REDUCCIÓN DE DIMENSIONES PARA CLASIFICACIÓN DE DATOS MULTIDIMENSIONALES USANDO MEDIDAS DE INFORMACIÓN

RESUMEN
El enorme desarrollo tecnológico ha creado un concepto de información con áreas extensas de aplicación para tareas referentes al entrenamiento de sistemas automáticos. Este trabajo propone una metodología basada en el análisis de componentes independientes (ICA), que incluye el uso de medidas de información, para realizar reducción de dimensiones en conjuntos de datos multidimensionales. La metodología usa un principio de relevancia, con el fin de hallar la representación reducida mientras se conserva la estructura relacionada a la información inicial. La metodología se compara y se conjuga con un sistema básico de selección de características cuya función de evaluación usa la medida de entropía logrando mejores resultados en la clasificación.

PALABRAS CLAVES: Reducción de dimensiones, Entropía, Información, ICA.

ABSTRACT
The enormous technological development has created a concept of information that has a big field of action for tasks related with the training of automatic systems. This paper proposes a methodology that is based in the Independent Component Analysis (ICA) that includes the use of information measures to perform a dimensionality reduction in multidimensional data sets. The methodology uses a principle of relevance that tries to find a reduced representation while the structure related with the initial information is maintained. The methodology is compared and mixed with a basic system of feature selection which evaluation function uses the measure of entropy. Results show that the classification rate increase.

KEYWORDS: Dimensionality reduction, Entropy, Information, ICA.

[1]**ANDRÉS J. HERNÁNDEZ**
Ingeniero Electrónico
Universidad Nacional de Colombia
andresjulian.hernandez@gmail.com

[1]**EDILSON DELGADO TREJOS**
Ingeniero Electrónico, M.Sc.
Docente ocasional
Universidad Nacional de Colombia sede Manizales
edelgadot@unal.edu.co

[2]**JORGE RIVERA PIEDRAHITA**
Ingeniero Electrónico, M.Sc.
Profesor Asistente
Universidad Tecnológica de Pereira
j.rivera@utp.edu.co

[1]**GERMÁN CASTELLANOS**
Ing. en Telecomunicaciones, Ph.D.
Profesor asociado
Universidad Nacional de Colombia sede Manizales
gcastell@telesat.com.co

[1]Grupo de desarrollo en electrónica y robótica.
[2]LIDER.

1. INTRODUCCIÓN

La reducción de dimensiones es frecuentemente usada como una etapa de preproceso en el entrenamiento de sistemas, y consiste en escoger un subconjunto de variables, de tal manera, que el espacio de características quede óptimamente reducido de acuerdo a un criterio de evaluación, cuyo fin es distinguir el subconjunto que representa mejor el espacio inicial de entrenamiento. Como cada característica que se incluye en el análisis, puede incrementar el costo y el tiempo de proceso de los sistemas, hay una fuerte motivación para diseñar e implementar sistemas con pequeños conjuntos de características. Sin dejar de lado, que al mismo tiempo, hay una opuesta necesidad de incluir un conjunto suficiente de características para lograr un alto rendimiento. Esto ha motivado el desarrollo de una variedad de técnicas que ofrecen encontrar el subconjunto óptimo a partir de un conjunto inicial de características. Sin embargo, muchas de esas técnicas pueden solamente manipular ciertos tipos de datos [1].
El uso de la teoría de información para la selección de características relevantes ha sido explorada mediante varios trabajos, presentándose como una metodología alternativa al momento de analizar conjuntos de datos multidimensionales con variables redundantes que, además de afectar la precisión de los clasificadores, son difíciles de distinguir [1]. Debido a esto, se han generado varios modelos basados en métricas de información para reducir dimensiones y mejorar la tasa de acierto en la clasificación [2].
En el reconocimiento de patrones, y en general, en problemas de clasificación, métodos como el Análisis de Componentes Principales (PCA) y el Análisis de Componentes Independientes (ICA) han sido ampliamente estudiados [3]. Estos métodos realizan un mapeo del espacio de características a un espacio transformado que puede ser de menor dimensión. PCA no es adecuado cuando la dirección de máxima variabilidad no corresponde a la dirección de máxima separabilidad de los datos [4]. ICA representa los datos multivariantes en componentes estadísticamente independientes, y para este trabajo, se incluye un criterio de optimización basado en medidas de información, buscando que la estructura informativa en los datos transformados se mantenga y que la tasa de acierto en la clasificación se incremente. En [5] los autores utilizan ICA para la selección de un subconjunto de componentes que apunten al mayor poder discriminatorio dentro de un contexto de clasificación de rostros. Se utiliza el método ICA partiendo de un gran subconjunto de características, además se realiza una reducción preliminar del espacio

