left\|\boldsymbol{\Lambda}_{\hat{k}+1} - \boldsymbol{\Lambda}_{\hat{k}}\right\| < \xi$

En el Algoritmo 2, $\mathbf{G}_{1\hat{k}}$ contiene los m mayores valores propios de $\mathbf{Q}_{\hat{k}}$ y $\mathbf{H}_{\hat{k}}$ sus vectores propios. Se elige m de manera que los restantes valores propios contenidos en $\mathbf{G}_{2\hat{k}}$ sean todos pequeños y de tamaño similar.

2.1.3. Análisis probabilístico de componentes principales

En general, PCA no contempla un modelo probabilístico para los datos observados, pero es posible asumir la densidad de probabilidad de los datos durante la estimación; esquema que se conoce como PPCA y tiene como principales ventajas [10]:

- Permite obtener las proyecciones para los componentes principales cuando hay datos perdidos.
- Puede utilizarse como modelo general de densidad Gaussiano, donde los estimados de máxima verosimilitud para los parámetros asociados con la matriz de covarianza, pueden calcularse eficientemente a partir de los componentes principales. Esto es útil en

reducción de dimensión, sistemas de clasificación y detección de anomalías.

Es posible determinar las direcciones principales de un conjunto de datos observados a través de la estimación de los parámetros de máxima verosimilitud en un modelo de variables latentes,

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \quad (6)$$

donde la distribución de las perturbaciones no observadas $\boldsymbol{\varepsilon}$ se asume isotrópica (a diferencia del análisis factorial en el cual la matriz de varianzas de $\boldsymbol{\varepsilon}$ es en general diagonal),

$$\boldsymbol{\varepsilon} \sim N_p\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \quad (7)$$

El Algoritmo 3 presenta una síntesis de PPCA.

Algoritmo 3. Cálculo de PPCA

1: Calcular la descomposición en valores y vectores propios de $\mathbf{S}$

2: Calcular el estimador de máxima verosimilitud σ^2_{ML} en la forma:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{p-m}\sum_{j=m+1}^{p}\lambda_j \quad (8)$$

3: Calcular la matriz de parámetros $\mathbf{W}_{\mathrm{ML}}$ usando,

$$\mathbf{W} = \mathbf{V}_m\left(\mathbf{U}_m - \sigma^2\mathbf{I}\right)^{1/2}\mathbf{H} \quad (9)$$

4: Calcular la transformación de reducción de dimensión de los datos $\mathbf{x}$ empleando la *media posterior*

$$\langle \mathbf{z}_i \rangle = \mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\left(\mathbf{x}_i - \boldsymbol{\mu}\right) \quad (10)$$

empleando,

$$p\left(\mathbf{z}\middle|\mathbf{x}\right) = \left(2\pi\right)^{-m/2}\left|\sigma^{-2}\mathbf{M}\right|^{1/2} \times \exp\left[-\frac{1}{2}\left\{\mathbf{z} - \mathbf{M}^{-1}\mathbf{W}\left(\mathbf{x}-\mu\right)^{\mathrm{T}}\right\} \times \left(\sigma^{-2}\mathbf{M}\right)\left\{\mathbf{z} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\left(\mathbf{x}-\mu\right)\right\}\right] \quad (11)$$

y,

$$\sigma^2\mathbf{M}^{-1} = \sigma^2\left(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1} \quad (12)$$

$\mathbf{V}_m$ es $(p\times m)$ y contiene los vectores propios de $\mathbf{S}$. $\mathbf{U}_m$ es una matriz diagonal $(m\times m)$ con los respectivos valores propios. $\mathbf{H}$ es una matriz $(m\times m)$ arbitraria ortogonal de rotación.

