Visualización y métodos kernel: Integrando inteligencia natural y artificial

Juan C. Alvarado-Pérez[1,2], Diego H. Peluffo-Ordóñez[3,4] & Roberto Theron[2]

juan.alvarado@aunar.edu.co[1], dhpeluffoo@unal.edu.co[2],theron@usal.es[3]
[1] Corporación Universitaria Autónoma De Nariño, Colombia,
[2] Universidad de Salamanca, España
[3] Universidad Técnica de Norte, Ecuador,
[4] Universidad de Nariño, Colombia.

Resumen

Los enormes volúmenes de datos, generados por la actividad académica, científica, empresarial e industrial, entre muchas más, contienen información muy valiosa, lo que hace necesario desarrollar procesos y técnicas robustas, de validez científica que permitan explorar esas grandes cantidades de datos de manera óptima, con el propósito de obtener información relevante para la generación de nuevo conocimiento y toma de decisiones acertadas.

La robustez y altas capacidades de procesamiento computacional de las maquinas modernas son aprovechadas por áreas como la inteligencia artificial que si se integra de forma holística con la inteligencia natural, es decir, si se combina sinérgicamente los métodos sofisticados de análisis de datos con los conocimientos, habilidades y flexibilidad de la razón humana, es posible generar conocimiento de forma más eficaz. La visualización de información propone formas eficientes de llevar los resultados generados por los algoritmos a la comprensión humana, la cual permite encontrar tendencias y patrones ocultos de forma visual, los cuales pueden formar la base de modelos predictivos que permitan a los analistas producir nuevas observaciones y consideraciones a partir de los datos existentes, mejorando el desempeño de los sistemas de aprendizaje automático, haciendo más inteligibles los resultados y mejorando la interactividad y controlabilidad por parte del usuario. Sin embargo, la tarea de presentar y/o representar datos de manera comprensible, intuitiva y dinámica, no es un tarea trivial; uno de los mayores problemas que enfrenta la visualización es la alta dimensión de los datos, entendiendo dimensión o como el número de variables o atributos que caracterizan a un objeto. Una solución efectiva son los métodos de reducción de dimensión (RD) que permiten representar los datos originales en alta dimensión en dimensiones inteligibles para el ser humano (2D o 3D). En la actualidad, los métodos *kernel* representan una buena alternativa de RD debido su versatilidad y fácil implementación en entornos de programación. En este trabajo se presenta una breve descripción y forma de uso de un método generalizado conocido como análisis de componentes principales basado en kernel (KPCA).

Palabras claves: Inteligencia artificial, inteligencia natural, kernel PCA, reducción de dimensión.

Abstract.
The large amount of data generated by different activities -academic, scientific, business and industrial activities, among others- contains meaningful information that allows developing processes and techniques, which have scientific validity to optimally explore such information. Doing so, we get new knowledge to properly make decisions. The robustness and high computational processing capabilities of modern machines are used by areas such as artificial intelligence, if holistically integrates with the natural intelligence, in other words, if is synergistically combines sophisticated data analysis methods as well as the knowledge, skills and flexibility of human reasoning, it is possible to discover knowledge in a more effective way. "Information Visualization" is an efficient way to bring the results generated by the algorithms to human understanding in order to find hidden trends and patterns belonging visually to the predictive model, that allow analysts to produce new observations and considerations from existing data, improving the performance of machine learning systems, making it more intelligible results and improving interactivity and controllability by the user. Nonetheless, the task of presenting and/or represent data in an understandable, intuitive and dynamic way, is not a trivial task, one of the biggest problems that visualization faces is the high dimension -being dimension the number of variables or attributes that characterize an object. An effective solution is the dimensionality reduction methods (RD) that can represent the original, high-dimensional data as a space whose dimension is intelligible to humans (2D or 3D). Currently, the kernel methods represent a good alternative because of their versatility and ability to make RD algorithms easily implementable in programming environments. This paper presents a brief description and method of use of a generalized method, so-called kernel principal component analysis (KPCA).

Keywords: Artificial intelligence, dimensionality reduction, kernel PCA, natural intelligence.

1. Introducción

Los grandes volúmenes de datos que provienen de diversas fuentes y que al ser combinadas pueden generar nuevos conjuntos de datos de estructura inconsistente e impredecible ha creado una nueva era informática llamada *Big Data* (Japkowicz & Stefanowski, 2016) (Sparks, Ickowicz, & Lenz, 2016), la cual conlleva nuevos desafíos en cuanto al análisis eficiente de la información. Si bien es cierto que en cuanto mayor es el tamaño del conjunto de datos, mayor es su riqueza, también es verdad que al incrementar su volumen (alta dimensión y numero registros) se incrementa la dificultad para detectar patrones, ya que las técnicas y métodos de aprendizaje automático (*Machine Learning*) pueden resultar improcedentes al conllevar tiempos de procesamiento muy elevados y/o formular modelos que pueden resultar muy abstractos y de escasa comprensión para los usuarios, sobre todo para aquellos no expertos (Keim, Mansmann, & Thomas, 2009). Intuitivamente, podría pensarse que una alternativa para enfrentar este problema es la representación de los datos en un espacio de dimensión menor que preserve la estructura del espacio original. Este

enfoque se conoce técnicamente como reducción de dimensión (RD) (Lee & Verleysen, 2007) que puede aplicarse bajo dos propósitos: Como técnica de preprocesamiento al comprimir los datos (Wang & Zhao, 2016) haciendo más eficiente la aplicación de algoritmos de *Machine learning* o *Data Mining*, de modo que el rendimiento de un sistema de reconocimiento de patrones puede ser mejorado, y a su vez como técnica de visualización inteligible representando los datos de alta dimensión, en una dimensión perceptible para los humanos (Sedlmair, Munzner, & Tory, 2013).

