IMPLEMENTACIÓN DE UNA INTERFAZ DE VISUALIZACIÓN DE DATOS EFICIENTE E INTERACTIVA A PARTIR DE UNA PERSPECTIVA GEOMÉTRICA

JOSE ALEJANDRO SALAZAR CASTRO

YESID CAMILO ROSAS NARVÁEZ

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
SAN JUAN DE PASTO
2015

IMPLEMENTACIÓN DE UNA INTERFAZ DE VISUALIZACIÓN DE DATOS EFICIENTE E INTERACTIVA A PARTIR DE UNA PERSPECTIVA GEOMÉTRICA

JOSE ALEJANDRO SALAZAR CASTRO

YESID CAMILO ROSAS NARVÁEZ

TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE INGENIERO ELECTRÓNICO

DIRECTOR

PhD. DIEGO HERNÁN PELUFFO ORDÓÑEZ

INGENIERO ELECTRÓNICO

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
SAN JUAN DE PASTO
2015

NOTA DE RESPONSABILIDAD

“La Universidad de Nariño no se hace responsable por las opiniones o resultados obtenidos en el presente trabajo y para su publicación priman las normas sobre el derecho de autor.”

Acuerdo 1. Artículo 324. Octubre 11 de 1966, emanado del honorable Consejo Directivo de la Universidad de Nariño.

NOTA DE ACEPTACIÓN:

_____
_____
_____
_____
_____
_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

San Juan de Pasto, 11 de noviembre de 2015

DEDICATORIA

“A mi madre y a mi hija por ser el apoyo incondicional y la motivación para seguir adelante hasta alcanzar el éxito, a mi profesor, director y amigo Diego Hernán Peluffo Ordóñez por su confianza en mí y por ser el guía en este maravilloso camino del aprendizaje. A mi pareja y a mi familia por su apoyo incondicional. A Dios por darme el privilegio de tener la salud y el entendimiento necesario para salir adelante y lograr mis objetivos”.

JOSE ALEJANDRO SALAZAR CASTRO

“A Dios por regalarme la fortaleza y conocimiento para seguir adelante perseverando esta meta tan anhelada. A mis padres y a mi pareja por ser ese pilar en mi formación académica y personal, ellos me impulsaron a seguir adelante hasta alcanzar el éxito, a mi profesor, director y amigo Diego Hernán Peluffo Ordóñez por su confianza en mí y por ser el guía en este maravilloso camino del aprendizaje”.

YESID CAMILO ROSAS NARVÁEZ

AGRADECIMIENTO

“Agradezco a Dios por darme la vida y la posibilidad de poder estudiar, aprender y aplicar mis conocimientos. A mi madre, por ser el apoyo y una de las razones más grandes para salir adelante y cumplir mis metas. A mi director y amigo, Diego Hernán Peluffo Ordoñez, por depositar su confianza en mí y despertar el espíritu investigativo, por su colaboración en momentos difíciles y por su apoyo incondicional. A mi hija, por ser la razón más grande que tengo para alcanzar mis metas y cumplir mis sueños. A mi pareja, por ser quien en momentos difíciles me escuchaba y ayudaba para continuar con mi formación profesional. A mi familia, por sus palabras de aliento y apoyo incondicional que me brindaron motivándome a seguir adelante y alcanzar mis metas propuestas”.

JOSE ALEJANDRO SALAZAR CASTRO

“Agradezco a Dios por brindarme la posibilidad de poder estudiar y consolidarme como una persona llena de muchos valores, también le agradezco por permitirme aprender y llevar a cabo mis conocimientos, agradezco a mis padres, por ser el apoyo y una de las razones más grandes para salir adelante y cumplir mis metas. A mi director y amigo, Diego Hernán Peluffo Ordoñez, por depositar su confianza en mí y despertar el espíritu investigativo, por su gran colaboración en momentos difíciles y por su apoyo incondicional. A mi pareja, por acompañarme en los momentos más difíciles durante mi carrera, por guiarme por el camino correcto hacia el éxito, el cual se ve reflejado en este trabajo. A mi familia, por sus palabras de aliento y apoyo incondicional que me brindaron motivándome a seguir adelante y alcanzar mis metas propuestas”.

YESID CAMILO ROSAS NARVAEZ

CONTENIDO

Pág.

LISTA DE FIGURAS

LISTA DE ANEXOS

RESUMEN

El uso de métodos de minería de datos y técnicas de visualización de datos por separado conlleva a la necesidad de expertos para la interpretación de resultados, esto es algo negativo porque involucra un incremento en tiempo, costos y trabajo para llegar a la etapa última del procesamiento de datos, la cual es determinar la información útil inmersa en los datos analizados. La integración de los métodos de minería y las técnicas de visualización es una necesidad latente; aunque existen herramientas que realizan dicha integración, el diseño de un sistema de análisis visual que se adapte adecuadamente a las necesidades y requerimientos de un usuario particular es aún un problema abierto. Una de las formas de hacer una visualización inteligible de grandes volúmenes de datos es reducir la dimensión de los mismos de tal forma que se obtengan representaciones fácilmente interpretables por el ser humano.

En esta tesis se propone implementar una interfaz de usuario de fácil manejo que permita una interacción con los resultados de la representación de los datos realizando una mezcla de métodos de reducción de dimensión. Para este fin, se propone un modelo que realice dicha mezcla a través de un enfoque geométrico, mediante el cual un usuario –no necesariamente experto– intuitivamente manipule dinámicamente la representación visual de los datos usando los parámetros de un polígono geométrico.

ABSTRACT

The use of data mining methods and data visualization techniques separately imvolve to the need for experts to interpret the results, this fact is something negative because it increase processing time, computational cost and work for reaching the main goal of the data processing, which is to infer meaningful information about the analyzed data. The integration of data mining methods and information visualization techniques is a latent need, despite having been proposed tolos to perform this task, the design of visual analysis system that adjust properly to the needs and requirements of a particular user is still an open issue. One way to make an intelligible visualization of big data volumes is reducing its dimensión, so that we can obtain easily interpretable representations by the human.

In this thesis, we propose to implement a easy handling user interface that allows for an interaction with the results of data representation by doing a mixture of dimension reduction methods. To this end, we introduce a model able to perform this mixture through a geometric approach, whereby a user –no necessarily expert- can handle intuitively and dynamically the visual representation of the data using the parameters of a geometric shape.