2.1.4. Análisis de componentes independientes

Es un método para encontrar variables latentes a partir de datos multivariados, que busca componentes estadísticamente independientes y no gaussianos [11].
El modelo ICA libre de ruido está dado por:

$$\mathbf{x} = \mathbf{W}\mathbf{z} \quad (13)$$

donde los componentes independientes z_k, son mutuamente independientes, si y sólo si:

$$p\left(z_1, z_2, \ldots, z_m\right) = p\left(z_1\right)p\left(z_2\right)\cdots p\left(z_m\right) \quad (14)$$

El modelo ICA es generador, describe cómo los datos provienen de la mezcla de componentes z_k.

En la estimación *uno a uno* de los componentes independientes, se halla un vector $\mathbf{b}$, tal que la combinación $\mathbf{b}^{\mathrm{T}}\mathbf{x}$ sea igual a uno de los componentes independientes $\pm\mathbf{z}_k$. Este procedimiento se itera para hallar varias componentes independientes. Debido a que en la labor de reducción de dimensión se parte de la hipótesis que la cantidad de componentes encontrados es menor que la cantidad de variables observadas, es suficiente hallar sólo algunas componentes (m).

El Algoritmo 4 presenta un procedimiento de punto fijo para la estimación de varios componentes del modelo ICA con base en la kurtosis.

Algoritmo 4. Cálculo de ICA

1: Seleccionar el número m de componentes a estimar. Establecer el contador de iteraciones en $\hat{k} = 1$.

2: Tomar un vector aleatorio $\mathbf{b}_{\hat{k}}$ de norma 1.

3: Hacer una iteración de:

a: Realizar la normalización de la matriz de datos $\mathbf{X}$ (media cero y matriz de covarianza identidad $\mathbf{I}$).

b: Hacer

$$\mathbf{b}_{\hat{k}} = E\left\{\mathbf{x}\left(\mathbf{b}_{\hat{k}-1}^{\mathrm{T}}\mathbf{x}\right)^3\right\} - 3\mathbf{b}_{\hat{k}-1} \quad (15)$$

c: Divida $\mathbf{b}_{\hat{k}}$ entre su norma.

d: Si $\left|\mathbf{b}_{\hat{k}}^{\mathrm{T}}\mathbf{b}_{\hat{k}-1}\right|$ no es un valor próximo a 1, hacer $\hat{k} = \hat{k} + 1$ y regresar al paso 2. En otro caso el algoritmo converge y la salida es $\mathbf{b}_{\hat{k}}$

4: Hacer la ortogonalización

$$\mathbf{b}_{\hat{k}} \leftarrow \mathbf{b}_{\hat{k}} - \sum_{k=1}^{\hat{k}-1}\left(\mathbf{b}_{\hat{k}}^{\mathrm{T}}\mathbf{b}_k\right)\mathbf{b}_k \quad (16)$$

5: Normalizar $\mathbf{b}_{\hat{k}}$ dividiéndolo por su norma.

6: Si $\mathbf{b}_{\hat{k}}$ no ha convergido regresar al paso 3.

7: Hacer $\hat{k} = \hat{k} + 1$. Si $\hat{k} < m$ regresar al paso 2.

2.2. Selección de características

La selección de características se estructura con base en dos etapas básicas: el algoritmo de búsqueda de variables y la función de evaluación para dicha búsqueda.

2.2.1. Selección secuencial hacia adelante

Es una técnica de búsqueda abajo-arriba. Selecciona primero la mejor variable según algún criterio $\mathcal{J}$, luego se combina la variable original con cada una de las variables restantes, entonces, se busca la pareja que aporta el mayor valor de evaluación y se escoge como nuevo con-

junto de partida. A continuación se combina esta pareja con cada una de las variables restantes, formando ternas, se selecciona la terna que dé un mayor valor en el criterio de evaluación. El proceso se repite una y otra vez en la misma forma. La búsqueda se detiene cuando un conjunto de más variables no mejore los resultados de la función de costo para un conjunto de menos variables.

