SELECCIÓN EFECTIVA DE CARACTERÍSTICAS PARA BIOSEÑALES UTLIZANDO EL ANALISIS DE COMPONENTES PRINCIPALES

RESUMEN

Este articulo presenta algunos resultado parciales de una reciente investigación que comparó varias técnicas lineales y no lineales del análisis multivariado de datos con el objeto de seleccionar y extraer de manera efectiva un grupo de características basadas en señales electrocardiográficas orientadas a la identificación del infarto agudo de miocardio.

Específicamente en este artículo se presentan los resultados obtenidos al aplicar el método lineal de análisis en componentes principales para generar un subespacio de características de menor dimensión que el original. Se presentan también los resultados obtenidos al evaluar la precisión de la clasificación de estados funcionales normales y patológicos del miocardio utilizando un clasificador bayesiano. Además se estimó también su costo computacional.

PALABRAS CLAVES: PCA, Extracción de características, Selección de características, Clasificador Bayesiano.

JORGE HERNANDO RIVERA
Ingeniero Electrónico, Ms.C
Profesor Asistente
Ingeniería Física
Universidad Tecnológica de Pereira
j.rivera@utp.edu.co

CESAR CASTELLANOS
Ingeniero Electrónico, Ph.D
Profesor Titular
Ingeniería Electronica
Universidad Nacional de Colombia
gcastell@telesat.com

JOSE SOTO MEJIA
Fisico, Ph.D.
Profesor Titular
Ingeniería Industrial
Universidad Tecnológica de Pereira
jomejia@utp.edu.co

Grupo de investigación LIDER
Tecnología Eléctrica

ABSTRACT

In this article some partial results of comparison results from a recent investigation are presented, in this investigation a comparation between linear and non linear methods from multivariate analysis is made with the main purpose of selection and feature extraction from electrocardiographic signals, this all oriented to identification of accute infarction of the myocardium.

Specifically this article summarizes the results from having applied the multivariate method of analysis known as analysis of principal components to generate a subspace of characteristics of minorless dimension that the original one.

The precision of the classification of normal and pathological functional states of the myocardium using a Bayesian classifier was also computed. Its associated computational cost was also estimated.

KEYWORDS: PCA, feature extraction, feature selection, bayesian classifier.

1. INTRODUCCIÓN

Este articulo presenta algunos resultado parciales de la reciente investigación [1] que comparó varias técnicas lineales y no lineales del análisis multivariado de datos (Análisis de Componentes Principales (PCA), PCA Probabilístico, Análisis de Componentes Independientes (ICA), Kernel PCA y Kernel ICA) con el objeto de seleccionar y extraer de manera efectiva un grupo de características basadas en señales electrocardiográficas orientadas a la clasificación de estados funcionales normales y patológicos del infarto agudo de miocardio. El desempeño de cada técnica, en términos de la precisión de la clasificación, así como su costo computacional, fue probado experimentalmente en pruebas con bases de datos convencionales y con bases de datos correspondientes a características extraídas de señales biomédicas reales [1].

En particular este artículo presenta los resultados obtenidos al aplicar el método lineal de análisis en componentes principales para generar un subespacio de características de menor dimensión que el original. Para el subespacio generado por éste método se presentan también los resultados obtenidos al evaluar la precisión de la clasificación de estados funcionales normales y patológicos utilizando un clasificador bayesiano. Además se estimó también su costo computacional.

En la sección 2 “Justificación”, se resalta la importancia de seleccionar y extraer de manera efectiva un grupo de características basadas en señales electrocardiográficas orientadas a la clasificación de estados funcionales normales y patológicos del infarto agudo de miocardio.

La sección 3 “Análisis de Componentes Principales” presenta los fundamentos teóricos del método matemático utilizado en la selección efectiva de las características ECG para la identificación de infarto agudo del miocardio.

La sección 4 “Esquema de trabajo” describe el marco experimental utilizado.

En la sección 5 “Resultados” se presentan los resultados obtenidos al aplicar el método lineal de análisis en componentes principales para generar un subespacio de características de menor dimensión que el original y la evaluación de la precisión de clasificación obtenida utilizando un clasificador bayesiano.