1. INTRODUCCIÓN

El campo de la visualización de la información (*Info Vis*) tiene por objetivo desarrollar formas graficas de representar datos de modo que la información pueda ser más utilizable e inteligible para el usuario. La reducción de dimensión (RD) se convierte en una etapa determinante en el diseño de sistemas de minería de datos (*data mining*) o sistemas de reconocimiento de patrones para el tratamiento de conjuntos de datos de alta dimensión [1]. Para lograr el desarrollo de esta etapa es necesario aplicar métodos RD, cuya finalidad es la extracción de un conjunto de datos en baja dimensión que deriva de la información relevante (llamada información embebida o integrada) de un conjunto de datos de entrada que contienen alta dimensión (ver Figura 1), con la finalidad de mejorar el desempeño de un sistema de minería de datos o de reconocimiento de patrones y a su vez lograr una representación de datos más inteligible [2]. Aunque los métodos RD son frecuentemente desarrollados bajo determinados parámetros de diseño y criterios de optimización preestablecidos, se necesita un mayor desarrollo relacionado a la inclusión de propiedades de procedimientos de la Info Vis, como son la interactividad con el usuario y la controlabilidad del sistema [3]. Según lo anterior, se puede determinar que es factible mejorar la RD importando algunas propiedades de las técnicas de Info Vis [4]. Esto es la premisa en la que se basa este trabajo de investigación.

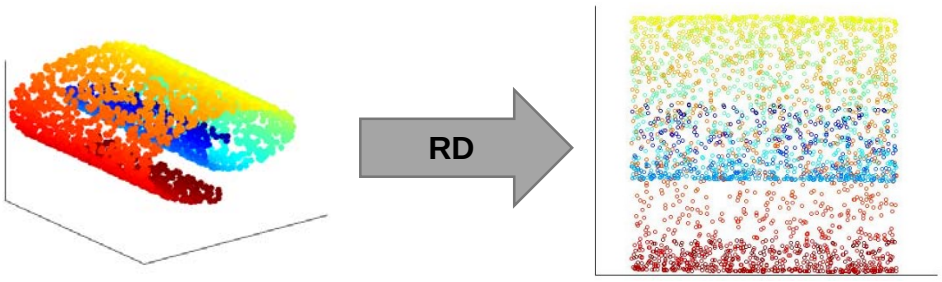


Figura 1. Explicación gráfica de la visualización de los resultados de la reducción de dimensión. En el ejemplo se reduce un rollo suizo 3D a una representación en 2D

Esta tesis presenta una propuesta novedosa para unir el campo de la minería de datos con la visualización de la información a partir de la reducción de dimensión, con el fin de aprovechar las propiedades especiales de la Info Vis dentro de la estructura RD. En particular, las propiedades de la Info Vis que son de interés para este trabajo de investigación son la controlabilidad y la interactividad, las cuales harían que la RD resulte significativamente más entendible y manejable para el usuario especialmente no experto en este campo [5]. Así, estas dos propiedades permiten al usuario tener la libertad de seleccionar la mejor forma de representar los datos mediante la interacción con la máquina. Específicamente, se ha propuesto

una estrategia geométrica para ajustar los factores de ponderación para combinar linealmente los métodos de RD. Esto se realiza a partir de aproximaciones Kernel de métodos convencionales los cuales son combinados para lograr una mescla de los Kernel. El modelo propuesto y los métodos convencionales anteriormente dichos serán explicados con mayor detalle en la sección 4.

1.1. JUSTIFICACIÓN

En la actualidad existe la necesidad de integrar los sistemas de descubrimiento de conocimiento en grandes conjuntos de datos (DCBD) con el fin de rescatar la información más importante que se encuentra inmersa en una serie de variables contenidas en grandes volúmenes de información, de tal forma que sea fácilmente interpretable a través de una visualización clara y coherente de estos resultados del análisis [6]. Las herramientas de DCBD forman parte de un área activa de investigación que se encuentra en constante evolución [7, 8], por tanto, los resultados esperados de este proyecto de trabajo de grado representan un importante aporte en el área de visualización y minería de datos garantizando eficiencia y aplicabilidad. Las diferentes herramientas que hoy en día se han implementado utilizan diferentes técnicas que permiten la clasificación de datos teniendo en cuenta el alcance y las etapas del DCBD. Algunas se enfocan únicamente en la minería de datos teniendo por objetivo la clasificación de información sin tener muy cuenta la representación visual de los datos ni los requerimientos del usuario que debe realizarla [6]. Otras herramientas existentes integran la visualización con los sistemas de minería de datos con el fin de realizar una aplicación completa sin embargo exigen un grado de conocimiento teórico para poder interpretar los resultados finales.

Una técnica de minería de datos visual no es justamente una técnica de visualización, siendo enfocada a explotar los datos en algunas fases de un proceso de minería de datos, pero un algoritmo de minería de datos con análisis visual puede desempeñar un mejor papel [9]. La implementación de una interfaz interactiva basada en la combinación de métodos de reducción de dimensión desarrollada dentro de un algoritmo de análisis visual es una nueva metodología dentro de la minería de datos, y por consiguiente dentro del DCBD, que permitirá a un usuario interactuar con los métodos, sin que necesariamente se tenga conocimiento previo sobre estos, y obtener resultados deseados. Así se obtendría una nueva forma de realizar la visualización de datos a partir de la reducción de dimensión, orientada a la integración interactiva y eficiente de estas dos formas de conocimiento, es decir: Trabajo conjunto entre métodos de minería de datos y análisis visual.

Lo que se busca con el desarrollo de este proyecto es realizar una interfaz completa de débil acoplamiento es decir hay una interacción entre usuario y maquina sin necesidad de que el usuario conozca el algoritmo de funcionamiento, dicha interfaz

debe ser muy intuitiva para el usuario y de fácil manejo, que integre el uso de técnicas de minería de datos a partir de la reducción de dimensión y el análisis visual eficiente de los resultados obtenidos. Con esto se obtendrá un sistema completo en el que los usuarios dispondrán de una herramienta que permitirá acceder a los datos, interactuar con los métodos, controlar parámetros para mezclar los métodos y obtener resultados visuales de una forma eficiente y dinámica. El aspecto más importante que tiene este proyecto es que será una base de trabajo para un proyecto de grado doctoral de la Universidad de Salamanca – España. La interfaz será un prototipo para trabajos futuros encaminados a la implementación de una herramienta completa de DCDB que integrará el análisis visual y los procesos de minería de datos. Además se cuenta con asesorías de expertos nacionales y extranjeros.