2.2.2. Selección secuencial hacia atrás

Al igual que la técnica anterior, la idea es construir conjuntos diferentes iteración tras iteración, con la diferencia que ahora se inicia con el conjunto completo de características de dimensión *p*, y en cada iteración se remueve una variable. La variable que se elimina es aquella que al no estar presente en el subconjunto a evaluar, hace que la función de costo reporte el mayor valor entre todos los subconjuntos evaluados en la misma iteración. El algoritmo se detiene cuando el valor de $\mathcal{J}$ no supera cierta cota preestablecida. Entonces, se selecciona el último subconjunto que al ser evaluado haya superado el umbral.

2.2.3. Selección secuencial flotante

A diferencia de los casos anteriores, este procedimiento permite tanto adicionar como eliminar características al subconjunto ya elegido. Básicamente, en una iteración $\hat{k}$, se adiciona la característica que maximice el criterio $\mathcal{J}$, posteriormente se elimina del subconjunto actual aquella variable que más reduzca el valor de $\mathcal{J}$, si es la última variable agregada, entonces el conjunto no se modifica y se adiciona una nueva variable; en caso contrario se remueve la característica del subconjunto y se continúan removiendo características siempre y cuando $\mathcal{J}$ no decrezca. Luego se agrega nuevamente una característica y se continúa el proceso. La búsqueda se detiene cuando al incrementar o decrementar características el valor de $\mathcal{J}$ no es mejorado.

2.2.4. Funciones de evaluación con base en matrices de dispersión

Se han desarrollado algunas medidas que basan su funcionamiento en matrices entre-clases e intra-clase, las cuales son matrices de dispersión. Sea $\mathbf{S}$ la matriz de covarianza estimada y sea $\mathbf{S}_l$ la matriz de covarianza estimada de la clase l [6],

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}} \tag{17}$$

$$\mathbf{S}_l = \frac{1}{n_l}\sum_{i=1}^{n}\mathcal{C}_{li}\left(\mathbf{x}_i - \bar{\mathbf{x}}_l\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}_l\right)^{\mathrm{T}} \tag{18}$$

donde,

$$\mathcal{C}_{li} = \begin{cases} 1, \text{ si } \mathbf{x}_i \in \omega_l \\ 0, \text{ otro caso} \end{cases} \tag{19}$$

siendo ω_l la etiqueta de la clase l y $n_l = \sum_{i=1}^{n}\mathcal{C}_{li}$.

Además, $\bar{\mathbf{x}}_l$ es la estimación de la media de la clase ω_l y $\bar{\mathbf{x}}$ es la estimación de la media para un número dado de K clases.

Sea $\mathbf{S}_W$ la matriz de dispersión intra-clase,

$$\mathbf{S}_W = \sum_{l=1}^{K}\frac{n_l}{n}\mathbf{S}_l \tag{20}$$

y sea también, $\mathbf{S}_B$ la matriz de covarianza entre clases estimada,

$$\mathbf{S}_B = \sum_{l=1}^{K}\frac{n_l}{n}\left(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}\right)^{\mathrm{T}} \tag{21}$$

tal que se cumple, $\mathbf{S}_W + \mathbf{S}_B = \mathbf{S}$.

La distancia promedio entre clases puede definir una medida de separación entre dos conjuntos de datos, usando una distancia euclidiana cuadrática se puede tener,

$$\mathcal{J}_1 = \mathrm{tr}\left\{\mathbf{S}_W + \mathbf{S}_B\right\} = \mathrm{tr}\left\{\mathbf{S}\right\} \tag{22}$$