Las técnicas de RD, en primera instancia, se basaron en métodos lineales de identificación de patrones, los cuales son los más simples de detectar, sin embargo su simplicidad esta correlacionada con la rigidez del modelo, que puede resultar muy general y de escasa adaptabilidad a nuevos conjuntos de datos. La gran ventaja que presentan las técnicas de reconocimiento de patrones lineales es que resultan muy intuitivas y cercanas a los sistemas de percepción humana (Ware, 2012) y por tanto, es posible modelar la realidad intrínseca de los datos de forma sencilla y significativa. Sin embargo, los conjuntos de datos en el mundo real no siempre se presentan con un comportamiento lineal, de hecho, la mayoría de variables y sus relaciones tienen un comportamiento complejo y caótico por lo que surgieron técnicas para la detección de patrones de naturaleza no lineal. Dichos métodos funcionan de manera muy eficiente en bases de datos de tamaño moderado, no obstante, en aquellas de tamaño considerable, estas técnicas resultan improcedentes, por lo que se hace necesario recurrir a métodos aproximados, los cuales tienen como fundamento procedimientos de optimización convexa que comúnmente llevan a óptimos locales lejanos de la solución óptima. En respuesta y como solución a los diversos problemas de las técnicas tanto lineales como no lineales ha emergido un nuevo enfoque denominado "Aprendizaje basado en kernel*"* que integra de forma sinérgica lo mejor de los dos preceptos. En esencia dicho enfoque aplica métodos lineales de reconocimiento de patrones a un espacio de muy alta dimensión (en donde los datos son fácilmente separables) y posteriormente los proyecta a un espacio de menor dimensión, lo cual permite reconocer patrones no lineales en un conjunto de entrada. En consecuencia, poseen las ventajas de los algoritmos lineales (sencillez, estabilidad, eficiencia computacional,...) y al mismo tiempo, gozan de la flexibilidad de los algoritmos de detección de patrones no lineales gracias a un cambio en la representación de los datos. Así, el algoritmo en realidad está detectando patrones no lineales en el espacio original valiéndose de la aplicación de métodos lineales en un espacio de alta dimensión.

En la actualidad muchas áreas del conocimiento utilizan las técnicas de RD espectrales basadas en kernel, por ejemplo: El diagnóstico médico (Hu et al., 2016), datos climáticos (Pulkkinen, 2016) detección de fallos (Liu, Zhang, Yu, & Zeng, 2016) (Chen, Ding, Zhang, Li, & Hu, 2016) reconocimiento biométrico (Ambika, Radhika, & Seshachalam, 2016) bioinformática (Fu, 2014), entre otras. Como se evidencia, los datos de entrada sobre los que se puede aplicar los métodos kernel pueden ser muy diversos, haciendo que estos sean realmente versátiles. No obstante, su verdadera versatilidad consiste en la gran cantidad de

algoritmos que pueden utilizar los métodos kernel, puesto que es aplicable a cualquier procedimiento de aprendizaje no supervisado que tome como punto de partida una matriz de distancias. Actualmente, los métodos de RD se han enfocado en criterios elaborados los cuales están orientados a la preservación de la topología de los datos, representada en una matriz de similitud o afinidad, entre estos métodos podemos encontrar a *Laplacian Eigenmaps* (LE) (Belkin & Niyogi, 2003) y Locally Linear Embedding} (LLE) (Roweis & Saul, 2000), los cuales son de tipo espectral. Otros métodos emergentes tienen una connotación probabilística ya que se basan en divergencias gracias a que la matriz de similitud normalizada puede ser vista como una distribución de probabilidad, entre estos métodos se pueden distinguir *Stochastic Neighbour Embedding* (SNE) (Bunte, Haase, Biehl, & Villmann, 2012), y sus variantes y mejoras, tales como t-SNE que utilizan una distribución t-student y JSE que usa la divergencia de Jensen-Shanon (Lee, Renard, Bernard, Dupont, & Verleysen, 2013). Algunos métodos utilizan criterios de conservación de la varianza y la distancia (Borg, 2005), entre estos métodos encontramos a *Principal component analysis* (PCA) y *Classical Multidimensional Scaling* (CMDS), respectivamente. Cada técnica kernel posee algunas propiedades particulares que las hacen adecuadas para el análisis de cierto tipo de datos pero inadecuadas para otro, por ejemplo, algunos métodos serán más adecuados para representar una topología global mientras que otros representaran mejor una topología local (Peluffo-Ordóñez, Lee, & Verleysen, 2014). La elección correcta del kernel contribuye al reconocimiento de patrones más adecuado y ajustado a los requerimientos del usuario en torno a su área específica de conocimiento, es decir, la integración holística de la inteligencia natural (IN) con la inteligencia artificial(IA) (Alvarado-Pérez & Peluffo-Ordóńez, 2015) (Alvarado-Pérez, Peluffo-Ordóñez, & Therón, 2015).