Fecha de Recepción: 31 Agosto de 2006
Fecha de Aceptación: 30 Noviembre de 2006

de busqueda mediante PCA. En [6] se utiliza ICA para la selección de características dentro para un sistema de análisis de texturas, el espacio inicial de características se reduce en una primera instancia mediante PCA, después se aplica a los componentes obtenidos un filtro basado en el principio de ICA para después utilizar una maquina de soporte vectorial para la clasificación. Los resultados de [6] demuestran un rendimiento superior del método propuesto sobre un esquema que solo utiliza PCA para la selección de características. ICA también se ha utilizado para la reducción dimensional con el objetivo de detectar enfisemas en imágenes médicas [7], con resultados favorables sobre esquemas de selección de caracterísitacas similares [7].
Particularmente en este trabajo, se realiza tanto la comparación como la conjugación entre un procedimiento que usa el análisis de componentes independientes con función de evaluación basado en negentropía y un algoritmo básico de selección de características cuya función de evaluación usa la entropía como medida de aceptación o rechazo. Para la validación de los procedimientos, se usa un conjunto de parámetros que caracterizan voces de niños que evidencian hipernasalidad y registros de voces normales provenientes de niños con edades similares (45 registros por clase).

2. MODELO DE REPRESENTACIÓN ECONÓMICA

La representación económica, en el reconocimiento de patrones, ha surgido debido a la necesidad de dar solución a los tres problemas siguientes: a) Reducir el costo y la complejidad de los clasificadores, b) Mejorar la precisión del modelo de clasificación y c) Mejorar la visualización y la interpretación de estructuras contenidas en el espacio de representación [2].

2.1. Reducción de Dimensionalidad

Definición 2.1.1 (Reducción de dimensión): Sea el espacio inicial de análisis, o de entrenamiento, dado por el conjunto $\mathbf{X}=[\mathbf{\Phi}_1,\mathbf{\Phi}_2,\ldots,\mathbf{\Phi}_p]$, de manera que cada columna $\mathbf{\Phi}_i$ está asociada a una característica del *conjunto de características iniciales* ξ de dimensión *p*. Se llama *conjunto reducido de características* ξ de dimensión *q*, con *q<p*, tan pequeña como sea posible, asociado con un espacio de representación reducida $\mathbf{Z}$, que satisface las siguientes condiciones: hay una función o un mapeo de reducción de variables y existe una reconstrucción o mapeo suave, no singular que debe contener aproximadamente toda la información del espacio inicial de análisis.

2.2. Relevancia

La diferencia entre el número de elementos del conjunto inicial de características y el conjunto resultante después de utilizar un procedimiento de selección de características, es determinada por el criterio de evaluación o *criterio de relevancia*. En [8], las características se clasifican en tres categorías: características fuertemente relevantes, débilmente relevantes y características irrelevantes. Un subconjunto óptimo debe incluir todas las características fuertemente relevantes, ninguna de las características irrelevantes, y el subconjunto de características débilmente relevantes que se requiera para el adecuado desempeño del clasificador.

2.3. Entropía

En [9] se introduce este concepto como una medida de desorden presente en los datos de análisis. Sea f una función [7]:

$$f(k) = -k\log(k) \tag{1}$$

donde $0 \le k \le 1$. *f(k)* es una función no negativa cuyo valor es igual a cero para $k = 0$ y $k = 1$, y positivo para valores dentro del intervalo (0, 1). Para el conjunto inicial de entrenamiento $\mathbf{X}=[\mathbf{\Phi}_1,\mathbf{\Phi}_2,\ldots,\mathbf{\Phi}_p]$, la entropía de un vector columna $\mathbf{\Phi}$ puede ser descrita como [11]:

$$H(\mathbf{\Phi}) = \sum_i f\,(p(\mathbf{\Phi}=\phi_i)) \tag{2}$$

2.4. Negentropía

La negentropía o entropía diferencial es una medida de información que puede interpretarse como la medida de aleatoriedad de una variable, resaltando que puede tener valores negativos. La entropía convencional no puede ser negativa porque la función *f* es no negativa en el rango [0,1], y las probabilidades necesariamente permanecen en ese rango.
La negentropía J se define como [10]:

$$J(\mathbf{\Phi}) = H(\mathbf{\Phi}_{gauss}) - H(\mathbf{\Phi}) \tag{3}$$

donde $H(\mathbf{\Phi}_{gauss})$ es un vector aleatorio gaussiano con la misma matriz de covarianza $\mathbf{\Sigma}$ que $\mathbf{\Phi}$. Su entropía puede ser evaluada de la siguiente manera:

$$H(\mathbf{\Phi}_{gauss}) = \frac{1}{2}\log\left|\det\mathbf{\Sigma}\right| + \frac{n}{2}\left[1+\log 2\pi\right] \tag{4}$$

donde *n* es la dimensión de $\mathbf{\Phi}$.

2.5. Análisis de Componentes Independientes

El Análisis de Componentes Independientes (ICA) es una herramienta que busca separar fuentes estadísticamente independientes a partir de mezclas obtenidas mediante diferentes sensores [13]. Existen diferentes versiones del algoritmo, pero el método usado en este trabajo, para el conjunto inicial de entrenamiento $\mathbf{X}_{n\times p}=[\mathbf{\Phi}_1,\mathbf{\Phi}_2,\ldots,\mathbf{\Phi}_i]$, funciona bajo las siguientes suposiciones [14]: las fuentes son estadísticamente independientes, solo una de las fuentes puede tener una distribución gaussiana y las señales capturadas en los sensores $\{\mathbf{\Phi}_j = \phi_j(t),\, j = 1,\ldots,p \wedge t = 1,\ldots,n\}$ son combinaciones lineales de las fuentes $s_j(t)$. Para dos fuentes y dos sensores, el problema puede ser formulado, así:

$$\begin{bmatrix}\phi_1(t)\\ \phi_2(t)\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}s_1(t)\\ s_2(t)\end{bmatrix} \tag{5}$$

con el objetivo de encontrar $a_{11}, a_{12}, a_{21}, a_{22}, s_1(t), s_2(t)$ para $\phi_1(t)$ y $\phi_2(t)$ conocidas. De forma general, para un número de sensores igual al de fuentes, se tiene:

$$\mathbf{x} = \begin{bmatrix}\phi_1(t)\\ \vdots\\ \phi_p(t)\end{bmatrix}, \mathbf{A} = \begin{bmatrix}a_{11} & \cdots & a_{1p}\\ \vdots & \ddots & \vdots\\ a_{p1} & \cdots & a_{pp}\end{bmatrix}, \mathbf{s} = \begin{bmatrix}s_1(t)\\ \vdots\\ s_p(t)\end{bmatrix} \tag{6}$$

de tal manera, que las combinaciones lineales de las fuentes se pueden reescribir como **x** = **As**. Cuando **A** es encontrada, **s** puede ser calculada como $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$. La solución encontrada cuenta con la limitación de que tanto los signos como el orden de las fuentes no pueden ser determinados [15], [16].

Sea, $y_j(t)$, la estimación de un componente independiente ($\approx s_j(t)$) y $\mathbf{w}_j$ una fila de la matriz estimada de transformación inversa $\mathbf{W} \approx \mathbf{A}^{-1}$. Un componente independiente se estima de acuerdo a la expresión $y_j(t) = \mathbf{w}_j^{\mathrm{T}}\mathbf{x}$.

Bajo ciertas condiciones, la distribución que resulta de la suma de variables aleatorias estadísticamente independientes, es mucho más cercana a la gaussiana que cualquiera de las distribuciones de las variables originales [10]. Este teorema resulta básico para los algoritmos de ICA. Como $\phi_j(t)$ es una suma ponderada de los componentes en **s** y los componentes son no gaussianos e independientes, el problema se reduce a encontrar $\mathbf{w}_j$ de tal manera que la no gaussividad de $y_j(t)$ sea maximizada [17]. Una manera de medir la no gaussividad, es por medio de la siguiente aproximación de negentropía [18]:

$$J(y) \propto \left(\mathrm{E}\{G(y)\} - \mathrm{E}\{G(v)\}\right)^2 \tag{7}$$

donde $\propto$ denota proporcionalidad; *v* es una variable gaussiana con media cero y varianza unitaria, de manera que el término $\mathrm{E}\{G(v)\}$ es una constante; $G(y)$ es una función no cuadrática, cuya elección depende del problema, y las comúnmente usadas son: $G_1(y) = \dfrac{y^4}{4}$, $G_2(y) = \log(\cosh(y))$ y $G_3(y) = -\exp\left(-\dfrac{y^2}{2}\right)$. El principal interés es hallar $\mathbf{w}_j$ que maximice $J(\mathbf{w}_j^{\mathrm{T}}\mathbf{x})$.

3. MARCO EXPERIMENTAL

3.1. Base de Datos

Corresponde a una muestra de registros de voz que evalúa 90 niños. Médicos especialistas asignaron dos etiquetas de clase: registro normal y registro de paciente hipernasal (45 registros por clase). Cada grabación está conformada por cinco palabras seleccionadas: /*coco*/, /*gato*/, /*jugo*/, /*mano*/ y /*papá*/.

Las señales fueron tomadas en condiciones de bajo nivel de ruido ambiental usando un micrófono dinámico unidireccional (cardiode). El rango de adquisición de todas las señales de voz está entre (−1, 1). Es importante anotar que la palabra /*papá*/ de esta base de datos, se considera ruidosa debido a que los ni˜nos no fijaron el acento en la misma sílaba. Los momentos de análisis por cada característica corresponden a los parámetros de posición, escala y forma [20]. En total se analizan 128 variables de representación por cada palabra. Por lo tanto, el espacio inicial de entrenamiento que extrae la información de los registros de voz, a la entrada de los clasificadores, se incrementa hasta 640. Sobre este espacio inicial de características se realiza un preproceso de datos (remoción de valores anómalos, verificación de normalidad y transformación de distribuciones), que tiene como objetivo la disminución de la influencia, entre otras razones, de errores sistemáticos de adquisición, fallas ocasionales de los dispositivos, etc. Se considera también, el control de la homogeneidad de las propiedades estadísticas de las características. Por tanto, cada palabra queda con el número de características que se muestra en la Tabla I.

3.2. Algoritmos

Análisis de Componentes Independientes basado en Negentropía. La base de este algoritmo es el cálculo de los componentes independientes usando el algoritmo de Mastica para estimar varios componentes independientes, con ortogonalización simétrica [10]. Una fase de preprocesamiento es útil antes de intentar estimar $\mathbf{W}$, la matriz de separación de los componentes. Primero, las señales mezcladas deben ser centradas substrayendo su valor medio $\mathrm{E}(\mathbf{x})$:

$$\hat{\mathbf{x}} = \mathbf{x} - \mathbf{E}\{\mathbf{x}\} \tag{8}$$

Así, las señales comienzan su proceso de blanqueamiento, lo cual significa que son transformadas linealmente de manera que los componentes quedan no correlacionados y con varianza unitaria. El blanqueamiento requiere de la descomposición en eigenvalores de la matriz de covarianza,

$\mathbf{VDV}^{\mathrm{T}} = \mathrm{E}\{\hat{\mathbf{x}}\hat{\mathbf{x}}^{\mathrm{T}}\}$. $\mathbf{V}$ es la matriz de eigenvectores ortogonales y $\mathbf{D}$ es una matriz diagonal con los correspondientes eigenvalores.

El blanqueamiento es realizado mediante la multiplicación con la matriz de transformación, quedando $\mathbf{P} = \mathbf{VD}^{-\frac{1}{2}}\mathbf{V}^{\mathrm{T}}$, quedando:

$$\tilde{\mathbf{x}} = \mathbf{Px} \tag{9}$$

La covarianza de los datos blanqueados $\mathrm{E}\{\hat{\mathbf{x}}\hat{\mathbf{x}}^{\mathrm{T}}\}$ es la matriz identidad, y la matriz de mezcla $\tilde{\mathbf{A}} = \mathbf{PA}$ es ortogonal $\left(\tilde{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}^{-1}\right)$. La matriz para extraer los componentes independientes de $\tilde{\mathbf{x}}$, es ahora notada como $\tilde{\mathbf{W}}$, de manera que $\mathbf{W} = \tilde{\mathbf{W}}\mathbf{P}$. Solo para resumir la notación: $\mathbf{s} \approx \mathbf{y} = \mathbf{Wx} = \tilde{\mathbf{W}}\mathbf{Px}$.