Sin embargo, el criterio $\mathcal{J}_1$ no es el más adecuado para la selección de características, porque simplemente refleja la varianza total, y no analiza los datos de cada clase de forma separada. El propósito de los criterios es hallar un conjunto de variables para el cual la dispersión intra-clase sea pequeña mientras la dispersión entre-clases sea grande en algún sentido. Usualmente se emplean los siguientes criterios,

$$\mathcal{J}_2 = \mathrm{tr}\left\{\mathbf{S}_W^{-1}\mathbf{S}_B\right\} \tag{23}$$

$$\mathcal{J}_3 = \frac{|\mathbf{S}|}{|\mathbf{S}_W|} \tag{24}$$

$$\mathcal{J}_4 = \frac{\mathrm{tr}\left\{\mathbf{S}_B\right\}}{\mathrm{tr}\left\{\mathbf{S}_W\right\}} \tag{25}$$

2.2.5. Función de evaluación con base en MANOVA

En MANOVA, el criterio de evaluación estadístico es la separabilidad entre clases, que se realiza mediante una prueba de hipótesis sobre la igualdad o desigualdad de los vectores de promedios entre las clases. Se asume que los datos se generan con base en el siguiente modelo:

$$\mathbf{x}_{kj} = \mathbf{m}_k + \varepsilon_{kj},\ \mathbf{m}_k = \mathbf{m} + \boldsymbol{\alpha}_k \tag{26}$$

siendo j la observación y k la clase, $\mathbf{m}_k$ es el vector de medias para cada clase y ε_{kj} es la respectiva perturbación del modelo, $\mathbf{m}$ es la media global de las clases y $\boldsymbol{\alpha}_k$ es la perturbación sobre esta media global.

La comparación de los vectores de medias de las k clases para encontrar diferencias significativas, se realiza mediante la prueba hipótesis:

$$\begin{aligned} &H_0 : \mathbf{m}_1 = \mathbf{m}_2 = \cdots = \mathbf{m}_K \\ &H_1 : \exists \text{ al menos un par } \mathbf{m}_k \neq \mathbf{m}_i; \forall\ k, i \in \{1, \ldots, K\} \end{aligned} \tag{27}$$

donde K es el número de clases.

La estadística de Wilks es comúnmente usada al interior de MANOVA, para probar la hipótesis H_0, que corresponde a la relación de verosimilitud dada por:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} \quad (28)$$

la cual es conocida como Λ de Wilks. Siendo **H** la matriz de hipótesis que puede entenderse como una medida de dispersión entre la media de las clases, mientras que la matriz de error **E** se relaciona con la medida de dispersión entre las observaciones para cada clase. La hipótesis H_0 se rechaza si la dispersión entre los patrones es mayor que la dispersión de las observaciones dentro de los patrones, y así, $\Lambda \in [0,1]$ tiende a cero. Por otro lado, el Λ de Wilks puede ser similar a un estadístico F, pero de manera inversa. Un valor grande del estadístico F rechaza H_0 [8].

2.2.6. Función de evaluación wrapper

La función de evaluación tipo *wrapper* es aquella en la que se evalúa directamente el rendimiento de clasificación, en el caso particular, la función de costo $\mathcal{J}$ está dada por el porcentaje de acierto en entrenamiento de un clasificador bayesiano sobre distribuciones gaussianas, lo que corresponde a una función cuadrática.

3. MARCO EXPERIMENTAL

3.1. Base de datos[1]

Esta base de datos pertenece a la *Universidad de Las Palmas de Gran Canaria* y contiene grabaciones de audio de 180 individuos (hombres y mujeres), repartidas entre 93 pacientes con disfonía y 87 pacientes sin anomalías de voz. El contenido de las grabaciones corresponde a la fonación de la vocal /*a*/ del idioma español, de forma sostenida y no susurrada. El formato de grabación consiste en audio digital con una frecuencia de muestreo de 22050Hz y resolución de 16 bits. La caracterización de las señales se llevó a cabo con base en 4 dominios que se emplean frecuentemente en el procesamiento de señales de voz: dominio temporal (18 variables), dominio espectral (53 variables), dominio cepstral (42 variables) y dominio del modelo inverso (31 variables). En este sentido, sobre cada vocal son calculadas en total 144 características.