1.2. CONTRIBUCIONES DE ESTA TESIS

En ocasiones, la aplicación de métodos de minería de datos y reducción de dimensión arroja resultados que no son fácilmente interpretables por los usuarios, sino que resultan ser abstractos por lo que se debe tener la supervisión de un experto que conozca el funcionamiento de dichos métodos para poder determinar información realmente útil. Con el desarrollo de esta tesis se desarrollara una interfaz de usuario que permita una interacción y control entre el usuario y la máquina con el fin de que el usuario pueda manipular e interpretar fácilmente los resultados. Uno de los factores más importantes de la interfaz propuesta es la controlabilidad que permitirá que los métodos de reducción de dimensión se mezclen de acuerdo a la interacción del usuario con la máquina, quien -aún sin conocer específicamente los métodos que se han aplicado- podrá obtener resultados confiables, involucrando un costo computacional bajo.

Por otra parte, esta tesis representa un aporte en el área de minería de datos en términos de realizar una visualización eficiente permitiendo a un usuario, no experto o sin previo conocimiento de los métodos, obtener resultados visuales más naturales o inteligibles mediante el uso de una interfaz interactiva e intuitiva de fácil manejo que requiera de un costo computacional adecuado y que responda eficientemente a las necesidades planteadas.

1.3. ORGANIZACIÓN DEL DOCUMENTO

Este documento está compuesto por 7 secciones, dichas secciones se distribuyen así: Introducción, objetivos, marco teórico, metodología, resultados y discusión, conclusiones y recomendaciones. El contenido es el siguiente:

- En la sección 2 se presenta los objetivos que se plantearon como logros con el desarrollo de esta investigación.

- En la sección 3 se presenta la revisión bibliográfica y un recorrido conceptual donde se abordan conceptos sobre diferentes temáticas fundamentales para el desarrollo de esta tesis tales como: minería de datos, reducción de dimensión, visualización de datos y reconocimiento de patrones e inteligencia artificial.

- En la sección 4 se describe las metodologías diseñadas, adicionalmente se describen las bases de datos empleadas. Los resultados de los experimentos se discuten en la sección 5.

- Finalmente en la sección 6 se presenta las conclusiones de esta investigación, además se menciona el posible trabajo a futuro.

2. OBJETIVOS

En esta sección se plantea los objetivos esperados con el desarrollo de esta tesis.

2.1. OBJETIVO GENERAL

Desarrollar e implementar una interfaz de usuario para el análisis visual de bases de datos a través de un enfoque matemático-geométrico y programación secuencial que permita la visualización eficiente e interactiva de grandes volúmenes de información.

2.2. OBJETIVOS ESPECÍFICOS

- Diseñar un modelo matemático-geométrico para realizar la combinación de métodos de reducción de dimensión, orientado a la visualización interactiva de datos.

- Desarrollar un algoritmo de programación secuencial para realizar eficientemente un análisis visual de grandes volúmenes de datos.

- Implementar una interfaz de usuario interactiva que permita realizar la puesta en funcionamiento del modelo y algoritmo desarrollados.

3. MARCO TEÓRICO

El gran poder de procesamiento de las máquinas y su bajo costo de almacenamiento ha permitido un gran crecimiento en las capacidades de generar y agrupar datos en cantidades enormes, estos grupos de datos contienen una gran cantidad de información oculta la cual es sumamente importante [9, 10], el Descubrimiento de Conocimiento en Bases de Datos (DCBD o KDD -por sus siglas en inglés-) es básicamente un proceso automático en el que partiendo de los datos se combinan el descubrimiento y el análisis, este proceso conlleva a extraer patrones en forma de reglas o funciones, con el fin de que el usuario realice el respectivo análisis, el DCBD se lleva a cabo a partir de tres etapas las cuales son: El preprocesamiento, la realización de la minería de datos (*data mining*) y la presentación de resultados que se realiza mediante la visualización [9, 10, 11].

Las técnicas comunes de tratamiento de datos no permiten recuperar esta información oculta en su totalidad o sencillamente no tienen la capacidad de tratarlos, como dicha información es de gran importancia estratégica es necesario aplicar técnicas de recuperación de la información como la minería de datos [9, 10]. Hoy en día, el crecimiento de datos ha generado una alta demanda en el desarrollo de procesos que permitan entender estos volúmenes de información, esto se realiza eficazmente mediante la minería de datos, pero grandes volúmenes de datos pueden generar similares conjuntos de reglas o patrones. Estas formas de representación del conocimiento requieren de analistas con habilidades en la interpretación de patrones y reglas para extraer verdaderamente el conocimiento subyacente [6]. Lo anterior, es una de las razón por las que surgen las técnicas de reducción de dimensión las cuales permiten mitigar en cierta forma el problema de la dimensión de estos resultados, permitiendo reducir por ejemplo de 5000 variables a tan solo 5 o 4, pero aun así, tales variables pueden ser abstractas, por lo que estas técnicas también necesitan de un experto para su interpretación [12]. En la actualidad, se han desarrollado herramientas de visualización y exploración inteligente de datos que permiten comprender de una mejor forma la gran cantidad de reglas y parámetros obtenidos de la aplicación de la minería de datos, mientras se interactúa con múltiples presentaciones visuales de la información [13].

La información visual representa un papel muy importante en la minería de datos, ya que el objetivo de esta área es lograr descubrir conocimiento inmerso en datos, tal conocimiento no se puede determinar sino por métodos de minería, pero si este conocimiento no es fácilmente interpretable, se aumenta la inversión de tiempo, dinero y entendimiento (que supone la presencia de un experto en el tema). En la actualidad existen herramientas que, en general, implican etapas de pre procesamiento, uso de métodos de minería de datos, post procesamiento y/o la visualización. Sin embargo no todas las herramientas integran todas las etapas mencionadas, terminando en resultados abstractos de la información. Asimismo, las herramientas que integran todas las etapas no tienen especial énfasis en la

visualización, por lo que los resultados, a pesar de que involucran un análisis visual, tienden también a ser abstractos [6, 14]. Mediante el uso de métodos reducción de dimensión se puede transformar los datos en representaciones visuales de objetos en 1D, 2D o 3D que son más inteligibles para el ser humano.