Fecha de Recepción: 30 Agosto de 2006
Fecha de Aceptación: 12 Abril de 2007

Finalmente, la sección 6 "Conclusiones y Trabajos Futuros", presenta las conclusiones parciales obtenidas al utilizar el análisis en componentes principales y hace una relación de las otras técnicas que fueron evaluadas y cuyos resultados se publicaran en detalle en otros artículos de esta revista.

2. JUSTIFICACIÓN

El diagnóstico del Infarto Agudo del Miocardio se basa en la caracterización de las señales biomédicas del electrocardiógrafo.
Los sistemas automatizados de identificación evalúan grupos conformados por un gran número de diferentes características-variables- extraídas del fenómeno de estudio. En la mayoría de estos procesos se contemplan amplios conjuntos de características que conllevan al empleo de grandes recursos computacionales, tanto en la etapa de caracterización como en las posteriores de almacenamiento y procesamiento de los datos.
Uno de los principales problemas con los conjuntos de datos de alta dimensión es que no todas las variables medidas son relevantes en términos de representación [2]. Además, el aumento del número de características no esta relacionado en proporción directa con la buena capacidad de clasificación para una posterior detección [3].

Por lo anterior, es aconsejable reducir la dimensión de los datos-numero de variables-, manteniendo lo más que se pueda de la estructura original de los mismos. Así, un número limitado de características simplifica la representación tanto del patrón de caracterización como la de los parámetros de clasificación, resultando una extracción y análisis menos denso, permitiendo tener un clasificador más rápido y con menos carga computacional.

El análisis multivariado de datos ofrece varios métodos estadísticos para reducir la dimensión del espacio de características [4] [5], de tal manera, que se descarten las variables que no ofrecen separabilidad, con el objetivo de hacer mas preciso y eficiente el sistema de detección, sumando a ello, la disminución en el tiempo de computo requerido para el entrenamiento del sistema clasificador.
Sin embargo, una reducción exagerada en el número de características podría llevar a una perdida en el poder discriminante empobreciendo la precisión del sistema de reconocimiento. Con base en lo anterior, se puede intuir que la elección de un adecuado proceso para realizar la reducción de la dimensión, es una decisión que ubica en una balanza la disminución del número de variables de análisis contra la precisión de los resultados de clasificación.

El proceso de reducción de dimensión puede entonces entenderse como la transformación del espacio de variables originales a un espacio de dimensión menor. Dichas transformaciones pueden ser combinaciones lineales o no lineales de las características originales.

El entrenamiento de sistemas automáticos, usados en la detección de patologías a partir de señales biomédicas, implica obtener la efectiva selección de variables o combinaciones de las mismas que contengan la información suficiente contenida en la señal para su adecuada representación.

Es así como en la reciente investigación ya citada [1], se compararon varias técnicas lineales y no lineales del análisis multivariado de datos (Análisis de Componentes Principales (PCA), PCA Probabilístico, Análisis de Componentes Independientes (ICA), Kernel PCA y Kernel ICA) con el objeto de seleccionar y extraer de manera efectiva un grupo de características basadas en señales electrocardiográficas orientadas a la clasificación de estados funcionales normales y patológicos del infarto agudo de miocardio. Además, el desempeño de cada una de las técnicas, en términos de la precisión de la clasificación, así como su costo computacional, fue probado experimentalmente en pruebas con bases de datos convencionales y con bases de datos correspondientes a características extraídas de señales biomédicas reales.

3. ANALISIS DE COMPONENTES PRINCIPALES

Los métodos para reducir la dimensionalidad tienen como objetivo extraer las características que son relevantes para la clasificación planteada.
En la investigación ya citada [1] de acuerdo a la operación de mapeo realizado al espacio inicial de características se evaluaron, para la selección de características, las siguientes técnicas lineales y no lineales:

Métodos lineales

- Análisis de componentes principales (PCA)
- PCA Probabilístico
- Análisis de componentes independientes (ICA)

Métodos no lineales

- Kernel PCA
- Kernel ICA

En particular, en lo subsiguiente de esta sección se resumen los fundamentos teóricos acerca de la técnica del Análisis en Componentes Principales cuyos resultados se presentan en la sección 5 "Resultados" de éste artículo.

3.1 El método de Análisis en Componentes Principales

El Análisis de Componentes Principales (PCA - Principal Component Analysis) tiene por objetivo principal reducir

la dimensión de un conjunto de variables, tratando de mantener la mayor cantidad de información que sea posible. Esto se logra mediante la transformación a un nuevo conjunto de variables las cuales son no correlacionadas y se ordenan de modo tal que unas pocas (las primeras) retengan la mayor cantidad de variación presente en el conjunto original de variables [6].