3.2. Preproceso de datos

Inicialmente, la base de datos fue sometida a un preproceso que consistió en identificar aquellas variables que contienen datos no convergentes o errores de medida, dichas variables son eliminadas del conjunto inicial de características. Seguidamente, se continua con la identificación de datos atípicos; en este caso en particular, debido al número reducido de observaciones con que se cuenta, no se eliminan las observaciones identificadas como atípicas, sino que se buscan y eliminan las variables que poseen más de un 10% de valores atípicos. Finalmente, la etapa de preproceso termina con la verificación univariada de la distribución normal de las variables con base en la prueba de *Kolmogorov-Smirnov* con un nivel de significancia de 0,05. La prueba de normalidad se realiza para cada una de las clases, aquellas variables que no posean distribución normal se eliminan de todas las clases.

3.3. Extracción o selección de características

El conjunto de variables preprocesadas obtenido a partir del procedimiento anteriormente descrito se utiliza como entrada para cada una de las técnicas de extracción y selección presentadas en la Sección 2. Los conjuntos de variables resultantes de cada una de las técnicas aplicadas son utilizados en el clasificador final.

3.4. Clasificación

Como algoritmo de decisión se emplea un clasificador sencillo, en particular un clasificador bayesiano sobre distribuciones gaussianas. Con el objetivo de evaluar el desempeño de cada uno de los conjuntos de variables extraídas o seleccionadas, se compara la tasa de errores en validación y la desviación estándar de la tasa de errores de validación. Para la estimación de los errores de validación se emplea la estrategia de validación cruzada *leave-M-out*, la cual consiste en generar L conjuntos, que corresponden a particiones aleatorias del conjunto de n observaciones en pares de entrenamiento-validación donde se retienen M observaciones para validar (se entrena con $n-M$ observaciones). Para este trabajo $L=100$, $n-M=70\%$ y $M=30\%$.

4. RESULTADOS

La Tabla 1 presenta los resultados comparativos de cada una de las técnicas de reducción, bajo los criterios de: precisión del sistema en la etapa de clasificación y capacidad de reducción.

Técnica	Error medio de validación [%]	Desviación del error	Porcentaje de reducción del espacio de características
PCA	16.01	0.0452	69.23
PPCA	16.01	0.0452	69.23
FA	16.13	0.0491	30.77
ICA	15.68	0.0436	76.92
SFS-J1	24.78	0.0542	0
SFS-J2	24.78	0.0542	0
SFS-J3	24.78	0.0542	0

[1] *Nota:* También se realizan pruebas sobre otras bases de datos de señales electrocardiográficas y señales de voz, pero por cuestión de espacio no es posible presentar esos resultados acá, ver [1].

SFS-J4	17.77	0.0471	96.15
SFS-MANOVA	17.77	0.0471	96.15
SFS-*wrapper*	10.54	0.0396	76.92
SBS-J1	24.78	0.0542	0
SBS-J2	24.78	0.0542	0
SBS-J3	24.78	0.0542	0
SBS-J4	17.77	0.0471	96.15
SBS-MANOVA	26.09	0.0512	96.15
SBS-*wrapper*	47.09	0.0591	96.15
SFFS-J1	24.78	0.0542	0
SFFS-J2	24.78	0.0542	0
SFFS-J3	24.78	0.0542	0
SFFS-J4	17.77	0.0471	96.15
SFFS-MANOVA	17.77	0.0471	96.15
SFFS-*wrapper*	10.54	0.0393	76.92

Tabla 1. Comparación de técnicas de reducción de dimensión.