3.1. MINERÍA DE DATOS

La minería de datos se define como un proceso de descubrimiento de patrones, tendencias y significativas relaciones al examinar grandes volúmenes de datos para determinar información inmersa (también conocida como información oculta) en tales datos. El análisis de los datos ha cambiado debido al uso generalizado de herramientas informáticas, por tanto nace la necesidad de la utilización de técnicas especializadas, que se encuentran contenidas en la minería de datos. Dichas técnicas extraen automáticamente el descubrimiento del conocimiento (DC) contenido en la información, almacenándolo de modo ordenado en grandes bases de datos (BD). Las técnicas de minería de datos son el resultado de un largo proceso de investigación. Esta evolución comenzó cuando grandes volúmenes de información fueron almacenados por primera vez en computadoras, dando paso a mejoras en el acceso a los datos, e impulsando el desarrollo de tecnologías que permitan a los usuarios navegar a través de los datos en tiempo real.

La minería de datos toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, ya que esta da origen a una información prospectiva y proactiva. Los componentes esenciales de la tecnología en minería de datos han tenido una evolución durante décadas, en áreas de investigación como; estadísticas, inteligencia artificial y aprendizaje de máquinas. Estas técnicas tienen como finalidad el descubrir patrones, perfiles y tendencias de interés a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos [15] según el requerimiento de usuarios. En la actualidad, la madurez de estas técnicas hizo que estas tecnologías fueran prácticas para los entornos de base de datos actuales [10].

3.1.1. Big data

Dentro de este marco el *big data* hace referencia a todo aquello que tiene que ver con grandes volúmenes de información como se puede apreciar en la Figura 2. Estos volúmenes son analizados a alta Velocidad, y pueden presentar una compleja variabilidad en cuanto a la estructura de su composición, los conceptos fundamentales agrupados que han definido este nombre han sido: volumen, visualización, variabilidad y velocidad. También es importante comprender que además de los datos estructurados, existen los no estructurados que son aquellos

que provienen de fuentes de información conocidas como: la Web, cámaras de móviles, videos o imágenes, redes sociales, sensores de las ciudades y edificios, entre otros. La variedad de su origen y la rapidez con la que se incrementa su volumen, son algunos de los factores que habían dificultado su análisis hasta el momento [16].

Figura 2. Fuentes de información de Big data
Fuente: http://blog.nimbeo.com/index.php/tag/big-data/

El *big data* también está relacionado con la minería de datos, debido a que esta última que intenta descubrir patrones en grandes volúmenes de datos. Tanto la minería de datos como el *big data* parten de *Machine Learning*, el cual también utiliza los métodos de la Inteligencia Artificial y la Estadística para analizar los patrones en las bases de datos con las que trabaja.

3.1.2. Minería de datos y reducción de dimensión.

En el *data mining* se automatizan los procesos cuya finalidad es la de encontrar información predecible en grandes bases de datos, esto como parte de un proceso de predicción automatizada, así las herramientas de *data mining* analizan las bases de datos para identificar modelos ocultos o inmersos en los datos con tan solo realizar un paso, a este paso se le denomina descubrimiento automatizado de modelos previamente desconocidos [10].

Según [10], los pasos que se deben seguir cuando se busca realizar un proyecto de minería de datos como el de la Figura 3, no dependen de la técnica de extracción de conocimiento que se utilice ya que estos siempre serán los mismos. Por consiguiente se dice que las fases del proceso de minería de datos son:

- **Filtrado de datos:** Por lo general, los datos contenidos en la fuente de datos no suelen ser idóneos, y muchas veces no es posible utilizar algoritmos sobre dichos datos. Por tal motivo, mediante un preprocesamiento se filtran los datos y se obtienen muestras de los mismos, o se reducen el número de valores posibles.

- **Selección de variables:** A pesar de que los datos ya se han filtrado, en la mayoría de los casos se obtienen datos con alta dimensión o con un gran número de variables, es por eso que se debe realizar una selección de características para reducir dicho tamaño eligiendo las variables más influyentes en el problema, pero esto se debe realizar teniendo en cuenta la calidad del modelo de conocimiento que resulta del proceso.

- **Extracción de conocimiento:** Son algoritmos que se utilizan para generar modelos de conocimiento que representan patrones de comportamientos que se pueden observar en las variables del problema o en relaciones de asociación entre tales variables.

- **Interpretación y evaluación:** Ya con el modelo resultante, se debe realizar una validación en la que se compruebe la veracidad de las conclusiones que resultan y que son lo suficientemente satisfactorias.

Con las técnicas de minería de datos obtenemos modelos de conocimiento que representan patrones de comportamiento que permiten arrojar conclusiones, pero si ninguno de los modelos alcanza los resultados que espera el usuario, se debe modificar alguno de los pasos anteriores con la finalidad de producir nuevos modelos que aporten conclusiones según los requerimientos del usuario.

Figura 3. Fases a seguir para el proceso de minería de datos

Una etapa que se utiliza con frecuencia para el preprocesamiento de datos es la reducción de dimensión o RD, la cual toma un subconjunto de variables de forma que el espacio de características sea reducido de manera óptima según criterios de valuación cuya finalidad será diferenciar el subconjunto que permita representar de la mejor manera el espacio inicial de entrenamiento. Las características que son incluidas dentro del análisis pueden incrementar costos y tiempo en el proceso que se desarrolla en los sistemas, por lo tanto es necesario invertir tiempo, dinero y

esfuerzos para poder diseñar e implementar sistemas que trabajen con pequeños conjuntos de características. En cuanto a la selección de variables, también existe una necesidad relacionada con la inclusión de un conjunto que contenga las características suficientes para lograr un alto rendimiento, estas dos necesidades se han convertido en los pilares para el desarrollo de diversas técnicas cuyo objetivo es encontrar el subconjunto óptimo partiendo de características iniciales de análisis [17]. Es por esto que la RD significa un gran acople en la minería de datos a partir de la Info Vis con la finalidad de realizar las fases anteriores de una forma más eficiente en relación a fases-modelos-resultados se refiere.

3.2. VISUALIZACIÓN DE DATOS

Entiéndase a la visualización como una comunicación entre un usuario y una computadora, en donde el factor más relevante es la utilidad de la información por sobre cómo se diseña y se presenta. Si se analiza la visualización dentro de la computación se aprecia que está fuertemente relacionada con las interfaces de usuario, los gráficos y la minería de datos, entre otras. La percepción en las personas juega un papel fundamental en el área de la visualización, ya que la comprensión por parte del usuario puede mejorar notablemente teniendo en cuenta la calidad y cantidad de la información mostrada [17].

