Maestría en Electrónica

Reconocimiento de Patrones

Práctica 1: Análisis de Componentes Principales

II Cuatrimestre, 2018

Esteban Martínez Valverde
estemarval@gmail.com

1. Investigación PCA

El análisis de componentes principales (PCA, por sus siglas en inglés), es un procedimiento matemático que transforma un número de variables (posiblemente) correlacionadas en un número más pequeño de variables no correlacionadas llamadas componentes principales. El primer componente principal representa la mayor variabilidad posible en los datos [2].

El PCA se puede considerar como una rotación de los ejes del sistema de coordenadas de la variable original a unos nuevos ejes ortogonales, llamados .[ej]es principales". Como se observa en la Figura 1, el punto y_{1m} es la proyección del punto (x_{1m}, x_{2m}) en el eje definido por Y_1.

Tradicionalmente, el PCA se realiza en una matriz simétrica cuadrada. Puede ser una matriz SSCP (sumas puras de cuadrados y productos cruzados), matriz de covarianza (sumas de cuadrados escaladas y productos cruzados), o sumas de cuadrados de matriz de correlación y productos cruzados a partir de datos estandarizados).

El análisis de componentes principales es similar a otro procedimiento multivariante llamado *Factor Analysis*. A menudo se confunden y muchos científicos no entienden la diferencia entre los dos métodos o qué tipos de análisis son los más adecuados. La diferencia reside en que el *Factor Analysis* es un procedimiento estadístico para identificar las interrelaciones que existen entre un gran número de variables, es decir, para identificar cómo se relacionan conjuntos de variables[2].

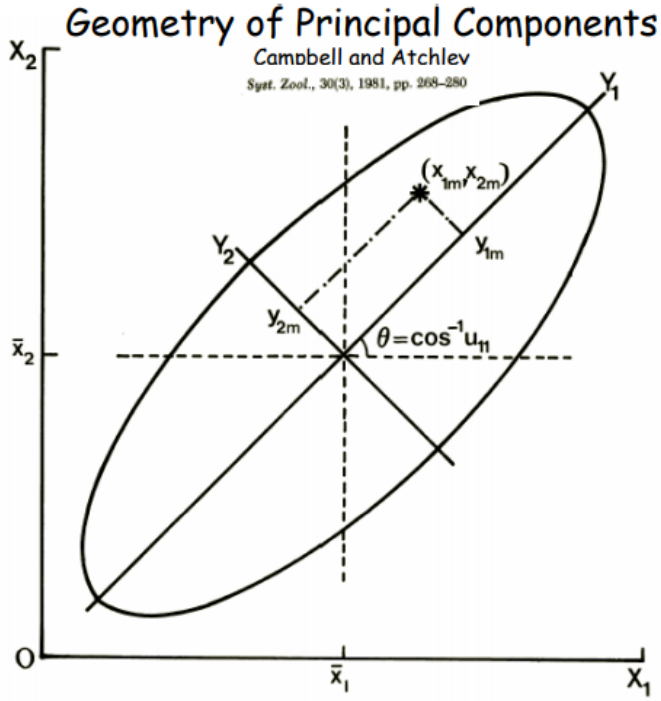

Figura 1: Representación idealizada de un diagrama de dispersión para dos variables [3].

2. Casos de uso de PCA

A continuación se mencionan dos aplicaciones que utilizan PCA:

2.1. Detección y visualización de ataques a redes informáticas

Los datos de tráfico de red recopilados para el análisis de intrusión son típicamente de gran dimensión, lo que dificulta su análisis y visualización. El PCA se aplica a los ataques de red seleccionados de los conjuntos de datos de detección de intrusos DARPA 1998, conocidos como: ataques de *Denial-of-Service* y *Network Probe*. Se propone un método para identificar un ataque basado en las estadísticas generadas. La visualización de la actividad de la red y las posibles intrusiones se logra utilizando Bi-plots, que proporciona un resumen de las estadísticas [5].

