Reducción de la dimensionalidad con componentes principales y técnica de búsqueda de la proyección aplicada a la clasificación de nuevos datos

Dimension Reduction with Main Components and Projetion Search Techniques Applied to the Classification of New Data

Eliana Mirledy Toro Ocampo
Ingeniera Industrial Universidad Tecnológica de Pereira. Profesora Catedrática de la facultad de Ingeniería Industrial de la Universidad Tecnológica de Pereira. Magíster en Ingeniería Eléctrica en el área de Optimización. Candidata a Magíster en Investigación de Operaciones y Estadística.
Correo electrónico: elianam@utp.edu.co

Lucas Paul Pérez Hernández
Ingeniero Electricista Universidad Tecnológica de Pereira. Profesor Catedrático de la Facultad de Ciencias Básicas de la Universidad Tecnológica de Pereira. Candidato a Magíster en Ingeniería Eléctrica.
Correo electrónico: eliana@ohm.utp.edu.co

María Elena Bernal
Ingeniera de Sistemas de la Universidad Nacional a Distancia. Profesora Catedrática de la Facultad de Ingeniería de la Universidad Tecnológica de Pereira. Candidata a Magíster en investigación de Operaciones y Estadística.
Correo electrónico: mbernal@utp.edu.co

Clasificación del artículo: investigación

Fecha de recepción: 13 de abril de 2007

Fecha de aceptación: 17 de julio de 2007

Palabras clave: análisis discriminante, redes neuronales, técnica de componentes principales, técnica de búsqueda de proyección.

Key words: discriminator analysis, neural networks, principal component analysis (CP), and projection pursuit (BP).

Resumen

En muchas ocasiones el investigador se ve enfrentado a una gran cantidad de datos que describen un fenómeno. Cuando se ha caracterizado el conjunto de datos y se requieren clasificar nuevos individuos aparecen técnicas tales como el Análisis Discriminante, aunque no siempre es posible aplicarla; por esta razón las redes neuronales aparecen como una técnica alternativa para discriminar conjuntos de datos. En este artículo se muestran los resultados obtenidos al entrenar y validar una red neuronal con las componentes principales de una base de datos

multivariada y con las proyecciones obtenidas por medio de la técnica de búsqueda de proyección, la tasa de error de clasificación es el parámetro que mide la calidad de las respuestas y el nivel de aprendizaje de la red. Para evaluar, comparar y validar los resultados se tomó una base de datos compuesta por 20 variables y 24.474 datos. Se obtuvieron excelentes resultados, en los que se destacan los obtenidos usando búsqueda de la proyección.

Abstract

In this paper two dimensional data reduction techniques were compared: Principal Component Analysis – CP *(from "Componentes Principales" in Spanish)*, and Projection Pursuit – BP *(from "Búsqueda de Proyección" in Spanish)*. Both CP and BP were used in different data bases. The results obtained with these techniques were used as artificial neural network inputs in order to classify new objects. The best results were obtained when the neural network was feed with indexes taken from BP, due to the fact that this methodology takes the whole information and represents it through two indexes which group the totality of it; on the other hand, CP discards some part of the information in order to diminish the dimensionality of the original database. To evaluate the quality of the responses the error rate was taken as the parameter of classification.

Furthermore, the strategy suggested is an alternative for the cases in which the nature of data does not allow to perform Discriminator Analysis to classify new objects.

* * *

1. Introducción

Describir cualquier situación real, como por ejemplo, las características físicas de una persona, la situación política y económica de un país, las propiedades de una imagen, el rendimiento de un proceso, las motivaciones del comprador de un producto, entre otras, requiere tener en cuenta simultáneamente muchas variables. El análisis de los datos multivariantes comprende el estudio estadístico de variables medidas en elementos de una población con objetivos tales como: resumir los datos mediante un conjunto de nuevas variables, encontrar grupos en los datos si existen, clasificar nuevas observaciones en los grupos definidos.

Estas técnicas tienen aplicaciones en todos los campos científicos. En las ciencias económicas y empresariales se utilizan para cuantificar el desarrollo de un país, construir tipologías de clientes e identificar las dimensiones del desarrollo económico. En Ingeniería para controlar procesos de fabricación, diseñar máquinas más inteligentes que reconozcan formas, caracteres o imágenes y construir clasificadores que aprendan interactivamente del entorno. En ciencias de la computación para desarrollar sistemas de inteligencia artificial y redes neuronales más eficientes que resuman información y diseñen sistemas que clasifican automáticamente mediante reconocimiento de patrones. En medicina para construir procedimientos automáticos de ayuda diagnóstica y reconocimiento de tumores en imágenes digitales. En psicología para interpretar resultados de pruebas sicotécnicas y construir escalas. En sociología y ciencia política para analizar encuestas de actitudes y opiniones, y para identificar el peso de distintos factores en comportamientos sociales y políticos [1].