5. DISCUSIÓN Y CONCLUSIONES

En general, todas las técnicas de extracción de características tienen un desempeño similar, tanto desde el punto de vista de separabilidad, como de reducción. Sin embargo, el análisis factorial tuvo el más bajo desempeño en cuanto a capacidad de reducción, dificultando la posibilidad de interpretar sus resultados; debido a que la explicación de 3 o más factores es una labor muy compleja, porque se desconocen las relaciones causa efecto de las variables y los factores en el experimento. Por otra parte, los algoritmos de búsqueda **SFS** y **SFFS** reportan mejores resultados de clasificación y capacidad de reducción en comparación con el algoritmo **SBS.** Lo anterior, es ocasionado al interior de los algoritmos de búsqueda por errores en la estimación de las funciones de costo, situación que suele ocurrir cuando la dimensión original del espacio de variables es muy grande con relación al número de observaciones, o cuando los datos están altamente correlacionados; estas dos situaciones, son altamente factibles cuando la búsqueda inicial se realiza sobre el conjunto completo de características, como ocurre en el algoritmo **SBS**. De acuerdo a lo anterior, no es adecuado emplear el algoritmo **SBS** cuando el número de observaciones por clase es menor que el número de variables.

En otro sentido, la técnica de selección de características tipo *wrapper* es adecuada, siempre y cuando el tipo de clasificador que se emplee en el sistema completo de reconocimiento automático de patrones, sea del mismo tipo que el clasificador usado como función de costo al interior del algoritmo de selección, porque las variables seleccionadas son totalmente dependientes del hiperplano de separación construido por el clasificador en la etapa de selección. En conclusión, el trabajo realizado prueba una metodología de **reducción** de dimensión, estructurada y desarrollada con clara identificación de los requisitos y restricciones que exige cada una de las técnicas utilizadas. Se implementaron 3 esquemas de búsqueda heurística: SFS, SBS y SFFS; cada uno de los esquemas se configuró empleando 6 funciones de costo diferentes, correspondientes a medidas univariadas y multivariadas que permiten determinar el grado separación entre clases. También se implementaron 4 técnicas de **extracción de características** para reducción de dimensión. La diferencia entre estas técnicas consiste en la capacidad de interpretar o analizar los componentes resultantes después de la reducción.

6. AGRADECIMIENTOS

Los autores desean agradecer la colaboración del Ph. D. Jesús B. Alonso, profesor de la Universidad de las Palmas de Gran Canaria por facilitarnos la base de datos para realizar las pruebas. Este trabajo se realiza dentro del proyecto *"Identificación automatizada de hipernasalidad en niños con LPH por medio del análisis acústico del habla"*, financiado por la Vicerrectoría de Investigaciones de la Universidad Nacional de Colombia.

7. BIBLIOGRAFÍA

[1] G. Daza-Santacoloma, "Metodología de reducción de dimensión para sistemas de reconocimiento automático de patrones sobre bioseñales," M. Sc. Thesis, Depto. IEEyC, Universidad Nacional de Colombia, Manizales, 2006.

[2] A. K. Jain, R. P. W. Duin, and J. Mao, "Satatistical patern recognition: A Review," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[3] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd ed. San Diego, CA, USA: ELSEVIER Academic Press, 2003.

[4] H. Brunzell and J. Eriksson, "Feature reduction for classification of multidimensional data," *Pattern Recognition*, vol. 33, pp. 1741–1748, 2000.

[5] I. K. Fodor, "A survey of dimension reduction techniques," Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, USA, Tech. Rep., 2002.

[6] A. R. Webb, *Statistical Pattern Recognition*, 2nd ed. Indianapolis, IN, USA: John Wiley & Sons, 2002.

[7] D. Peña, *Análisis de datos multivariantes*, C. F. Madrid, Ed. Madrid, España: McGraw-Hill, 2002.

[8] A. C. Rencher, *Methods of multivariate analysis*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2002.

[9] I. Doltsinis, Ed., *Stochastic analysis of multivariate systems in computational mechanics and engineering*, 1st ed. Barcelona, Spain: CIMNE, 1999.

[10] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, S. Haykin, Ed. New York, NY, USA: John Wiley & Sons, Inc, 2001.