Como la varianza de los componentes independientes no puede ser determinada (limitación 1), es natural restringir la varianza de $\tilde{y}_j = \mathbf{w}_j^{\mathrm{T}}\tilde{\mathbf{x}}$ a la unidad. Cuando los datos son blanqueados, esto es equivalente a restringir a un valor unitario la norma de $\mathbf{w}_j$. La expresión que describe la aproximación de la negentropía $J(y)$, ocasiona un problema de optimización: encontrar el $\mathbf{w}_j$ que maximice $J(\mathbf{w}_j^{\mathbf{T}}\tilde{\mathbf{x}})$ bajo la restricción de que $\|\mathbf{w}_j\| = 1$.

De la teoría de optimización se tiene que el extremo de $\mathrm{E}\{G(y)\}$ es encontrado donde la función del gradiente de Lagrange es cero. Como la restricción $\|\mathbf{w}_j\| = 1$ es equivalente a $\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1 = 0$, la función de Lagrange está dada por:

$$\mathrm{L}(\mathbf{w}, \boldsymbol{\lambda}) = \mathrm{E}\{G(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - \boldsymbol{\lambda}(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1) \tag{10}$$

El gradiente con respecto a $\mathbf{w}$ es:

$$\mathrm{L}'(\mathbf{w}, \boldsymbol{\lambda}) = \mathrm{E}\{\mathbf{x}g(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - 2\boldsymbol{\lambda}\mathbf{w} \tag{11}$$

donde $g(y)$ es el gradiente de $G(y)$, de manera que se obtiene la solución:

$$\mathrm{E}\{\mathbf{x}g(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - 2\boldsymbol{\lambda}\mathbf{w} = 0 \tag{12}$$

Los gradientes de las funciones G, más usadas en la literatura, están dados por $g_1(y) = y^3$, $g_2(y) = \tanh(y)$ y $g_3(y) = y\exp\left(-\frac{y^2}{2}\right)$.

En el algoritmo de FastICA, la ecuación $L_w(\mathbf{w}, \boldsymbol{\lambda}) = 0$, es resuelta iterativa mente utilizando el método de Newton. Una solución para cada componente es buscada, de manera que la optimización tiene que ser realizada para un componente a la vez. Los $\mathbf{w}_j$ son ortogonales, de manera que para impedir que la misma solución sea encontrada más de una vez, una decorrelación Gram-Schmidt puede ser realizada en cada iteración. Suponer un punto de partida $\mathbf{w}_j$ es necesario para inicializar el método de Newton, y el orden en que los componentes son encontrados dependen de este supuesto punto de partida. El algoritmo debe continuar hasta que un adecuado criterio de parada sea logrado. Normalmente se utilizan como criterio que $\mathbf{w}_j$ converja a la solución o que demasiadas iteraciones hayan sido realizadas.

Selección de características basada en la Entropía. La base del algoritmo es eliminar todas las características que tienen la carga informativa más pequeña con respecto a las demás, tomando como principal argumento la cantidad de información. La métrica usada es la entropía. El primer paso consiste en estimar la función de densidad de probabilidad de cada vector característico. Luego, se construye una matriz con los valores de probabilidad, donde cada valor del vector de características tiene un valor de probabilidad. En el siguiente paso del algoritmo, se calcula la entropía de cada una de las Características utilizando los valores de su función densidad de probabilidad estimados en el paso anterior. Ahora, usando como referencia el rango de valores situados entre el máximo y el mínimo, se sitúa un umbral que permita eliminar características que de acuerdo a su valor de entropía, superen o no el umbral que ha sido fijado experimentalmente.

3.3. Procedimiento General

En primer lugar se hizo normalización geométrica a cada vector de características. Este escalamiento permite evitar problemas que pueden presentarse por las diferentes escalas de magnitud que tienen las características. Después de realizar la transformación usando ICA, y/o la selección de características usando el segundo algoritmo descrito en la sección **3.2.**, los datos se ubican a la entrada de un clasificador Bayesiano de Naive. Este clasificador da el número de observaciones que fueron clasificadas correctamente y el porcentaje que este número representa sobre el número total de observaciones de la matriz inicial de entrenamiento $\mathbf{X}$.