El resto del documento se organiza de la siguiente manera: La sección 2 describe la forma de integrar la IA e IN y a su vez la identificación de patrones de forma lineal y no lineal. La sección 3 presenta una aproximación geométrica a la técnica de reducción de dimensión denominada PCA. La sección 4 describe la solución kernel PCA como método de integración. Por último, en la sección 5 se presentan algunos comentarios finales a modo de conclusión y trabajo futuro.

2. Integración holística entre la IA e IN y lo lineal y no lineal

Los patrones son tendencias de información que se repite con alguna regularidad o frecuencia en un intervalo de tiempo y que además poseen ciertas características comunes. Algunos patrones son de tipo lineal, los cuales son los más simples de divisar en un conjunto de datos y se apoyan en las técnicas del álgebra lineal que funcionan de manera muy eficiente en bases de datos de tamaño moderado; no obstante, en bases de datos de tamaño considerable, es decir, de alta dimensión y gran cantidad de registros, dichas técnicas resultan improcedentes al conllevar elevados tiempos de procesamiento. Una solución emergente fue el desarrollo de métodos aproximadas para la detección de patrones lineales, los cuales tienen como fundamento los procedimientos de optimización convexos (*convex optimization*) (Nguyen

Cong, Rivero, & Morell, 2015), que además no generan óptimos locales que afecten a la selección de la solución idónea lejana del óptimo global. La gran ventaja que ofrecen las técnicas de reconocimiento de patrones lineales es que resultan muy intuitivas y cercanas a los sistemas de percepción humana ya que, a partir de ellas, es posible modelar la realidad intrínseca de los datos en base a las relaciones de proporcionalidad entre los mismos las cuales se encuentra profundamente arraigadas en la conceptualización humana y, por tanto, resultan de muy fácil interpretación. Las desventajas aludidas al reconocimiento lineal son: 1) Su sencillez interpretativa está ligada a la rigidez del modelo, 2) los conjuntos de datos no sólo presentan un comportamiento lineal, de hecho, la mayoría de variables y sus relaciones en el mundo real presenta un comportamiento complejo y caótico, por lo que surgieron técnicas para la detección de patrones de naturaleza no lineal.

Por el contrario, en los problemas de detección de patrones no lineales no es posible utilizar las soluciones analíticas por lo que se hace necesario recurrir a procedimientos aproximados. Lamentablemente, dichos procedimientos no gozan de las ventajas que tienen en el reconocimiento lineal, por ejemplo, es común generar óptimos locales lejanos de los óptimos globales, con los problemas que esto puede conllevar a la hora de determinar una solución eficaz. Teniendo en cuanta lo anterior, se hacía necesario el desarrollo de nuevas técnicas con ciertas características particulares como lo son: Eficiencia computacional, que es utilizar adecuadamente los recursos para lograr el objetivo. Robustez, que es la capacidad del sistema para ser relativamente insensible a la presencia de una cierta proporción de datos erróneos sin que esto afecte su comportamiento de forma significativa. Estabilidad, que es la propiedad de un algoritmo que permite detectar los mismos patrones en bases de datos alternativas, obviando las particularidades y evitando el sobreajuste (*overfitting*). Flexibilidad, permite determinar pautas y relaciones más complejas que las proporcionadas por los rígidos modelos lineales, permitiendo develar modelos caóticos e impredecibles, Sensibilidad, que le permite a un algoritmo ser capaz de detectar patrones diferentes en bases de datos diferentes, esto entra en conflicto con la propiedad de robustez y por tanto se hace necesario dotar de mecanismos de regularización que permitan adecuar tanto la generalización como el sobreajuste a los criterios del analista y a los requerimientos de cada conjunto de datos.
Los patrones detectados deben ser congruentes y significativos para el analista, sin embargo si en la búsqueda de dichos patrones no se tiene en cuenta un criterio de referencia de lo que se desea encontrar, los resultados pueden ser muy abstractos y lejanos a la interpretación humana. Una forma de solventar esta situación es dotando a los métodos de una capacidad de interacción con el usuario, de tal forma que lo integre de manera activa al proceso de reconocimiento de patrones, en donde el analista puede guiar el proceso limitando el espectro de patrones que espera encontrar, además de minimizar el riesgo de *overfitting*. Por tal razón es imprescindible contar con el conocimiento de un experto en la disciplina a la que corresponde la base de datos objeto de estudio, puesto que utilizando su criterio se pueden establecer de forma idónea las expectativas de los patrones proclives a ser detectados en su respectiva área de conocimiento, al igual que determinar qué patrones resultarían ilógicos.

De esta manera, es posible adaptar el algoritmo a las condiciones óptimas del contexto de la base de datos analizada.