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. La técnica PCA presenta una doble utilidad: permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general de dimensión p (posible identificación de variables latentes), además, permite transformar las variables originales que generalmente están correlacionadas, en nuevas variables no correlacionadas que facilitan la interpretación.
[7].

Sea **X** la matriz original de datos de dimensión n × p. Las filas corresponden a las observaciones y las columnas a las variables, donde la media de cada una de las variables es cero.

$\mathbf{X} = \{x_{ij}\}$; donde $i = 1, \ldots, n$ representa la observaciones y $j = 1, \ldots, p$, representa la variable.

El propósito es hallar un subespacio de dimensión *m*, $m < p$, tal que al proyectar los puntos sobre dicho subespacio, los puntos conserven su estructura con la menor distorsión posible.
En una primera aproximación, se desearía proyectar todos los puntos observados sobre un subespacio de dimensión uno (una recta), de tal forma que todos los puntos mantengan, lo más posible, sus posiciones relativas.
Si se considera el punto x_i y una dirección $a_1 = (a_{11}, a_{12}, \ldots, a_{1p})^T$, definida por el vector a_1 de norma la unidad, la proyección del punto x_i sobre esta dirección es el escalar: $z_i = a_1^T x_i$

La primera componente se obtiene de manera que su varianza sea máxima, sujeta a la restricción de que la suma de los pesos de a_1 al cuadrado sea igual a la unidad.

Debido a que las proyecciones Zi son variables con media cero, maximizar sus cuadrados es equivalente a maximizar su varianza, lo que significa encontrar la dirección de proyección que maximice la varianza de los datos proyectados, así:

$$\max_{a_1} \sum_{i=1}^{n} z_i^2 = \max_{a} \sum_{i=1}^{n} a_1^T x_i^T x_i a_1 \quad (1)$$

Considerando el vector de proyecciones,

$$Z_1 = (z_1, z_2, \ldots, z_n)^T = X_1 a_1 \quad (2)$$

se puede reescribir (1) como,

$$\max_{a_1} (z_1^T z_1 = a_1^T X^T X \; a_1) \quad (3)$$

Por otra parte, la media de z_1 es nula y su varianza es igual a :

$$\frac{1}{n} z_1^T z_1 = \frac{1}{n} a_1^T X^T X \; a_1 = a_1^T S \, a_1 \quad (4)$$

donde ***S*** es la matriz de covarianzas de las observaciones. Para maximizar (4) se utilizan multiplicadores de Lagrange, tal que $a_1^T a_1 = 1$.
En consecuencia incorporando la restricción se forma el siguientes lagrangiano:

$$L = a_1^T S \; a_1 - \lambda (a_1^T a_1 - 1) \quad (5)$$

Para maximizar el valor del lagrangiano lo derivamos respecto a a_1 e igualando a cero se tiene:

$$\frac{\partial L}{\partial a_1} = 2S \; a_1 - 2\lambda a_1 = 0 \quad (6)$$

Finalmente,

$$S \; a_1 = \lambda a_1 \Rightarrow (S - \lambda I) a_1 = 0$$

Esto significa que a_1 es un vector propio de la matriz ***S*** asociado al valor propio λ, que corresponde a la varianza de z_1. Por tanto, el vector propio asociado al mayor valor propio de ***S*** corresponde al primer componente principal.

En general, es posible hallar el espacio de dimensión *m* que mejor represente los datos, el cual esta dado por los vectores propios asociados a los *m* mayores valores propios de la matriz ***S***. Estas nuevas direcciones se denominan direcciones principales de los datos y las proyecciones de los datos originales sobre estas direcciones se conocen como componentes principales. Usualmente, la matriz **X** tiene rango *p* (también la matriz ***S***), existiendo entonces tantos componentes principales como variables, que se obtienen calculando los vectores propios $\lambda_1, \lambda_2, \ldots, \lambda_p$ de la matriz de covarianza ***S***

4. ESQUEMA DE TRABAJO

En esta sección se presentarán el esquema de trabajo utilizado para evaluar el método lineal de *análisis en componentes principales* para generar un subespacio de características de menor dimensión que el original. Para

el subespacio generado por éste método se evaluó la precisión de la clasificación utilizando un clasificador bayesiano. Además la estimación de su costo computacional a efecto de compararlo, posteriormente, con los otros métodos de reducción de dimensionalidad mencionados en la sección anterior.