Los trabajos que se presentan a nivel estadístico manejan el análisis de componentes principales, como una técnica de reducción de la dimensionalidad, y separadamente presentan el Análisis Cluster y el Análisis Discriminante como técnicas de agrupamiento para clasificar nuevos datos. Las redes neuronales se presentan como una herramienta para realizar pronósticos o para hacer asignación de nuevos objetos considerando como conjunto de entrada todas las variables disponibles que describan el fenómeno en estudio.

En este trabajo se propone la combinación de las técnicas de Análisis Multivariado con las redes Neuronales a fin de disminuir los errores de clasificación de nuevos datos, tomando como conjunto de entrada las componentes principales que expliquen en un buen porcentaje la varianza de los datos o los índices de proyección obtenidos con la técnica de Búsqueda de Proyección.

En este documento se muestran los resultados obtenidos al manipular una base de datos de grandes dimensiones, mediante la aplicación de dos técnicas de reducción: componentes principales y técnica de búsqueda de proyección. Con base en las respuestas obtenidas estas técinas van a ser toma como datos de entrada para una red neuronal "Back Propagation" y comparar la eficiencia con base en los errores obtenidos.

2. Técnicas de reducción

2.1. Componentes principales

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información. Éste es el principal objetivo de la técnica y fue desarrollado por Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901).

Las componentes principales son nuevas variables con las siguientes propiedades:

- Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas originales, y la varianza generalizada de los componentes es igual a la original.
- La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.

  La varianza de la variable h es λ_h y la suma de las varianzas de las variables originales es $\sum_{i=1}^{p}\lambda_i$ donde p es el número de variables. La proporción de variabilidad total explicada por la componente h es $\frac{\lambda_h}{\sum\lambda_i}$
- Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio.

  $$COV(z, x) = \lambda_i a_i$$

  Donde a_i es el vector de coeficientes de la componente z_i.
- La correlación entre una componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.
- Las r componentes principales $(r<p)$ proporcionan la predicción lineal óptima con r variables del conjunto de variables X.
- Si se estandarizan las componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

Las componentes principales consideran una nube de puntos compuesta por las observaciones de p variables (ver figura 1).

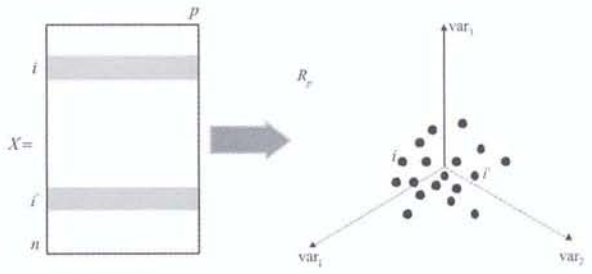

Figura 1. Representación gráfica de una nube de puntos

El objetivo de las componentes principales es proyectar la nube de puntos sobre un subespacio (un espacio bidimensional equivalente) que conserve el máximo de información de la nube original (ver figura 2).

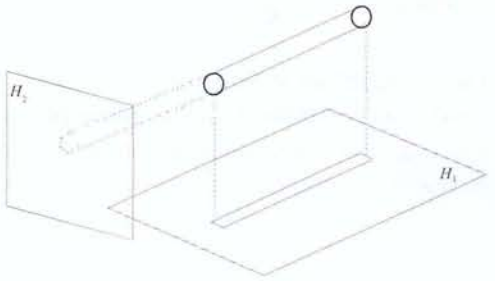

Figura 2. Proyección de la nube de puntos sobre un plano

Las componentes principales son nuevas variables que se representan a través de combinaciones lineales de las variables originales.

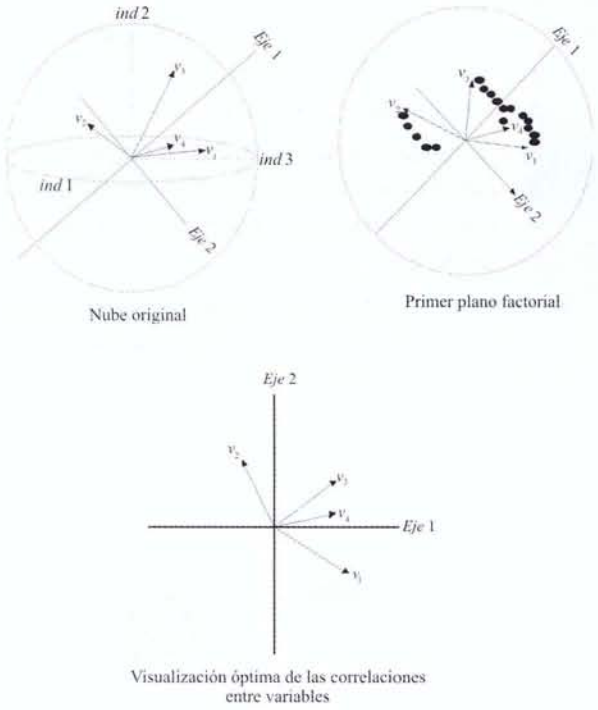

Figura 3. Representación de la nube de puntos original y de un plano factorial

2.2. Búsqueda de proyección

Freidman y Tukey (1974) describieron la búsqueda de proyección como una ruta de búsqueda para explorar una estructura multidimensional de datos no lineal examinando muchas posibles proyecciones en dos planos. La idea es que la proyección bidimensional ortogonal de los datos debe mostrar la estructura original de los datos. La técnica de búsqueda de proyección puede ser usada para obtener proyecciones en una dimensión, pero sólo se mostrará el caso bidimensional.