Dentro de la visualización de datos se hace uso de interfaces interactivas cuyo propósito principal es representar con mínima entropía visual una serie de datos a un usuario final. La visualización de la información debe cumplir con las siguientes características: ser interrelacional, transformar datos abstractos en información relevante, buscar la mínima pérdida de información en dicha transformación, y dirigirse intuitivamente a los usuarios que interactúan, transforman e interpretan esta información [18].

Teniendo en cuenta [11] y [19] las técnicas de visualización de datos se clasifican y se denominan así:

- **Proyección geométrica:**

Esta técnica provee soporte para las actividades en donde los usuarios necesitan encontrar proyecciones informativas de un conjunto de datos multidimensionales. La proyección geométrica incluye técnicas de estadística exploratoria que frecuentemente son usadas para el procesamiento de los datos, los componentes principales son el factor de análisis, y la escala multidimensional, así como los tradicionales gráficos de escala [20] en donde cada dos atributos son proyectados a lo largo de los ejes X y Y en el sistema de coordenadas cartesianas [18].

- **Coordenadas paralelas**

En esta técnica k-dimensiones de datos o espacios de objeto son mapeados sobre una vista bidimensional dibujando k espaciadamente en ejes paralelos a uno de los ejes visualizados. Cada eje es asociado con cada propiedad de los datos y es escalado linealmente asociado con el correspondiente rango de atributos de datos, que pueden o no ser normalizados. Cada ítem de datos es presentado con una línea poligonal que intersecta cada eje en el punto correspondiente a los ítems asociados al valor del atributo del dato. Esta técnica es efectiva para revelar gran cantidad de características de los datos, como sus diferentes distribuciones y las dependencias funcionales. Una limitación de esta técnica posee una limitación con un conjunto de datos pequeño ya que no se obtiene una buena representación, por tanto puede llevar a errar en la interpretación de las visualizaciones [21, 22, 23].

- **Visualización coordinada radial**

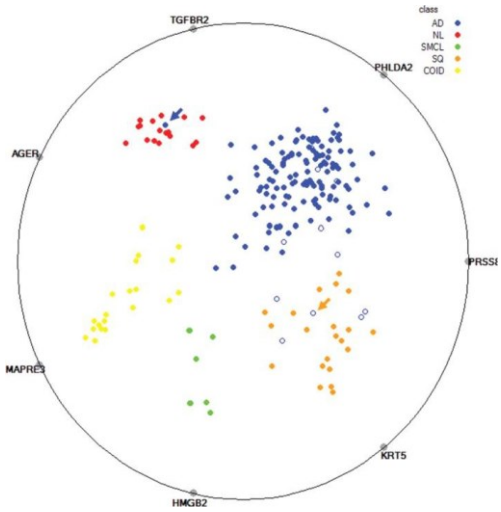
Esta técnica se basa en geometría, para la visualización de n dimensiones, n líneas se despliegan radialmente desde el centro de un círculo hasta su perímetro, y donde cada línea está asociada a un atributo de los datos de alta dimensión [24]. Como se muestra en la Figura 4a.

- **Basada en pixeles**

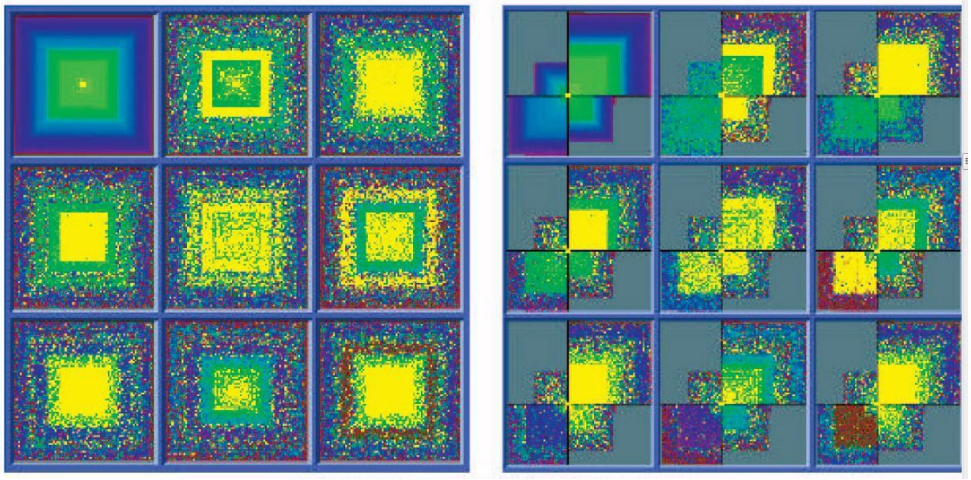
La técnica basada en pixeles es conveniente para grandes conjuntos de datos multidimensionales ya que es categorizada como “independiente de consulta” o “consulta dependiente”. Teniendo en cuenta la consulta independiente, los arreglos de los pixeles en las subventanas están fijos, independientemente de los valores de los datos. En la consulta dependiente, los ítems son establecidos con antelación y son computados usando alguna métrica. De tal forma en que el trazado de los colores de los pixeles se basa en el cómputo de las distancias para cada atributo, y los pixeles en cada subventana se distribuyen según las distancias totales al ítem de datos de la consulta [25]. En la Figura 4b se puede apreciar la visualización usando dos de los posibles arreglos de pixeles, denominados, espiral y arreglo de ejes (izquierda y derecha, respectivamente), producido desde el conjunto de datos ficticios con distribución normal y cinco clúster. El conjunto de datos tiene 7000 registros con ocho atributos.

- **Jerárquica**

Esta técnica es conocida como “mundos dentro de otros mundos”, ya que subdivide las k-dimensiones del espacio de datos y las organiza en subespacios de una forma jerárquica [26]. La técnica puede mapear la tabla de datos, definida en un espacio no jerárquico de k-dimensiones, sobre una perspectiva jerárquica en el espacio 2D.



(a) Visualización de conjunto de datos mediante una visualización coordinada radial. Fuente: [29].