4.1 Base de datos de características ECG

La base de datos ECG con registros electrocardiográficos de corazones normales y otros registros de corazones con patología de Infarto Agudo del Miocardio, hace parte del trabajo de investigación que viene realizando el grupo de control y procesamiento digital de señales "GCPDS" de la Universidad Nacional de Colombia sede Manizales.

Las características que presenta esta base de datos corresponde al análisis de la base de datos ST-T europea. Esta base comienza a desarrollarse en 1985 a través del proyecto "Concerted Action on Ambulatory Monitoring" de la Comunidad Europea, cuyo objetivo era definir una base de datos de ECG en pacientes ambulatorios y en la cuál participaron expertos de doce países [9].

Una descripción técnica detallada de cómo fue formada esta base de datos se encuentra en [10].
Inicialmente, la señal electrocardiográfica fue pre-procesada mediante técnicas como: Mediciones heurísticas, transformada wavelet (WT) con objeto de identificar las características que mejor discriminan el infarto agudo del miocardio, resultando en un total de 1009 características.
La base de datos consta de 1800 observaciones, 900 correspondientes a características extraídas de electrocardiogramas para corazones normales, y 900 correspondientes a características extraídas de electrocardiogramas correspondientes a patología de isquemia o infarto agudo del miocardio, para un total de las 1009 características arriba mencionadas [11].

4.2 El Subespacio de características generado y precisión de la clasificación

El subespacio de características que se evaluó fue generado utilizando inicialmente el primer componente principal, el siguiente subespacio fue generado con los dos primeros componentes, y así sucesivamente hasta obtener *p* subespacios de características. Para cada uno de estos subespacios se evaluó la precisión en la clasificación utilizando un clasificador bayesiano y se estimó también su costo computacional, determinado por el tiempo utilizado en la clasificación.

La separabilidad lineal de las bases de datos fue probada utilizando el algoritmo de kozinec [8], el cual se basa en la búsqueda de un hiperplano que separe en un subespacio de búsqueda dos clases de patrones. El encontrar dicho hiperplano implica la separabilidad lineal del subespacio.

Para el entrenamiento y la validación del clasificador bayesiano, se utilizó un 50% del total de las 1800 observaciones con que contaba la base de datos de características de la señal electrocardiográfica y el 50% restante fue utilizado en la validación del clasificador.

Para determinar la tendencia hacia la distribución gaussiana de las características de la base de datos se utilizó la prueba de hipótesis de Kolmogorov-Smirnov con un intervalo de confianza del 95%, además se realizó una prueba de kurtosis para determinar la gaussividad de las características.

Adicionalmente se realizó un análisis de correlación para el espacio inicial de características, para determinar gráficamente el nivel de correlación entre estas.
La base de datos posee dos clases: (1) ausencia o (2) presencia de infarto agudo del miocardio.

5.0 RESULTADOS

Con base en la prueba de hipótesis Kolmogorov-Smirnov se puede rechazar la hipótesis de gaussividad para cada una de las características que componen esta base de datos. El análisis de kurtosis muestra que en su mayoría las características que componen esta base de datos son super-gaussianas.

Con base en la grafica (1) de la matriz de correlación se muestra que las variables son altamente correlacionadas, además esta base de datos es linealmente separable, según los resultados obtenidos por el algoritmo de Kozinec.

.

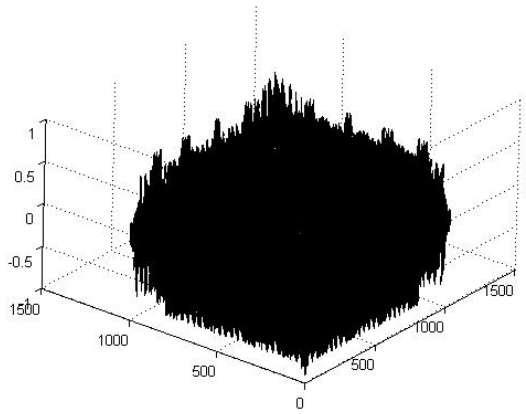

Figura 1. Gráfica de la matriz de correlación

En la tabla (1) se muestra los resultados de precisión en el clasificador bayesiano asíı como el costo computacional obtenido para el espacio inicial de características.