Extensiones de este método se describen en la literatura por los siguientes autores: Friedman (1987), Posse (1995), Huber (1985) y Jones y Sibson (1987). En este trabajo se desarrollará el método de Posse (1995). El análisis de datos explorando la búsqueda de proyección (PPEDA) logra encontrar muchas proyecciones interesantes, pero la calidad de la proyección se mide por un índice. En muchos casos el interés es la no normalidad, entonces el índice de proyección mide la salida desde la normalidad. El índice es conocido como el índice Chi cuadrado y es desarrollado en la metodología de Posse.

El método de PPEDA consiste básicamente de dos partes:

- Un índice de la búsqueda de proyección que mide el grado de la estructura (o salida desde la normalidad).
- Un método para encontrar la proyección que produce el mayor valor del índice.

Posse usa una búsqueda aleatoria para localizar el punto óptimo global del índice de proyección y lo combina con las estructuras retiradas de Freidman que logren una secuencia interesante en proyecciones bidimensionales. Cada proyección encontrada muestra una estructura que es menos importante (en términos del índice de proyección) que el anterior.

La notación para describir el método PPEDA es:

X es una matriz *nxd*, donde cada fila x_i corresponde a una observación d-dimensional y *n* es el Tamaño de la muestra.

Z es la versión esferada de ***X***

$\hat{\mu}$ es una muestra promedio de tamaño *1xd:*

$\hat{\mu} = \sum \frac{x_i}{n}$ $\hat{\sum}$ es la media de la matriz de covarianza

$$\hat{\sum}_{ij} = \frac{1}{n-1} \sum (X_i - \hat{\mu})(X_j - \hat{\mu})^T$$

α, β son vectores ortonormales ($\alpha^T \alpha = 1 = \beta^T \beta$ *y* $\alpha^T \beta = 0$) son vectores d-dimensionales que genera el plano de proyección.

$P(\alpha, \beta)$ es el plano de proyección generado por α, β.

Z_i^{α}, Z_i^{β} son la proyección esferada de las observaciones sobre los vectores α, β.

$z_i^{\alpha} = z_i^T \alpha \quad z_i^{\beta} = z_i^T \beta$

(α^*, β^*) denota el plano donde el índice es máximo.

$PI_{\chi^2}(\alpha, \beta)$ denota la proyección del índice Chi-cuadrado evaluado usando los datos de proyección sobre el plano espaciado por α y β.

ϕ_2 es la función de densidad normal bivariada.

c_k es la probabilidad evaluada sobre la *k-th* región usando la función normal estándar bivariada,

$$c_k = \int\int_{B_k} \phi_2 \, dz_1 \, dz_2$$

B_k es una caja en el plano de proyección.

I_{B_k} es la función indicadora de la región B_k, donde esta es cada una de las particiones del plano $\eta_j = \pi j / 36$, $j = 0,...,8$ es el ángulo por el cual el dato es rotado en el plano antes de ser asignado a una de las regiones B_k.

$\alpha(\eta_j)$ *y* $\beta(\eta_j)$

$\alpha(\eta_j) = \alpha \cos\eta_j - \beta \sin\eta_j$

$\beta(\eta_j) = \alpha \sin\eta_j + \beta \cos\eta_j$

C es un escalar que determina el tamaño del vecindario alrededor de (α^*, β^*) que es visitado en la búsqueda de planos a fin de encontrar mejores valores del índice de proyección.

V es un vector uniformemente distribuido sobre la unidad d-dimensional de la esfera.

γ especifica el número de iteraciones sin un incremento en el índice de proyección al mismo tiempo que el tamaño del vecindario es dividido.

m representa el número de búsquedas aleatorias para encontrar el mejor plano.