3. Análisis de componentes principales (PCA), aproximación geométrica

PCA es comúnmente aplicado en la reducción de dimensión de un conjunto de datos que en principio están descritos con un elevado número de variables. El objetivo de este método es encontrar la mejor representación en términos de componentes que brinden la mayor variabilidad de dichos datos en términos de mínimos cuadrados, realizando una proyección que corresponde a la varianza acumulada de cada observación. PCA en realidad no reduce la dimensión, puesto que en esencia las componentes generados deben representar el rango en su totalidad, sin embargo si se seleccionan las primeros componentes resultantes asociados a los mayores valores propios, es posible reducir la dimensión, reteniendo aquellos atributos del conjunto de datos que contribuyen más a su varianza, por tanto las características escogidas son las que presentan mayor separabilidad con respecto a la media de los datos. PCA construye una transformación lineal de los datos originales, generando un nuevo sistema de coordenadas, en donde la mayor varianza del conjunto de datos es capturada en el primer eje (denominado primer componente principal), la segunda varianza más grande en el segundo eje, y así sucesivamente; donde la medida de varianza la define una estimación de la matriz de covarianza de los datos. En síntesis, PCA busca minimizar el error cuadrático medio de la proyección de los datos sobre los vectores propios de la matriz de covarianza, sujeto a una condición de ortonormalidad, para encontrar un subespacio del espacio original en el que se encuentran representados los elementos, de dimensión reducida pero que, simultáneamente, permita una adecuada representación de los datos, con la mínima pérdida de información (Figura 1). Antes de aplicar PCA generalmente primero se normaliza y posteriormente se realiza el escalado de variables para que tenga media cero y rangos de valores comparables.

En términos geométricos, PCA intenta encontrar una superficie de dimensión inferior sobre la que pueda proyectar los datos de modo que la suma de los cuadrados de los segmentos conocida como error de proyección (líneas punteadas en la Figura 1) se minimice.

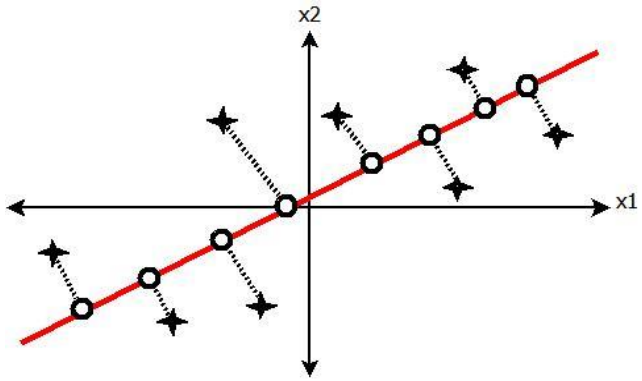


Figura 1: Proyección a un espacio de menor dimensión con el mínimo error.

En este ejemplo, la línea roja (componente 1) en la Figura 1, es la más adecuada para proyectar los datos, puesto que los errores de proyección son mínimos, por el contrario, si se eligiera otra línea de proyección (componente) el error cuadrático aumentaría (Figura 2).

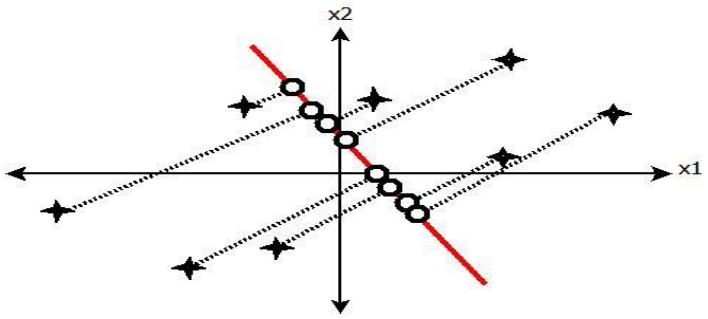


Figura 2: Proyección a un espacio de menor dimensión con alto margen de error.

A través del PCA es posible determinar las componentes de proyección de la distribución de puntos de la matriz de entrada **Y**, el cual tiene p dimensiones y n muestras, cuyo origen se encuentra en $\hat{\mathbf{y}}$, que es el vector medio de $y_1, y_2, \ldots, y_n$. Esto se realiza restando $\hat{\mathbf{y}}$ y calculando la rotación que minimice la suma de distancias a los ejes, o visto de otra manera, que maximice la proyección de los datos sobre los mismos ejes. Es posible rotar los ejes multiplicando cada vector p-dimensional y_i por una matriz ortogonal $\boldsymbol{A}$ asi: $z_i = \boldsymbol{A}\, y_i$. Teniendo en cuenta que $\boldsymbol{A}$ es ortogonal, la distancia al origen no cambia: $\boldsymbol{A}^{\mathrm{T}}\, \boldsymbol{A} = \boldsymbol{I}$, en consecuencia z_i es realmente una rotación cuyo resultado es el producto interno de **Y**: $z_i^{\mathrm{T}} z_i = (\boldsymbol{A} y_i)^{\mathrm{T}} (\boldsymbol{A} y_i) = y_i^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\, y_i = y_i^{\mathrm{T}} y_i$. Por tanto, el objetivo es encontrar una matriz ortogonal $\boldsymbol{A}$ que suministre unos nuevos parámetros Z denominados componentes principales, los cuales no deben estar correlacionados, con la intención de que cada componente agrupe un conjunto de variables similares entre sí pero diferentes a las variables de otros componentes, y por tanto cada componente aporta un tipo de información diferente al análisis. Para ello necesitamos que la matriz maestral de covarianzas de Z, $\boldsymbol{S}_z$, sea diagonal.