Precisión en clasificación	Costo computacional
50%	1569ms

Tabla 1. Precisión y costo del conjunto inicial de características

Se puede notar que para el conjunto inicial de características (sin aplicar el método de componentes principales) el clasificador pierde su poder discriminante. En la figura (2) se muestra los resultados para la precisión en la clasificación, así como el costo computacional para todos los subespacios generados por el método de componentes principales.

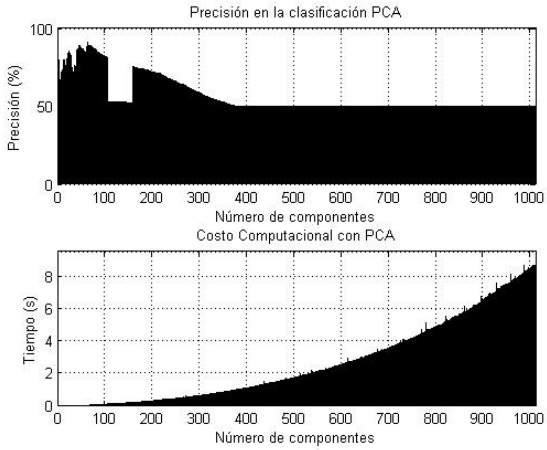

Figura 2. Precisión y costo computacional para PCA

En la figura 2 anterior se puede observar que cuando el numero de componentes del subespacio aumenta, el clasificador pierde su poder discriminante además el costo computacional se incrementa exponencialmente.
Para los subespacios generados la mejor clasificación se obtuvo con 64 componentes principales para una precisión en la clasificación del 90.33% con un costo computacional de 62 ms como se muestra en la tabla 2.

Componentes	Precisión	Costo computacional
64	90.33%	62 ms

Tabla 2. Mejor resultado en la clasificación

7. CONCLUSIONES Y TRABAJOS FUTUROS

La transformación del espacio de la base de datos con características ECG con la técnica de Análisis en Componentes Principales permitió establecer que la mejor clasificación se obtuvo con 64 componentes principales para una precisión en la clasificación del 90.33% con un costo computacional de 62 ms.

El esquema de trabajo presentado en la sección 4 será también el utilizado para la evaluación de las técnicas PCA Probabilístico, Análisis de componentes independientes (ICA) y los *Métodos no lineales:* Kernel PCA y Kernel ICA y cuyos resultados se publicaran en detalle en otros artículos de esta revista.

8. BIBLIOGRAFÍA

[1] Selección Efectiva de Características ECG Mediante Técnicas de Transformación no lineal: Identificación de Infarto Agudo del Miocardio. Jorge Hernando Rivera Piedrahita. Tesis de Magíster en Instrumentación Física. Facultad de Ciencias Básicas. Universidad Tecnológica de Pereira, 2006

[2] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In Mathematical Challenges of the 21st Century, 2000. URL: http://wwwstat.stanford.edu/_donoho/Lectures/AMS2000/Curses.pdf.

[3] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5:1205–1224, 2004.

[4] M. A. Carreira-Perpiñan. Continuous latent variable models for dimensionality reduction and sequential data reconstruction. PhD thesis, University of Sheffield, UK, 2001. URL: http://www.cse.ogi.edu/_miguel/papers/phd-thesis.pdf.

[5] R. Fried and V. Didelez. Latent variable analysis and partial correlation graphs for multivariate time series. Statistics & Probability Letters, 73:287–296, 2005.

[6] I. T. Jolliffe, Principal component analysis, 2nd ed., ser. Springer series in statistics.New York, NY, USA: Springer, 2002.

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

[8] K. B.N. *Recurrent algorithm separating convex hulls of two sets,*, chapter Learning algorithms in Pattern Recognition, pages 43–50. Moscow, Soviet Ratio, 1973.

[9] Y. Yang. Electrocardiogram (ecg) analysis using wavelet decomposition. Technical Report ECE/BMED.

[10] G. Castellanos. Identificación de estados funcionales en bioseñales: Voz, ecg, fonocardiografía. Technical report, Universidad Nacional de Colombia, Sede Manizales, Septiembre 2005.