2.2.1. Índice de proyección

La PP se realiza a través de la proyección de las variables en diferentes hiperplanos para encontrar el más interesante según el Índice de proyección Chi-cuadrado. Este procedimiento se realiza a través de dos etapas:

Etapa 1: Búsqueda de la no-normalidad de los datos. El plano es dividido en 48 regiones o cajas B_k distribuidas en anillos (figura1), cada una con un ancho angular de 45° y ancho radial $\sqrt{2\log 6/5}$, el cual garantiza que cada región tenga aproximadamente la misma probabilidad (1/48) para la distribución normal bivariante. El índice de proyección está dado por

$$PI_{\chi^2}(a_2, b_2) = \frac{1}{9} \sum_{J=1}^{8} \sum_{K=1}^{48} \frac{1}{C_K} \left[\frac{1}{n} \sum_{i=1}^{n} I_{B_K}\left(Z_i^{\alpha(\eta_j)}, Z_i^{\beta(\eta_j)}\right) - C_K \right]^2$$

$$(a_2, b_2)$$

$$PI_{\chi^2}(\alpha, \beta) = \frac{1}{9} \sum_{J=1}^{8} \sum_{K=1}^{48} \frac{1}{C_K} \left[\frac{1}{n} \sum_{i=1}^{n} I_{B_k}\left(Z_i^{\alpha(\eta_j)}, Z_i^{\beta(\eta_j)}\right) - C_K \right]^2 (1)$$

C_k : probabilidad evaluada sobre una región k usando distribución normal bivariada.

n : número de datos.

I_{Bk}: es la función indicadora de la región B_k, donde ésta es cada las particiones del plano.

Z_i^{α} : son las observaciones proyectadas de cada uno de los datos sobre los vectores α y β, donde estos dos últimos son dos vectores ortonormales que son la base del plano de proyección.

η_j: es el ángulo por el cual el dato es rotado en el plano antes de ser asignado a una de las regiones B_k.

Una de las ventajas del uso del índice Chi-cuadrado es que no se ve afectado en gran forma por outliers.

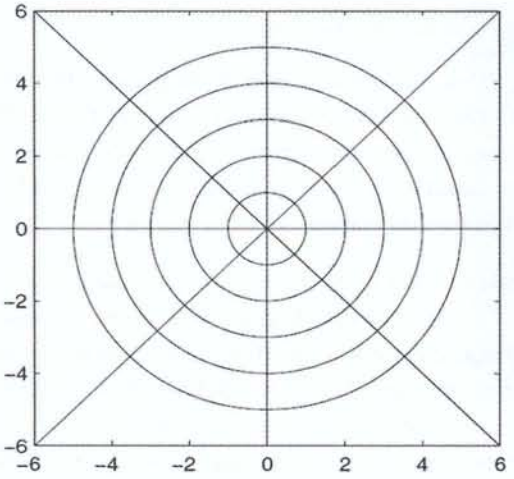

Figura. 4. División del plano donde serán proyectados los puntos del experimento

Buscar la proyección que da como resultado el mayor de los índices Chi-Cuadrado.

Etapa 2: búsqueda de la estructura. El algoritmo inicializa aleatoriamente los vectores α y β, para crear un primer mejor plano (α^* y β^*), luego se genera dos planos vecinos dados por las ecuaciones (2) y se evalúa el índice Chi-cuadrado para ellos, si uno de ellos presenta una mejoría en el índice éste será el nuevo mejor plano, de lo contrario se generan dos nuevos planos vecinos. Si después de cierto número de iteraciones no ha habido mejoría entonces se reduce el tamaño del vecindario de búsqueda a través de la disminución del parámetro c.

$$a_1 = \frac{\alpha^* + cv}{\left\|\alpha^* + cv\right\|} \quad b_1 = \frac{\beta^* - (a_1^T \beta^*) a_1}{\left\|\beta^* - (a_1^T \beta^*) a_1\right\|}$$
$$a_2 = \frac{\alpha^* - cv}{\left\|\alpha^* - cv\right\|} \quad b_2 = \frac{\beta^* - (a_2^T \beta^*) a_2}{\left\|\beta^* - (a_2^T \beta^*) a_2\right\|} \qquad (2)$$

2.2.2. Procedimiento para encontar el índice de proyección

- Se esferan los datos usando la siguiente transformación

$$Z_i = \Lambda^{-1/2} Q^T (X_i - \hat{\mu})$$

  Donde las columnas de Q son los vectores propios obtenidos desde $\hat{\sum}$, Λes una matriz diagonal que corresponde a los valores propios, y $X_i$es la i-ésima observación.
- Genera un plano aleatorio, (α_0, β_0). Éste es el plano incumbente (α^*, β^*)
- Evalúa el índice de proyección $PI_{\chi^2}(\alpha_0, \beta_0)$ para el plano inicial.
- Genera dos planos candidatos (a_1, b_1) y (a_2, b_2) con base en la ecuación (2).
- Evalúa el valor del índice de proyección en esos planos $PI_{\chi^2}(a_1, b_1)$ y $PI_{\chi^2}(a_2, b_2)$.
- Si uno de los planos candidatos tiene índice de proyección mayor, entonces ese plano se convierte ahora en el plano incumbente (α^*, β^*).
- Se repiten los pasos 4 al 6 mientras hayan mejoramientos en el índice de proyección.
- Si el índice no mejora en γ iteraciones, entonces decrece el valor de C a la mitad.
- Se repiten los pasos 4 al 8 hasta que C decrezca hasta un límite especificado por el analista.