$$S_z = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^{\mathrm{T}} = \begin{pmatrix} s_{Z1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{Zp}^2 \end{pmatrix}$$

Donde $\boldsymbol{S}$ es la matriz muestral de covarianzas de la matriz de entrada **Y**, la cual es simétrica y ha sido diagonalizada empleando una matriz ortogonal que contiene los vectores propios normalizados de $\boldsymbol{S}$, y consecuencia, la matriz diagonal resultante contiene los valores propios asociados. Teniendo en cuenta, que los datos han sido centrados, es decir, que el centro de masas coincide con el origen de $\mathbb{R}$, el promedio de los productos entre las puntuaciones de los individuos en una pareja de variables es igual a la covarianza entre dichas variables, por tanto: $(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) = \sum_{i=1}^{l} y_{ij} y_{ij'} = l\, cov(y_{.j'} y_{.j'})$ así $(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) = l\boldsymbol{C}$ siendo $\boldsymbol{C}$ la matriz de covarianzas, entre las variables de la matriz de entrada **Y**. La matriz $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ es de utilidad para calcular la dispersión de las observaciones en una dirección determinada del espacio vectorial $\mathbb{R}$. En síntesis, el problema se reduce a encontrar la matriz ortogonal $\boldsymbol{A}$ tal que diagonalice $\boldsymbol{S}$.

$$\boldsymbol{A} = \begin{pmatrix} a_1^{\mathrm{T}} \\ a_2^{\mathrm{T}} \\ \vdots \\ a_p^{\mathrm{T}} \end{pmatrix}$$

Los a_i elementos son los vectores propios de $\boldsymbol{S}$ que verifican $a_i^{\mathrm{T}} a_j = \delta ij$ (están normalizados y son ortogonales). Las componentes principales son las nuevas variables $Z_{\mathrm{i}} = a_{ii}^{\mathrm{T}} y_{\mathrm{i}}$ tal que, $z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p$. Los valores propios de $\boldsymbol{S}$ serán las varianzas muestrales de las componentes principales, los cuales deben ser ordenados de forma descendente $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Con la intención de establecer los d λ valores propios más significantes, donde d es la nueva dimensión a la que se redujo el sistema.

$$\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_2 \end{pmatrix} = \begin{pmatrix} s_{Z1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{Zp}^2 \end{pmatrix}$$

Como se puede observar, PCA en esencia es un problema de descomposición espectral de la matriz $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ o, lo que es equivalente, de la matriz de covarianzas $\boldsymbol{S}$; en otras palabras, las matrices $\mathbf{Y}^{\mathrm{T}}\,\mathbf{Y}$ y $\boldsymbol{S}$ comparten vectores propios y los valores propios de una se pueden obtener muy fácilmente a partir de los valores propios de la otra.

4. Kernel PCA

Como respuesta a los requerimientos anteriormente mencionados y como solución a los diversos problemas de las técnicas lineales y no lineales ha emergido un nuevo enfoque denominado "Aprendizaje con funciones kernel" (Sánchez, Osorio, & Suárez, 2008) que integra diversos mundos de forma sinérgica y holística, es decir, por un lado toma lo mejor del enfoque lineal y del no lineal y por otro lado, gracias a su característica de modularidad también permite integrar los mundos de la inteligencia artificial y la inteligencia natural como un equipo ideal para descubrir nuevo conocimiento. En efecto, los procedimientos de detección de patrones basados en funciones kernel son, en esencia, métodos de detección de patrones lineales aplicados a un espacio de muy alta dimensión y en consecuencia poseen las ventajas de los algoritmos lineales (sencillez, estabilidad, eficiencia computacional...). Al mismo tiempo, gozan de la flexibilidad de los algoritmos de detección de patrones no lineales gracias un cambio en la representación de los datos. A continuación se describe el procedimiento del Aprendizaje con funciones kernel (Figura 3):

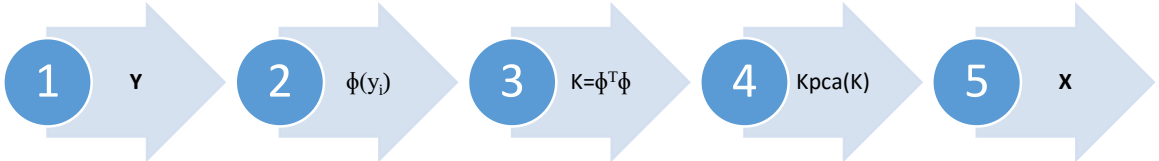


Figura 3: Procedimiento de aprendizaje *kernel* PCA. Se reduce la alta dimensión de los datos de entrada **Y** utilizando métodos espectrales cuyo resultado es la obtención de la matriz **X**, que tiene dimensión inferior.

1) Datos originales: Un conjunto de datos de entrada (*Input Space*) denominado **Y**, $\mathbf{Y} \in \mathbb{R}^{DxN}$ el cual se desea analizar y sobre el que se descubrirán patrones, en donde posiblemente los datos no son separables. No es necesario que este conjunto tenga una estructura algebraica concreta, es decir, puede tener asociado un espacio vectorial o no.