(b) Visualización de conjunto de datos por medio de la técnica basada en pixeles. Fuente: [25]

Figura 4. Clasificación de técnicas de visualización

- **Basada en gráficos**

Esta técnica permite visualizar gráficos de alta dimensión utilizando algoritmos específicos de capas, lenguajes de consulta, y técnicas de abstracción [27], con el fin de transmitir un significado claro y con respuesta rápida.

- **Basada en iconos**

Esta técnica de visualización hace un mapeo a cada ítem de datos multidimensional donde cada rasgo visual varía según los valores de los datos [28]. Las propiedades de los datos son mapeados en una posición 2D de la cara de un ícono sobre la representación y el resto de propiedades son mapeadas sobre las caras del ícono dando formas como nariz, boca, ojos, y las formas propias del rostro. Cabe resaltar que algunos rasgos son usualmente más representativos que otros para el ojo humano, por ejemplo, la gente usualmente pone más atención a los ojos que a las orejas.

- **Reducción de dimensión como técnica de visualización**

Una forma intuitiva de visualizar datos es mediante gráficos 2D o 3D lo que resulta en una visualización natural e inteligible para los seres humanos, por consiguiente, esto significa que los datos inicialmente de alta dimensión deberían ser representados dentro de un espacio de baja dimensión. En este sentido, la reducción de dimensión toma lugar, siendo una etapa importante tanto para el reconocimiento de patrones como para los sistemas de visualización de datos así como también en el diseño de sistemas que realizan el proceso de minería de datos, tal como se vio en la sección 3. De esta manera, se obtiene una visualización de datos más realística e inteligible para el usuario [2].

En general, diremos que la reducción de dimensión (RD) tiene por objetivo alcanzar una representación de datos dentro de una baja dimensión, sobre lo cual el desempeño de las tareas de clasificación son mejoradas en términos de exactitud así como también la naturaleza intrínseca de los datos es correctamente representada (ver Figura 1) [1]. En la reducción de dimensión se tiene una función (o mapeo) de reducción de variables con lo que se da una reconstrucción (o mapeo) suave, el cual es no singular y debe contener de forma aproximada toda la información del espacio que se busca analizar [17]. En otras palabras, la finalidad de la RD es la de deformar una matriz $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$ de alta dimensión, tal que $\boldsymbol{y}_i \in \mathbb{R}^D$, dentro de una matriz $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq i\leq N}$ de baja dimensión, siendo $\boldsymbol{x}_i \in \mathbb{R}^d$, donde *d<D*.

3.3. RECONOCIMIENTO DE PATRONES

El reconocimiento de patrones se define como un mecanismo para distinguir unas cosas de otras, relacionar cosas similares, formar grupos de cosas, describir objetos, tomar y explicar decisiones, etc.

Con el fin de automatizar este proceso es necesario resolver problemas que involucran cuestiones de medición, validación, procesamiento y en finalmente interpretación de la información teniendo en cuenta el entorno de estudio, por tanto,

el reconocimiento de patrones se convierte en un campo de estudio multi-disciplinario. El reconocimiento de patrones se ocupa de los procesos sobre ingeniería, computación y matemáticas relacionados con objetos físicos y/o abstractos [30, 31].

3.4. INTELIGENCIA ARTIFICIAL

En la actualidad ha surgido un campo innovador denominado inteligencia artificial, el cual hace uso de las capacidades de procesamiento computacional de las maquinas modernas para generar conocimiento de forma más eficaz, existe la necesidad de integrar sinérgicamente estos métodos sofisticados enfocados al análisis de datos con los conocimientos, habilidades y cualidades holísticas, flexibles y paralelas de la razón humana. El propósito es encontrar tendencias y patrones ocultos, que forman la base de modelos predictivos que permiten a los analistas producir nuevas observaciones y consideraciones a partir de los datos existentes mediante un proceso como el de la Figura 5.

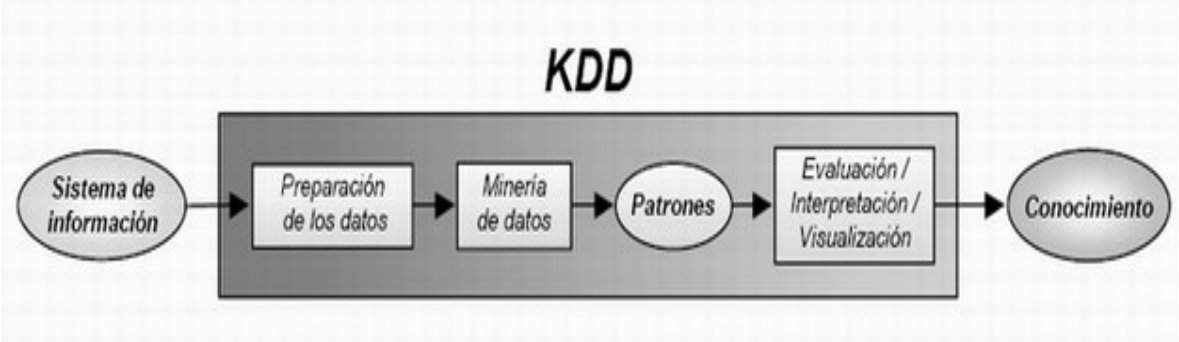


Figura 5. Diagrama del descubrimiento de conocimiento en bases de datos
Fuente: https://sites.google.com/site/todosobrebasededatos/mineria-de-datos

El aprendizaje autónomo hace uso del procesamiento computacional y la analítica visual, que a su vez recurre al pensamiento paralelo del ser humano, es decir la integración de la inteligencia artificial con la inteligencia natural como un equipo idóneo para extraer conocimiento grandes volúmenes de información. A ésta integración se la ha denominado KDD (*Knowledge Dicoverey in databases*), aunque generalmente el termino ha sido utilizado para referir solo las técnicas de minería de datos, aun cuando el concepto es mucho más amplio y abarca otras áreas informáticas.

4. METODOLOGÍA

Come se dijo antes, existe una necesidad latente por acoplar de alguna la minería de datos enfocada a la reducción de dimensión, con las propiedades características de la Info Vis. Si bien, lo métodos RD se desarrollan a partir de parámetros y criterios preestablecidos, se necesita proveer un sistema que permita al usuario una mayor interactividad y controlabilidad, para tal fin, hemos propuesto el desarrollo de una interfaz basada en un modelo matemático geométrico con el cual se trabajan los métodos de reducción de dimensión. Dichos métodos son mezclados a partir de la selección de un punto sobre la superficie de un polígono geométrico generado a partir del número de métodos, así, se obtendrán resultados acordes a las necesidades del usuario (no necesariamente experto) de forma más interactiva y permitiéndole controlar la mezcla de los métodos siendo los resultados obtenidos sean más realístico e inteligibles.