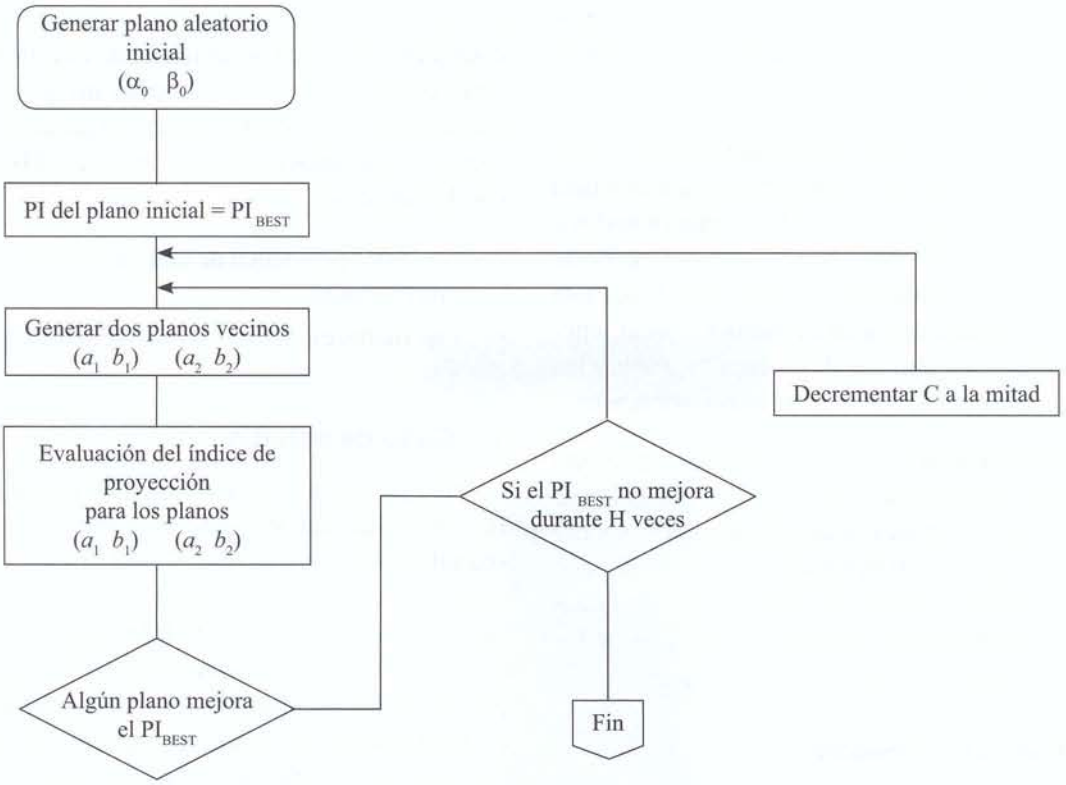

Diagrama 1. Algoritmo para encontrar el Índice de Proyección

3. Técnicas clasificación

3.1. Redes neuronales

Las neuronas es un sistema biológico que está formado por neuronas de entrada o censores conectados a una compleja red de neuronas que "calculan", o neuronas ocultas, las cuales, a su vez, están conectadas a las neuronas de salidas que, por ejemplo, son las encargadas de controlar los músculos. Por censores se entienden señales de los sentidos (oído, vista, etc.), las respuestas de las neuronas de salida activan los músculos correspondientes. En el cerebro hay una gigantesca red de neuronas "calculadoras" u ocultas que realizan la computación necesaria. De manera similar, una red neuronal artificial debe ser compuesta por censores del tipo mecánico o eléctrico.

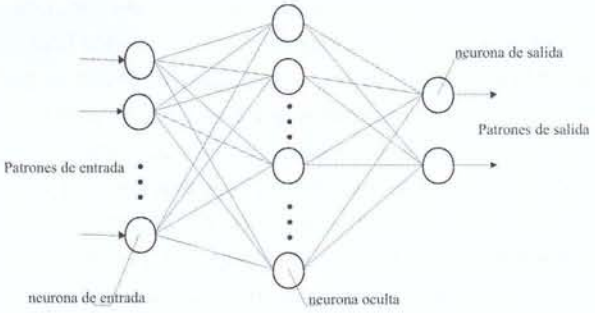

Figura 5. Red neuronal artificial típica

Las Redes Neuronales Artificiales (Artificial Neural Networks) son sistemas paralelos para el procesamiento de la información; están inspirados en el modo en el que las redes de neuronas biológicas del cerebro procesan la información, es decir, se han intentado plasmar los aspectos esenciales de una neurona real

a la hora de diseñar una neurona "artificial". Estos modelos realizan una simplificación, desentrañando cuáles son las relevancias del sistema.