2) Selección: Una función de mapeo que incrusta cada elemento del conjunto de entrada **Y** en un espacio de altísima dimensión **Φ**: $\phi(\cdot)$: $\mathbb{R}^D \rightarrow \mathbb{R}^{Dh}$

$$\mathbf{y}_i \rightarrow \phi(\boldsymbol{y}_i).$$

El objetivo es hacer que los datos sean fácilmente separables (Figura 4). De esta forma se modifica la representación del conjunto de datos original con el fin de facilitar la tarea de los algoritmos de búsqueda de patrones de tipo lineal. Este nuevo espacio debe tener una estructura algebraica de Espacio de Hilbert (Akhiezer & Glazman, 2013) y estar dotado de un producto escalar: $\phi(\boldsymbol{y}_i)^\top \phi(\boldsymbol{y}_j)$, el cual tiene una dimensión muy elevada (incluso infinita), la cual en principio, puede parecer un gran problema en términos de eficiencia computacional, ya que en aparecía tanto el cálculo de los mapeos y proyecciones por ϕ de los elementos de **Y,** como la búsqueda de patrones, así sean lineales, podría resultar muy costoso. La solución a este inconveniente es recurrir al criterio de Mercer, también conocido como truco kernel (*Kernel Trick*) (Leen, Dietterich, & Tresp, 2001) descrito posteriormente.

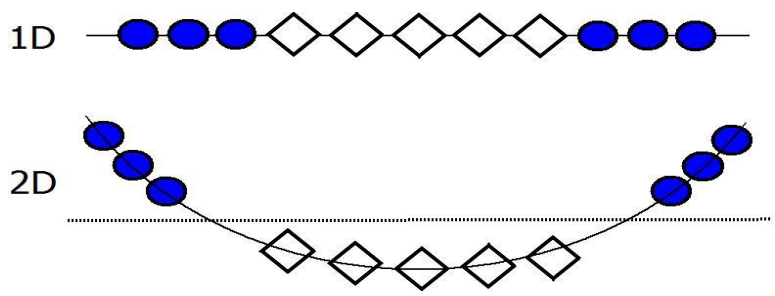


Figura 4: Incrustramiento de 1D a 2d. En primera instancia se tienen los datos dispuesto en una dimensión los cuales son llevados a dos dimensiones en donde son fácilmente separables linealmente (línea punteada).

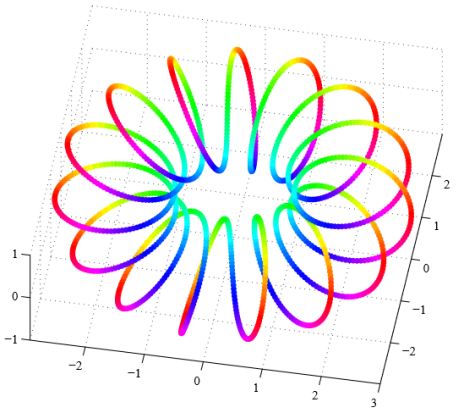
3) Formación de la matriz kernel: Una función kernel k(·,·) definida sobre el producto cartesiano del conjunto de entrada **Y** consigo mismo **K**: $\mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$, que hace corresponder a cada pareja del elementos **Y** el producto escalar en **Φ**, es decir, el mapeo generado por la función $\phi(\cdot)$, tal que: $\phi(\boldsymbol{y}_i)^{\mathrm{T}}\phi(\boldsymbol{y}_j) = k(\boldsymbol{y}_i, \boldsymbol{y}_i)$, organizando todos los productos escalares en una matriz **K** = $[k_{ij}]$, se obtiene la matriz kernel **K** = $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$, con la intención de encontrar las medidas de similitud. En síntesis, la función kernel puede ser entendida como una composición del mapeo generado por $\phi(\cdot)$ y de su producto escalar así: $\phi(\boldsymbol{y}_i)^{\mathrm{T}}\phi(\boldsymbol{y}_j)$, de tal forma que para cada pareja de elementos del conjunto **Y** se asigna directamente su producto escalar sin necesidad de transitar por el mapeo **Φ**. Así, la función kernel, junto con la versión kernel del algoritmo de búsqueda de patrones lineales y el Teorema de Mercer, permiten prescindir por completo tanto de la matriz de altísima

dimensión **Φ** como de la función de mapeo $\phi(\boldsymbol{y}_i)$ que, en muchas ocasiones, ni siquiera se conoce, por tanto, es suficiente con centrar la atención en la función kernel, k(·,·).