2.2. PCA aplicado a la teledetección *Remote Sensing*

Se utiliza el análisis de PCA para obtener información de la cobertura del suelo a partir de imágenes de satélite. Tres imágenes Landsat fueron seleccionadas a partir de dos áreas que se encuentran en los municipios de Gandia y Vallat (Valencia, España). Inicialmente, se utilizó sola una imagen Landsat del año 2005. Posteriormente, se utilizaron dos imágenes Landsat de los años 1994 y 2000. El principal objetivo es analizar los cambios más grandes en la cobertura de la tierra. Como resultado se obtuvo que el segundo componente principal de la imagen de área Gandia permitió la detección de la presencia de vegetación. El mismo componente en el área de Vallat permitió detectar un área forestal afectada por un incendio forestal. En consecuencia, en este estudio se confirmó la viabilidad del uso de PCA en teledetección para extraer la información territorial [4].

3. Aplicación de PCA

El MPI (Índice de Pobreza Multidimensional), es un indice de complementa las medidas monetarias de pobreza al considerar las privaciones superpuestas que sufren los individuos al mismo tiempo. El índice identifica privaciones en las mismas tres dimensiones que el HDI (Indice de desarrollo humano) y muestra el número de personas que son multidimensionalmente pobres (que sufren privaciones en el 33 % o más de los indicadores ponderados) y el número de privaciones ponderadas con las que los hogares pobres suelen lidiar [1].

Se tiene un Dataset que cuenta con el MPI de varios paises con enfoque rural y urbano. Los 8 atributos se listan acontinuación: Código País, Nombre País, MPI Urbano, *Headcount Ratio Urban*, Intensidad de deprivación Urbano, MPI Rural, *Headcount Ratio Rural*, Intensidad de deprivación Rural. Se agrega al Dataset un atributo, ”Nivel”, el cuál puede tener 2 valores: Nivel 1 (pobreza extrema) y Nivel 2 (pobreza). Estos valores fueron obtenidos de los estudios realizados por las Naciones Unidas.

De esta manera se realiza una implementación PCA para identificar de el nivel de pobreza de un país utilizando los 8 atributos del Dataset original.

En la Figura 2, se muestra el Dataset con el atributo *Nivel* agregado.

	ISO	Country	MPI Urban	Headcount Ratio Urban	Intensity of Deprivation Urban	MPI Rural	Headcount Ratio Rural	Intensity of Deprivation Rural	Nivel
0	KAZ	Kazakhstan	0.000	0.0	33.3	0.000	0.09	33.3	Nivel_1
1	SRB	Serbia	0.000	0.1	41.4	0.002	0.50	40.3	Nivel_1
2	KGZ	Kyrgyzstan	0.000	0.1	40.2	0.003	0.70	37.1	Nivel_1
3	TUN	Tunisia	0.000	0.1	35.6	0.012	3.18	38.7	Nivel_1
4	ARM	Armenia	0.001	0.2	33.3	0.001	0.39	36.9	Nivel_1

Figura 2: Carga del MPI *Dataset*, al *DataFrame* de Panda.

En la Figura 3, se muestra la selección de los *Features* para aplicar el PCA.

	MPI Urban	Headcount Ratio Urban	Intensity of Deprivation Urban	MPI Rural	Headcount Ratio Rural	Intensity of Deprivation Rural
0	-0.840300	-0.913202	-1.639400	-1.072205	-1.206570	-1.547421
1	-0.840300	-0.907769	-0.054480	-1.062216	-1.194186	-0.746509
2	-0.840300	-0.907769	-0.289283	-1.057222	-1.188145	-1.112640
3	-0.840300	-0.907769	-1.189361	-1.012271	-1.113237	-0.929575
4	-0.829574	-0.902337	-1.639400	-1.067211	-1.197508	-1.135524

Figura 3: *Features* escogidos para aplicar el PCA.