La definición más general considera a una Neural Network como un entramado o estructura formada por muchos procesadores simples llamados nodos o neuronas, los cuales están conectados por medio de canales de comunicación o conexiones. Cada una de ellas tiene una cantidad de memoria local, operando solamente con sus datos locales y sobre las entradas que recibe a través de esas conexiones.

Las Redes Neuronales llevan asociadas algún tipo de regla de aprendizaje o entrenamiento particular por la cual esas conexiones son ajustadas acorde con los ejemplos proporcionados. En otras palabras, éstas aprenden a partir de ejemplos, y muestran alguna capacidad para generalizar más allá de los datos mostrados [6 y 7].

3.2. Análisis discriminante

El análisis discriminante es una de las técnicas estadísticas usadas para probar las hipótesis de igualdad de medias de dos o más grupos. Esta técnica desarrollada por R. Fisher en 1963 se convierte en la más apropiada, cuando la variable dependiente es categórica (cualitativa); las variables independientes son métricas(cuantitativas). La idea básica de AD es encontrar una combinación lineal de las variables independientes que haga que se maximicen los puntajes promedio de las categorías de la variable dependiente. Esta combinación lineal se conoce con el nombre de Función Discriminante. En símbolos, X son las variables independientes y B son los coeficientes discriminantes. El objetivo es encontrar los valores de estos B, los cuales dan la FD (función discriminante) requerida.

Básicamente un análisis discriminante consiste en obtener funciones lineales de las variables independientes, denominadas funciones discriminantes que permitan clasificar a las muestras en una de las subpoblaciones o grupos establecidos por los valores de la variable dependiente. Al igual que cuando se realiza un ANOVA, hay ciertas condiciones que se deben verificar antes de realizar un análisis discriminante; "aunque el hecho de que algunas de estas condiciones no se cumplan no quiere decir que el análisis no tenga validez, especialmente en el caso de tamaños de muestra grandes" [16]. Estas condiciones son:

- Los datos provienen de una distribución normal multivariada.
- Las matrices de covarianza de los grupos son iguales.

4. Caso de estudio

La base de datos que se manipuló en este trabajo fue obtenida del Website disponible en www.spatial-econometrics.com, el cual provee a los investigadores en estadística bases de datos de difícil análisis para probar las metodologías desarrolladas dentro de marcos de proyectos de investigación o propias. A fin de comprobar cuál de las dos técnicas de reducción de la reducción de dimensionalidad presenta mejor eficiencia, se enfrentaran los resultados obtenidos con ambas por medio de un problema de clasificación. Este problema consiste en clasificar según un conjunto de datos cuál debe ser el área del terreno medido en acres.

Para la solución de este problema de clasificación se proponen dos topologías de redes neuronales artificiales como se muestra en la figura 6, una para realizar la clasificación usando los resultados obtenidos del análisis de componentes principales y el otro usando los resultados obtenidos del análisis de búsqueda de proyección.

El entrenamiento de ambas redes neuronales es realizado usando un algoritmo de aprendizaje de propagación hacia atrás (algoritmo "back propagation"). Las funciones de transferencia de cada una de las neuronas son de tipo logarítmica sigmoidal, debido a que los datos con los que se está tratando son todos mayores a cero. El número de capas ocultas y el número de neuronas en cada capa oculta, son usados como variables para realizar una búsqueda en malla, a fin de determinar la mejor configuración de estas variables que ofrezca el mínimo error. En

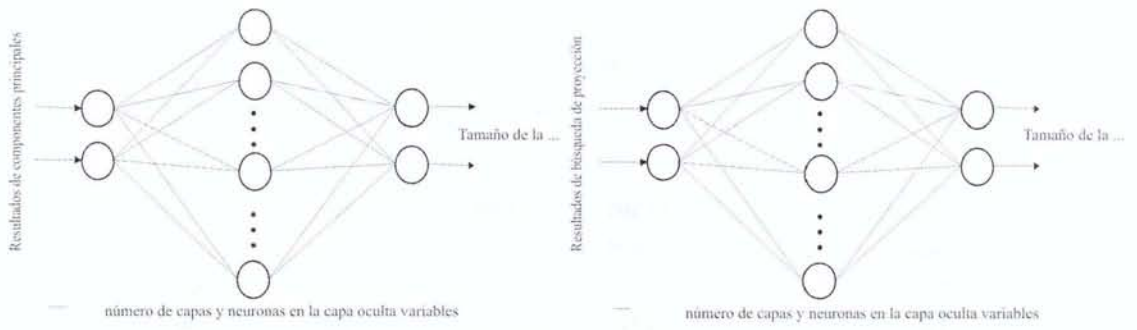

Figura 6. Redes neuronales empleadas

el proceso de búsqueda en malla se tienen valores de variación para α, que corresponde al número de capas ocultas en la neurona, entre 1 y 10, y para β, que corresponde al número de neuronas en cada capa oculta, entre 1 y 10, por lo que el número de operaciones que se van a realizar es de 100 para encontrar la mejor configuración ante la variación de dichos parámetros.