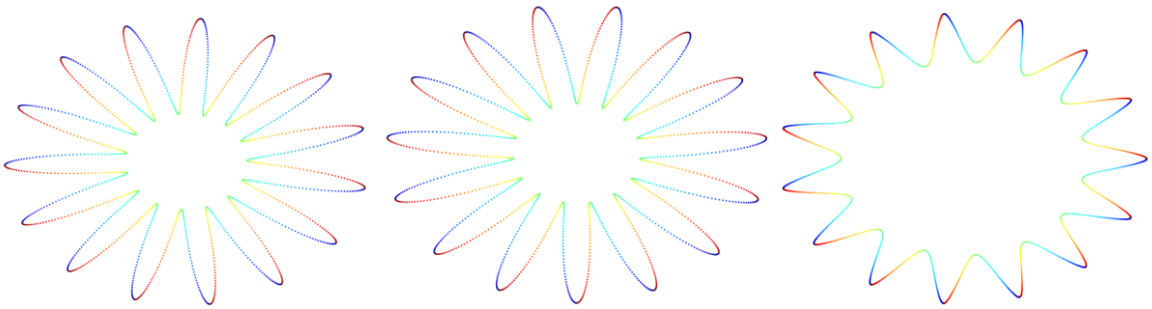
4) Algoritmo de reducción de dimensión: Un algoritmo rediseñado en versión kernel que permite detectar patrones lineales en el espacio de altísima dimensión **Φ**, de tal manera que no necesita disponer de los valores concretos de $\phi(\boldsymbol{y}_i)$ $\forall$ **y** $\in$ **Y** sino tan solo de los productos escalares entre el mapeo de los elementos de **Y**, en otras palabras, $\phi(\boldsymbol{y}_i)^{\mathrm{T}}\phi(\boldsymbol{y}_j)$. Este algoritmo es denominado (KPCA) ya que aplica PCA pero usando un kernel especifico en lugar de la matriz de covarianza. Como se puede observar, la determinación de un patrón lineal puede llevarse a cabo a partir, exclusivamente, de la información recogida en los productos escalares calculados para todas las parejas de elementos del espacio, los cuales esencialmente, recopilan la información sobre las normas de los elementos del espacio y los ángulos existentes entre ellos. Aunque en el proceso se pierde tanto la información sobre las coordenadas reales de los elementos en el espacio como la orientación del conjunto de datos en el espacio o la alineación de los elementos con las variables originales, sin embargo, generalmente dicha perdida no es relevante en la detección de patrones lineales, convirtiendo al conjunto de productos escalares en suficiente insumo para la determinación del patrón lineal.

5) Datos de salida: Finalmente se obtiene una matriz de datos reducidos denominada **X**, **X** $\in$ $\mathbb{R}^{dxN}$, lo cual significa que la matriz resultante es de dimensión inferior a la de los datos de entrada **Y**. En este caso al utilizar la técnica kernel de análisis de componentes principales es posible encontrar la dimensión asociada a las componentes que más información aporten al análisis como se explicó en la sección 3 y, por ende, reducir la dimensión.

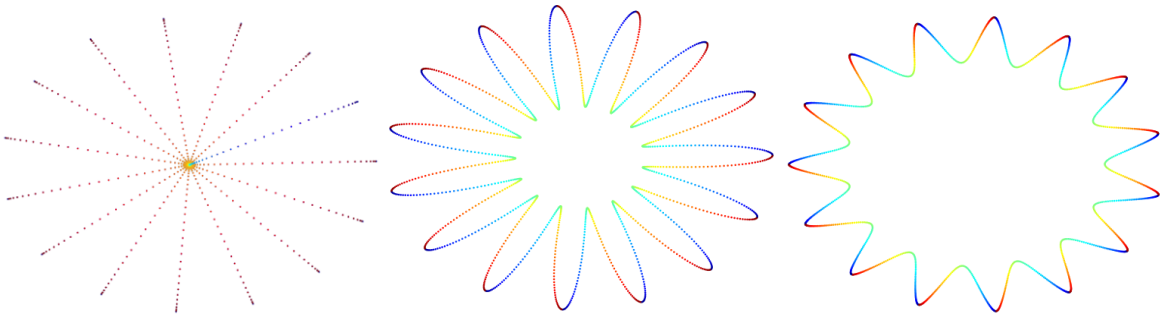
En síntesis, la elección de la función de mapeo $\phi(\cdot)$ hace que los patrones lineales detectados en el espacio de altísima dimensión **Φ** correspondan a patrones potencialmente no lineales en la matriz de entrada **Y**. Así, el algoritmo en realidad está detectando patrones no lineales en el espacio original **Y**. Por tanto la elección de la función de mapeo permite la incorporación de conocimiento experto y por tanto, aporta modularidad y adaptabilidad del algoritmo. Sin embargo y gracias al truco del kernel, la delimitación de patrones se traslada de la función de mapeo a la función kernel y es esta la que en realidad determina el conjunto de patrones que el analista puede esperar detectar en los datos de entrada **Y**, por tanto se convierte en la vía real de incorporación de conocimiento experto.



a) Datos de entrada: Toroide



b) LLE c) CMDS d) LE



f) KLLE g) KCMDS h) KLE

Figura 5: Representación en baja dimensión (2D) de los datos de entrada de un toroide (3D) utilizando métodos de reducción de dimensión originales y métodos kernel

En la Figura 5 se muestran los espacios de baja dimensión resultantes de aplicar algunos métodos de RD sobre un conjunto de datos articules que representa un toroide en 3D. Este conjunto de datos es simple y la tarea de reducción consiste, de algún modo, en aplanar el toroide, es decir, generar una representación plana (2D) de dicho toroide conservando la relación entre puntos vecinos en torno a su naturaleza topológica.

5. Conclusiones

En este documento, se presenta un enfoque de integración entre la inteligencia artificial y la natural basado en reducción de dimensión con métodos kernel, estableciendo las ventajas de aplicar técnicas lineales de reconocimiento de patrones en un espacio mapeado de muy alta dimensión, donde los datos pueden ser linealmente separables y posteriormente proyectando los resultados al conjunto original, identificando de esta manera patrones no lineales a partir de técnicas lineales, integrando de forma holística las ventajas del reconocimiento lineal y no lineal.