El atributo *Nivel* es seleccionado como *Target* para aplicar el PCA.

	principal component 1	principal component 2	Nivel
0	-2.940579	0.535639	Nivel_1
1	-1.973066	0.005235	Nivel_1
2	-2.211200	0.183007	Nivel_1
3	-2.444252	0.179670	Nivel_1
4	-2.757769	0.371897	Nivel_1

Figura 4: Aplicación de PCA a los *features* en una proyección en 2D.

En la Figura 4, se muestran los componentes principales (en 2 dimensiones) obtenidos tras aplicar PCA. Finalmente se obtiene una representación visual, la cuál se puede observar en la Figura 5.

El repositorio con el código fuente puede ser encontrado en el: Repositorio en *Github*

El *Jupyter notebook* del ejercicio de PCA, se puede encontrar con el nombre: *aplicacion_pca_poverty.ipynb*. Mientras que el *Dataset* se puede encontrar en la dirección: */data/MPI_national_labeled.csv*.

Referencias

[1] multidimensional-poverty-index-mpi @ hdr.undp.org.

[2] Matt Brems. A one-stop shop for principal component analysis.

[3] NCSU Bioinformatics Research Center Compute Cluster. Introduction to principal components and factoranalysis.

[4] M. Teresa Sebasti´a Jesus Mengual Javier Estornell, Jesus M. Mart´ı-Gavil´a. Principal component analysis applied to remote sensing.

[5] Khaled Labib and V. Rao Vemuri. An application of principal component analysis to the detection and visualization of computer network attacks.

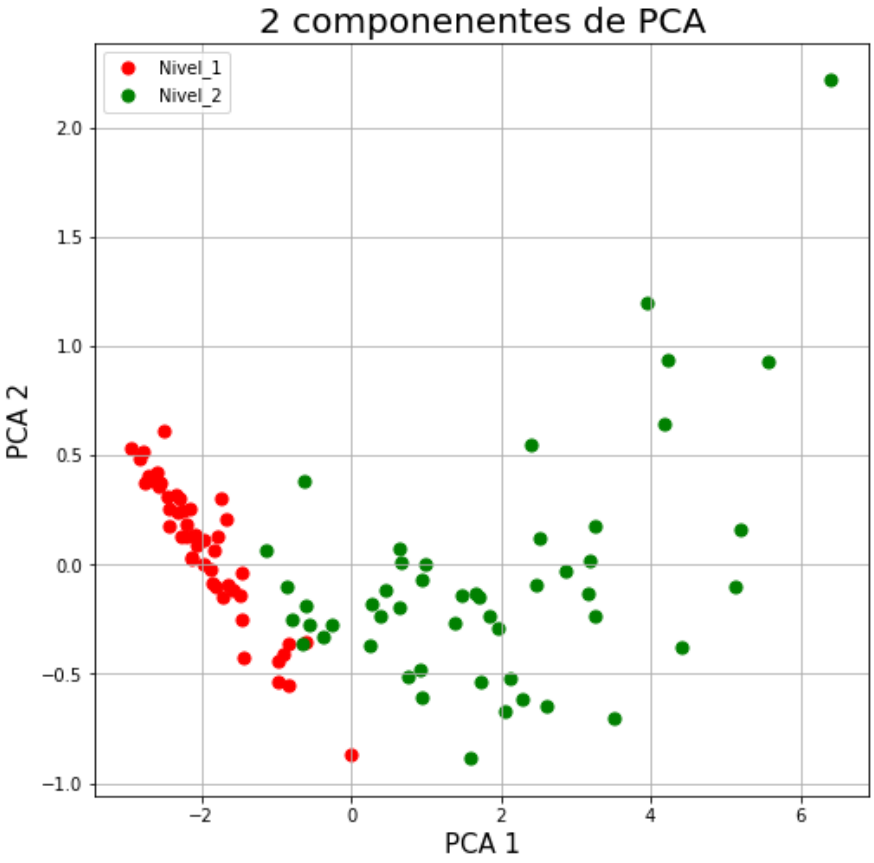

Figura 5: Visualización de la proyección en 2D de los componentes PCA.