4. RESULTADOS

El conjunto completo de características se ubicó a la entrada del clasificador obteniéndose los resultados de clasificación que se muestran en la Tabla 1. **Cin** es el número de características a la entrada del clasificador y **PC** es el porcentaje de precisión de clasificación usando validación cruzada. El resultado para la palabra /*papá*/ es esperado, debido a que esta palabra se considera ruidosa por los motivos expuestos en la sección **3.1.**

Palabra	Cin	PC(%)
/coco/	65	75.4
/gato/	67	68.2
/jugo/	59	69.7
/mano/	68	73.1
/papá/	43	54.6

Tabla1. Resultados de clasificación para el espacio inicial de entrenamiento completo.

Después de aplicar ICA al conjunto inicial de entrenamiento, usando tres funciones $G(y)$ diferentes, y ubicar este nuevo espacio de representación a la entrada del clasificador, se observa que los mejores resultados se obtienen usando la función $G(y) = -\exp\left(-\frac{y^2}{2}\right)$. En la Tabla 2 están subrayados los mejores resultados de clasificación. Es de destacar en este resultado, que la precisión de clasificación mejoró sustancialmente.
Para mejor comprensión de las tablas de resultados se utilizan las siguientes convenciones:

$$G(y) = \frac{y^4}{4} \Rightarrow (1)$$

$$G(y) = \log(\cosh(y)) \Rightarrow (2)$$

$$G(y) = -\exp\left(-\frac{y^2}{2}\right) \Rightarrow (3)$$

Palabra	Cin	Función	PC(%)
/coco/	65	(1)	92.2
		(2)	71.1
		(3)	<u>96.7</u>
/gato/	67	(1)	92.2
		(2)	81.1
		(3)	<u>93.3</u>
/jugo/	59	(1)	86.7
		(2)	87.8
		(3)	<u>95.6</u>
/mano/	68	(1)	91.1
		(2)	88.9
		(3)	<u>93.3</u>
/papá/	43	(1)	75.6
		(2)	72.2
		(3)	<u>83.3</u>

Tabla2. Resultados de Clasificación en el espacio de representación obtenido mediante ICA.

En la Tabla 3, se muestran los resultados obtenidos de clasificación para cada una de las palabras, después de la reducción de dimensiones basada en la eliminación de las características que están por debajo del umbral fijado en el 30% de la entropía total. Se observa que eliminando las variables redundantes mejora la tasa de acierto del clasificador, sin embargo, no es un incremento destacable.

Palabra	Cin	PC(%)
/coco/	62	86.7
/gato/	63	74.5
/jugo/	54	79.9
/mano/	61	84.4
/papá/	38	64.4

Tabla3. Resultados de Clasificación para el Espacio de Dimensión Reducida Basada en la Entropía.

En la Tabla 4, se exponen los resultados de clasificación que fueron obtenidos mediante el procedimiento híbrido construido así: primero actuó el algoritmo básico de selección de características basado en la entropía (que redujo el número de características), y luego en cascada, actuó el algoritmo FastICA basado en negentropía (el cual generó un espacio transformado de representación). Los mejores resultados se presentan subrayados.

Palabra	Cin	Función	PC(%)
/coco/	65	(1)	91.2
		(2)	95.6
		(3)	<u>96.7</u>
/gato/	67	(1)	93.4
		(2)	84.5
		(3)	<u>97.8</u>
/jugo/	59	(1)	86.7
		(2)	87.8
		(3)	<u>95.9</u>
/mano/	68	(1)	85.6
		(2)	80.1
		(3)	<u>91.2</u>
/papá/	43	(1)	68.9
		(2)	68.9
		(3)	<u>92.3</u>

Tabla4. Resultados de Clasificación para el Algoritmo Híbrido Construido.