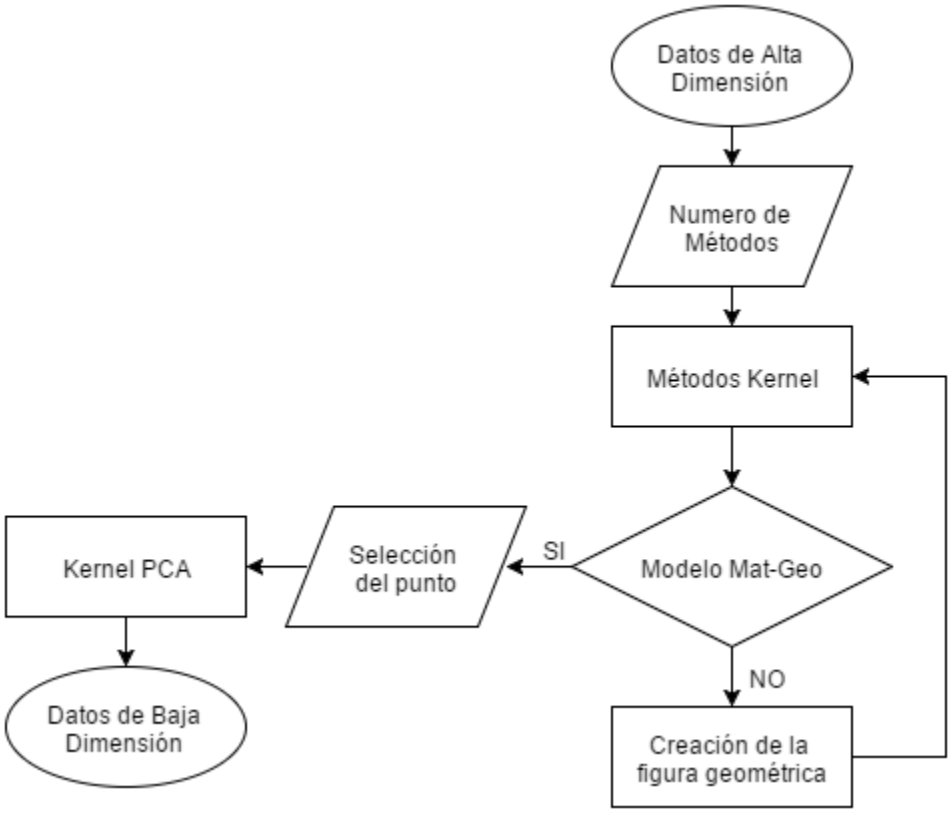


Figura 6. Diagrama de flujo del proceso de reducción de dimensión a partir de un modelo geométrico

En la Figura 6 se puede apreciar el planteamiento del proceso de funcionamiento de la interfaz. Esta consta de diferentes etapas las cuales serán descritas en las

subsecciones siguientes. Adicionalmente, la calidad de los datos obtenidos es cuantificada mediante una versión escalada de la tasa promedio de concordancia entre vecinos [32].

4.1. REDUCCIÓN DE DIMENSIÓN BASADA EN KERNEL

Para el desarrollo de esta tesis se consideran tres aproximaciones del Kernel para los métodos de DR espectrales [6], denominadas así; *Classical Multidimensional Scalling* (CMDS), *locally linear embedding* (LLE), and *graph Laplacian Eigenmaps* (LE). Kernel CMDS es la matriz de distancia doblemente centrada $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ así:

$$\boldsymbol{K}^{(1)} = \boldsymbol{K}_{\mathrm{CMDS}} = -\frac{1}{2}\left(\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N{}^T\right)\boldsymbol{D}\left(\boldsymbol{I}_N - \mathbf{1}_N\mathbf{1}_N{}^T\right), \tag{1}$$

donde la entrada *ij* de *D* será igual a $d_{ij} = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2$ y $\|\cdot\|^2$ teniendo en cuenta la norma euclidiana. El Kernel LLE se puede aproximar a partir de una forma cuadrática en términos de la matriz $\boldsymbol{W}$ que sostiene coeficientes lineales que suman 1 y de manera óptima se reconstruye los datos observados. Ahora se define una matriz $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ como $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{W})(\boldsymbol{I}_N - \boldsymbol{W}^T)$ y λ_{max} como el mayor valor propio de $\boldsymbol{M}$. La matriz Kernel para LLE es de la forma

$$\boldsymbol{K}^{(2)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \tag{2}$$

Como Kernel PCA es un problema de maximización de covarianza de datos de alta dimensión y representado por un Kernel, LE se puede expresar como la gráfica pseudo-inversa de Laplaciana $\boldsymbol{L}$ según

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LE} = \boldsymbol{L}^T, \tag{3}$$

donde $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, tal que $\boldsymbol{S}$ es una matriz de similitud y $\boldsymbol{D} = \mathrm{Diag}\,(\boldsymbol{S}\mathbf{1}_N)$ es el grado de la matriz. Todos los Kernel mencionados anteriormente son explicados ampliamente en [33]. La matriz de similitud $\boldsymbol{S}$ está formada de tal manera que el parámetro de ancho de banda relativo se estima manteniendo la entropía de la distribución con el vecino próximo con más o menos *log(K)* donde K es el número dado de vecinos como se explica en [34]. El número de vecinos se establece como K = 1% del tamaño de los datos.

Si bien, un Kernel RBF es también considerado: $\boldsymbol{K}^{(4)} = \boldsymbol{K}_{RBF}$ cuya *ij* entradas están dadas por $exp(-0.5||yi - yj\,||/\sigma^2)$ con σ = 0.1. Para todos los métodos, los datos de entrada se incrusta en un espacio de 2 dimensiones (*d* = 2).

En consecuencia, este enfoque se realiza considerando *M* = 4 Kernel. El Kernel resultante proporcionado aquí $\widehat{\boldsymbol{K}}$, así como los Kernels individuales

$\{\boldsymbol{K}^{(1)},\ldots,\boldsymbol{K}^{(M)}\}$ se prueban mediante la obtención de datos embebidos en Kernel PCA.