Las técnicas de reducción de dimensión pueden ser utilizadas en el proceso KDD (*Knowledge Discovery in Databases*) en diversas etapas. Como técnicas de preprocesamiento (*Data cleaning*) al eliminar ruido e información duplicada y redundante. También pueden ser consideradas en la etapa de procesamiento (*Data Mining*) como una técnica de descubrimiento de conocimiento en sí misma, al correlacionar exhaustivamente las variables y al develar agrupaciones implícitas (*Clusters*). Y finalmente, también es útil en la etapa de visualización, al llevar la representación de los datos a una dimensión comprensible por la percepción humana (2D O 3D). La aplicación de estos métodos también permite la compresión sustancial de dichos datos, preservando la información más relevante de los mismos mediante la consolidación de las componentes principales.

Como trabajo futuro, se propone la exploración de nuevos enfoques de aprendizaje basados en múltiples kernel mediante la combinación lineal parametrizada, enfocada a alcanzar un buen equilibrio entre la preservación de la estructura de datos (topología) y la visualización inteligible de los mismos.

Bibliografía

Akhiezer, N. I., & Glazman, I. M. (2013). *Theory of Linear Operators in Hilbert Space*. Courier Corporation.

Alvarado-Pérez, J. C., & Peluffo-Ordóñez, D. H. (2015). Artificial and Natural Intelligence Integration. En *Distributed Computing and Artificial Intelligence, 12th International Conference* (pp. 167–173). Springer.

Alvarado-Pérez, J. C, J. C., Peluffo-Ordóñez, D. H., & THERÓN, R. (2015). Bridging the gap between human knowledge and machine learning. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *4*(1), 54–64.

Ambika, D., Radhika, K., & Seshachalam, D. (2016). Periocular authentication based on FEM using Laplace–Beltrami eigenvalues. *Pattern Recognition*, *50*, 178–194.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, *15*(6), 1373–1396.

Borg, I. (2005). *Modern multidimensional scaling: Theory and applications*. Springer.

Bunte, K., Haase, S., Biehl, M., & Villmann, T. (2012). Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, *90*, 23–45.

Chen, Z., Ding, S. X., Zhang, K., Li, Z., & Hu, Z. (2016). Canonical correlation analysis-based fault detection methods with application to alumina evaporation process. *Control Engineering Practice*, *46*, 51–58.

Fu, Y. (2014). Kernel Methods and Applications in Bioinformatics. En N. Kasabov (Ed.), *Springer Handbook of Bio-/Neuroinformatics* (pp. 275-285). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-30574-0_18

Hu, C., Sepulcre, J., Johnson, K. A., Fakhri, G. E., Lu, Y. M., & Li, Q. (2016). Matched signal detection on graphs: Theory and application to brain imaging data classification. *NeuroImage*, *125*, 587–600.

Japkowicz, N., & Stefanowski, J. (2016). A Machine Learning Perspective on Big Data Analysis. En *Big Data Analysis: New Algorithms for a New Society* (pp. 1–31). Springer.

Keim, D. A., Mansmann, F., & Thomas, J. (2009). Visual Analytics : Scope and Challenges, *4404*(4404), 76-90.

Lee, J. A., Renard, E., Bernard, G., Dupont, P., & Verleysen, M. (2013). Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*.

Lee, J. A., & Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer Science & Business Media.

Leen, T. K., Dietterich, T. G., & Tresp, V. (2001). *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*. MIT Press.

Liu, Y., Zhang, Y., Yu, Z., & Zeng, M. (2016). Incremental supervised locally linear embedding for machinery fault diagnosis. *Engineering Applications of Artificial Intelligence*, *50*, 60–70.

Nguyen Cong, B., Rivero, J. L., & Morell, C. (2015). Aprendizaje supervisado de funciones de distancia: estado del arte. *Revista Cubana de Ciencias InformÃ!`ticas*, *9*, 14-28.

Peluffo-Ordóñez, D. H., Lee, J. A., & Verleysen, M. (2014). Recent methods for dimensionality reduction: A brief comparative analysis. En *European Symposium on Artificial Neural Networks (ESANN)*. Citeseer.

Pulkkinen, S. (2016). Nonlinear kernel density principal component analysis with application to climate data. *Statistics and Computing*, *26*(1-2), 471–492.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Sánchez, L. G., Osorio, G. A., & Suárez, J. F. (2008). Introducción a kernel ACP y otros métodos espectrales aplicados al aprendizaje no supervisado. *Revista Colombiana de Estadística*, *31*, 19–40.

Sedlmair, M., Munzner, T., & Tory, M. (2013). Empirical guidance on scatterplot and dimension reduction technique choices. *Visualization and Computer Graphics, IEEE Transactions on*, *19*(12), 2634–2643.

Sparks, R., Ickowicz, A., & Lenz, H. J. (2016). An Insight on Big Data Analytics. En *Big Data Analysis: New Algorithms for a New Society* (pp. 33–48). Springer.

Wang, L., & Zhao, C. (2016). Dimensionality Reduction and Compression Technique of HSI. En *Hyperspectral Image Processing* (pp. 257–281). Springer.

Ware, C. (2012). *Information Visualization: Perception for Design*. Elsevier.