5. CONCLUSIONES Y DISCUSIÓN

Se obtiene la reducción de dimensiones usando medidas de información, además de mejorar sustancialmente la precisión de clasificación, mediante el desarrollo de un procedimiento híbrido entre un algoritmo básico de selección de características basado en la entropía y el algoritmo FastICA basado en negentropía. A pesar que no se logró reducir en mucha proporción la dimensión del espacio inicial de características, la tasa de acierto se incrementó considerablemente, además de ser una

herramienta que ofreció un buen resultado para la palabra /*papá*/, que se cataloga como ruidosa.
El tiempo de cómputo usado en la ejecución del algoritmo híbrido no es exhaustivo, lo cual hace que el procedimiento sea fácilmente aplicable a otro tipo de datos, sin exigentes requerimientos de cómputo. Un problema presente en espacios de representación con dimensión alta, es que la cantidad de variables redundantes, aunque puede ser baja, pueden afectar apreciablemente la precisión de clasificación, así como sucede en el caso particular de este trabajo. La metodología que se propone en este artículo está orientada a preservar la información y minimizar la redundancia.

6. AGRADECIMIENTOS

Los autores manifiestan su agradecimiento a la Ing. Olga Beatriz Naranjo por sus aportes a este trabajo. Este trabajo se enmarca dentro del proyecto titulado *"Identificación automatizada de hipernasalidad en niños con LPH por medio de análisis acústico del habla"*, financiado por la DIMA con código 20201004208.

7. BIBLIOGRAFÍA

[1] M. Last, A. Kandel, and O. Maimon, "Information - theoretic algorithm for feature selection," *ELSEVIER - Pattern Recognition Letters*, no. 22, pp. 799–811, December 2001.

[2] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognition*, vol. 35, p. 835–846, 2002.

[3] I. Cohen, Q. Tian, X. Zhou, and T. Huang, "Feature selection using principal feature analysis," in *ICIP'02*, 2002.

[4] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[5] H.K. Ekenel , B. Sankur, "Feature selection in the independent component subspace for face recognition" Department of Electrical and Electronic Engineering, Bogazici University, Bebek 34342, Istanbul, Turkey, *Technical report,* 2004..

[6] X. Zeng, Y. Chen, "Selection of ICA features for texture classification" *International symposium on neural networks No2, Chongqing,* CHINE (30/05/2005).

[7] X. Zeng, Y. Chen, "Selection of ICA features for texture classification" *International symposium on neural networks No2, Chongqing,* CHINE (30/05/2005).

[8] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, p. 1205–1224, 2004.

[9] C. E. Shannon, "A mathematical theory of communication," *Bell Syst.Tech.J.*, vol. 27, pp. 379–423;623–656, July 1948.

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, I. John Wiley & Sons, Ed. Wiley - Interscience, 2001.

[11] R. M. Gray, *Entropy and Information Theory*, S. Verlag, Ed. Springer-Verlag. New York, 1990.

[12] O. Tibaduisa, M. Vallverdú, B. Giraldo, P. Caminal, D. Hoyer, and S. Benito, "Estudio de la información mutua en la variabilidad del patrón respiratorio," in *XXV Jornadas de Automática, Ciudad Real*, Septiembre 2004.

[13] A. Cichocki, "Blind signal processing methods for analyzing multichannel brain signals," *International Journal of Bioelectromagnetism*, vol. 6, no. 1, 2004.

[14] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, "Independent component approach to the analysis of eeg and meg recordings," in *IEEE Transactions on Biomedical Engineering,*, vol. 47, no. 5, May 2000.

[15] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," in *Neural Cumputation*, no. 9. Helsinki University of Technology: Laboratory of Computer and Information Science, 1997, pp. 1483–1492.

[16] S. Romero, M. A. Mañanas, J. Riba, A. Morte, S. Giménez, S. Clos, and M. J. Barbanoj, "Nuevo método automático de filtrado ocular en señales eeg espontáneas basado en análisis de componentes independientes," in *XXV Jornadas de Automática, Ciudad Real*, September 2004.

[17] Z. Wei, L. Jin, and Y. Jin, "Independent component analysis." Department of Statistics: Stanford University, May 2005.

[18] P. Hubka, V. Rosik, J. Zdinak, M. Tysler, and I. Hulín, "Independent component analysis of electrogastrographic signals," in *Measurement Science Review*, vol. 5, 2005.

[19] P. A. Regalia and E. Kofidis, "Monotonic convergence of fixed-point algorithms for ica," in *IEEE Transactions on Neural Networks*, vol. 14, no. 4, July 2003.

[20] G. Castellanos, O. D. Castrillón, and E. Guijarro, "Multivariate analysis techniques for effective feature selection in voice pathologies," in CASEIB, 2004.