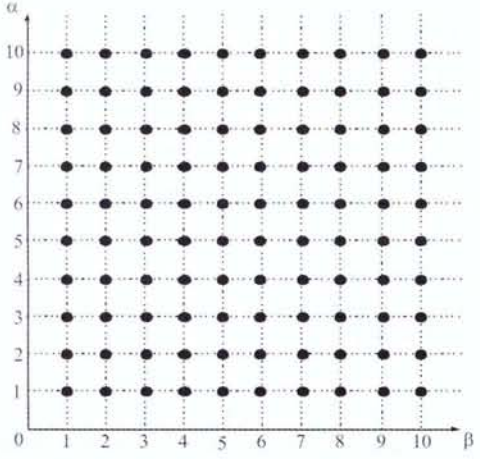

Figura 7. Espacio de búsqueda de la mejor configuración para la red neuronal

5. Resultados

La base de datos usada en este trabajo contiene veinte variables asociadas al campo, la agricultura, ganadería, y avicultura, de cada una de estas variables se tomaron 24.474 experimentos, por lo tanto, el número de datos que deben ser analizados es de 489.480 datos. Para realizar el Análisis Discriminante y la técnica de Componentes Principales se utilizó el software SPSS 11.0; para encontrar los índices de Proyección y codificar la red neuronal se utilizó el software Mat Lab 7.0.

Los resultados obtenidos con la técnica de Análisis Discriminante no fueron satisfactorios, a continuación se presentan los resultados obtenidos en el SPSS.

Resultados de la prueba

M de Box		509697,8
F	Aprox.	1212,165
	gl1	420
	gl2	1,3E+09
	Sig.	,000

Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.

El test M de Box sirve para contrastar la igualdad de las matrices de varianza covarianza, de tal manera que si el *p*-valor es menor de 0,05 se rechazaría la hipótesis de igualdad de matrices de varianzas-covarianza. (*p*-valor=0,000) es menor a 0,05. No se cumple el supuesto de homogeneidad de la varianza en las poblaciones analizadas. Por lo tanto, no se puede aplicar la técnica de Análisis Discriminate para clasificar nuevos individuos, razón por la cual es necesario utilizar otra técnica de clasificación y se procede a hallar los vectores que alimentarán la red neuronal (componentes principales y ángulos de proyección).

Usando el método de análisis de componentes principales se determina que para explicar el experimento con una exactitud de 99,9618% se deben usar las dos primeras componentes. Resultados obtenidos en el SPSS.

Matriz de componentes rotados(a)

	Bruta		Reescalada	
	Componente		Componente	
	1	2	1	2
Constante	27,509	-1,096	,409	-,016
Conservación	1,628	-,926	,099	-,056
Marcha	4,322	-,142	,359	-,012
Pasto	18,241	-1,309	,409	-,029
Rango de terreno	7,646	-1,251	,247	-,040
Madera	20,213	-1,714	,461	-,039
Suelo mejorado	2,354	-,388	,242	-,040
No mejorado	,778	-,190	,220	-,054
Barbecho	,185	-,675	,017	-,060
Otros terrenos	24,096	-,868	,448	-,016
FincaS	38,318	,834	,429	,009
Ganado	16,291	-1,780	,322	-,035
vacas lecheras	2,399	-,730	,196	-,060
Porcinos	1,480	-,745	,178	-,089
Ovejas	1,018	-,133	,195	-,026
Gallinas	1,683	-,110	,396	-,026
CaballoS	8,039	,389	,384	,019
Dueños	2199,954	11503,649	,188	,982
Área rural	98,770	-14,150	,462	-,066
Área de granjas	3058,831	298,834	,995	,097

Método de extracción: análisis de componentes principales.
Método de rotación: normalización Varimax con Kaiser.
La rotación ha convergido en 3 iteraciones.

Usando el método de búsqueda de la proyección se determina que si se reduce la dimensionalidad del experimento a un espacio bidimensional equivalente es posible representarlo en 100% usando dos variables. Se hicieron diferentes corridas y se varió el número de iteraciones totales de menor a mayor número, para determinar cuál configuración genera mejores planos de proyección bidimensionales, entendiendo como mejor plano aquellos que poseen el mayor índice de proyección. Los resultados de las diferentes corridas se muestran en la tabla 1.

Tabla 1. Resultados de los índices de proyección para diferente número de iteraciones

Numero de iteraciones	Índice de proyección para α^*	Índice de proyección para β^*
10	1,4328	1,1398
20	2,2304	1,3240
30	2,2558	1,4370
70	2,4930	1,5401

De la tabla anterior se deduce entonces que el mejor espacio bidimensional es el generado por los ejes obtenidos en 70 iteraciones, ya que éstos poseen el mayor índice de proyección de todo el conjunto. Teniendo los resultados de reducción de dimensionalidad entregados por las dos metodologías, se puede determinar la configuración de la red neuronal que mejor se desempeña en el proceso de clasificación para los dos casos. En la tabla 2 se muestran los resultados de algunas de las configuraciones que presentan errores bajos de entrenamiento, si se usan los datos obtenidos del análisis de componentes principales.

Tabla 2. Resultados de la búsqueda en malla para Componentes Principales

Capas ocultas	Neuronas en la capa oculta	Error
1	6	4.532
1	5	5.560
2	4	5.961
3	5	6.425

Según la tabla anterior, la topología de la red neuronal que se va a utilizar para realizar la clasificación es de dos entradas, una capa oculta, seis neuronas en la capa oculta y dos salidas. Para determinar la mejor configuración de la red neuronal se usan los datos obtenidos del análisis de búsqueda de proyección (ver tabla 3).

Tabla 3. Resultados de la búsqueda en malla para los índices de Búsqueda de Proyección

Capas ocultas	Neuronas en la capa oculta	Error
1	7	3.901
2	4	4.056
2	3	4.551
3	3	5.482

Según la tabla 3, la topología de la red neuronal que se va a utilizar para realizar la clasificación es de dos entradas una capa oculta, siete neuronas en la capa oculta y dos salidas.

Si se usan las dos topologías mencionadas anteriormente se procede a encontrar el error de clasificación con los datos que no fueron usados en el proceso de entrenamiento. Esta clasificación se realiza usando los datos encontrados del análisis de componentes principales y de la búsqueda de proyección. Para la determinación del error de clasificación se usa la siguiente expresión la cual combina los datos de validación y los datos que fueron bien clasificados, de la siguiente forma:

$$\%\ error = \frac{\text{Datos de validación} - \text{Datos bien clasificados}}{\text{Datos de validación}} * 100$$

El proceso de entrenamiento fue realizado usando 12.474 datos, y la clasificación usando los restantes 12.000 datos. Los resultados obtenidos para la clasificación se muestran en la tabla 4.

Tabla 4. Tasas de error de clasificación

Método	Datos de validación	Datos bien clasificados	% de error
Componentes principales	12.000	10.950	8,75
Búsqueda de proyección	12.000	11.652	3,04

En ésta se nota una mejor clasificación al usar los datos obtenidos utilizando el análisis de búsqueda de proyección.

6. Conclusiones

- Cuando el conjunto de variables está representada por valores en diferentes escalas es necesario tipificar la base de datos y tomar como base la matriz de covarianza para aplicar la técnica de componentes principales.
- Cuando no es posible aplicar el análisis discriminante a un conjunto de datos la técnica de redes neuronales se plantea como una alternativa eficiente para clasificar nuevos individuos.
- Si se comparan los resultados obtenidos teniendo como datos de entrada los componentes

principales o los ángulos de proyección α y β se obtienen mejores resultados con los ángulos de proyección; esto en razón a que la técnica de proyección considera absolutamente todos los datos y los componentes principales, sólo toma en cuenta un porcentaje de la información.

- Aunque la base de datos usada presenta una alta complejidad en cuanto a análisis y manejo de la información, las técnicas de reducción de dimensionalidad usadas permitieron hacer un adecuado proceso de clasificación, obteniendo excelentes resultados

Agradecimientos

Los autores expresan sus agradecimientos a la Universidad Distrital Francisco José de Caldas por facilitar los medios para esta publicación y a la Universidad Tecnológica de Pereira por su apoyo a las Maestrías en Ingeniería Eléctrica e Investigación de Operaciones y Estadística.

Referencias bibliográficas

[1] Peña Daniel. (2002) *Análisis de datos multivariantes*. Madrid: Editorial McGraw-Hill.

[2] Ferrán Aránez Magdalena. (2001) *Análisis estadístico con SPSS*. Editorial Osborne – McGraw-Hill.

[3] Jiménez L. and Landgrebe, D. A. (October 1995) Projection Pursuit in High Dimensional Data Reduction: Initial Conditions, Feature Selection and the Assumption of Normality. Presented at the IEEE International Conference on Systems, Man and Cybernetics, Vancouver Canada.

[4] Johnson A. Richard, Wichern Dean. (1998) *Applied Multivariate Statistical Analysis*. New Jersey: Editorial Prentice Hall.

[5] Díaz Monroy, Luis Guillermo. Estadística multivariada Inferencia y métodos. Universidad Nacional de Colombia.

[6] Toro Ocampo, Eliana, Molina, Alexander, Garcés, Alejandro. (2006) Pronóstico de bolsa de valores empleando técnicas inteligentes. Revista *Tecnura* de la Universidad Distrital Francisco José de Caldas Año 9 (18). Bogotá Semestre I de 2006.

[7] Toro Ocampo, Eliana; Mejía, Diego Adolfo; Salazar Isaza, Harold. *Pronóstico de ventas usando redes neuronales*. *Scientia et Técnica* (26). Universidad Tecnológica de Pereira, diciembre de 2004.