4.1.1. Análisis de componentes principales

Sea la matriz $\boldsymbol{Y}$ siendo $\boldsymbol{y}_i \in \mathbb{R}^D$, y conformada por D variables tales que $\boldsymbol{y}^{(l)} \in \mathbb{R}^N$ con $l \in \{1,\ldots, D\}$, y la matriz $\boldsymbol{X}$ siendo $\boldsymbol{x}_i \in \mathbb{R}^d$ compuesta por d variables denotada como $\boldsymbol{x}^{(\ell)} \in \mathbb{R}^N$ con $\ell \in \{1,\ldots, d\}$. Asumiendo un gran espacio de representación tridimensional $\boldsymbol{\Phi} \in \mathbb{R}^{\boldsymbol{D}_h \times N}$ de tal manera que $\boldsymbol{D}_h \gg D$, en el cual se calcula el producto interno mejorando la representación y visualización de los datos resultantes en comparación con la visualización que se obtiene de la observación directa de los datos. Por lo tanto, surge la necesidad de una representación Kernel con el fin de calcular el producto escalar en el espacio de alta dimensión desconocido. Sea $\Phi(\cdot)$ una función que mapea datos de una dimensión original a otra de una dimensión más alta, de manera que

$$\begin{aligned}\Phi(\cdot):\mathbb{R}^D &\rightarrow \mathbb{R}^{\boldsymbol{D}_h}\\ \boldsymbol{y}_i &\mapsto \Phi(\boldsymbol{y}_i).\end{aligned}$$

Por lo tanto, el *i*-ésimo vector columna de la matriz $\boldsymbol{\Phi}$ está dado por $\boldsymbol{\Phi}_i = \Phi(\boldsymbol{y}_i)$. Teniendo en cuenta la condición de Mercer o Kernel truncado, una función Kernel $k(\cdot,\cdot)$ permite estimar el producto escalar $\Phi(\boldsymbol{y}_i)^{\mathrm{T}}\Phi(\boldsymbol{y}_j) = k(\Phi(\boldsymbol{y}_i), \Phi(\boldsymbol{y}_j))$. Organizando todos los posibles productos punto en un matriz $\mathbf{K} = [k_{ij}]$, se obtiene la matriz Kernel

$$\boldsymbol{K} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}, \tag{4}$$

donde, $k_{ij} = k(\boldsymbol{y}_i, \boldsymbol{y}_j)$.

La formulación de Kernel PCA se lleva acabo centrándose en $\boldsymbol{\Phi}$. Esta condición se puede comprobar con la modificación algebraica del cálculo del producto escalar como se explicara adicionalmente. Para proyectar los datos, se utiliza una combinación lineal con una base d-dimensional. Un tipo de esta base puede estar dispuesta en una matriz de rotación ortonormal $\mathbf{W} \in \mathbb{R}^{\boldsymbol{D}_h \times N}$, de manera que $\mathbf{W} = [\mathrm{w}^{(1)},\ldots, \mathrm{w}^{(d)}]$ y $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}_d$, donde $\mathbf{w}^{(\ell)} \in \mathbb{R}^{\boldsymbol{D}_h}$ y la matriz $\mathbf{I}_d$ es una matriz identidad d-dimensional. Entonces la matriz de datos proyectados $\mathbf{X} \in \mathbb{R}^{d \times N}$ se puede calcular así:

$$\mathbf{X} = \mathbf{W}^{\mathrm{T}}\,\boldsymbol{\Phi}. \tag{5}$$

Generalmente, la proyección se realiza en un espacio de menor dimensión, lo que significa que los datos se proyectan con una representación de bajo rango de matriz de rotación $(d < D)$. No obstante, los datos pueden ser totalmente proyectadas

utilizando una configuración de base $d = D$. Además, a partir de la ecuación 5 una matriz de datos de menor rango $\widehat{\mathbf{\Phi}} \in \mathbb{R}^{D_h \times N}$ puede ser obtenida cuando $d < D$ por

$$\widehat{\mathbf{\Phi}} = \mathbf{WX}. \tag{6}$$

Entonces, se puede escribir $\widehat{\mathbf{\Phi}} = \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}$. El criterio de varianza se puede expresar como $E_{\mathbf{\Phi}}\{\|\mathbf{\Phi}_i - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}_i\|_2^2\}$ donde $\|\cdot\|_2$ y $E_{\mathbf{\Phi}}\{\cdot\}$ denotan la norma euclidiana, y el operador del valor esperado en relación a $\mathbf{\Phi}$, respectivamente. Suponiendo $E_{\mathbf{\Phi}}$como el promedio simple el objetivo de la función basada error cuadrático medio se puede escribir como

$$\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\|\mathbf{\Phi}_{\mathrm{i}} - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}_{\mathbf{i}}\|_2^2 = \frac{1}{\mathrm{N}}\left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_{\mathrm{F}}^2, \tag{7}$$

donde $\|\cdot\|_F$ es equivalente a la norma de Frobenius. Después se explica cómo resolver el problema de optimización y calcular el espacio integrado o embebido.

Una solución factible y eficiente es

$$\begin{gathered}\min\left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_{\mathrm{F}}^2 \\ \text{tal que } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}_{\mathrm{d}}, \mathrm{d} < \mathrm{D} \\ \mathbf{X} = \mathbf{W}^{\mathrm{T}}\mathbf{\Phi},\end{gathered}$$

es seleccionar $\mathbf{W}$ y $\mathbf{X}$ como los vectores propios asociados a los *d* mayores valores propios de $\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ y la matriz Kernel $\boldsymbol{K} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$, respectivamente.

El objetivo de la función puede expresarse como

$$\left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}) - 2\mathrm{tr}\left(\widehat{\mathbf{\Phi}}^{\mathrm{T}}\mathbf{\Phi}\right) + \mathrm{tr}\left(\widehat{\mathbf{\Phi}}^{\mathrm{T}}\widehat{\mathbf{\Phi}}\right). \tag{8}$$

Dado que $\mathrm{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}) = \left\|\widehat{\boldsymbol{\Phi}}\right\|_F^2$ es una constante y $\mathrm{tr}\left(\widehat{\mathbf{\Phi}}^{\mathrm{T}}\mathbf{\Phi}\right) = \mathrm{tr}\left(\widehat{\mathbf{\Phi}}^{\mathrm{T}}\widehat{\mathbf{\Phi}}\right)$, la siguiente dualidad tiene lugar:

$$\left\|\widehat{\mathbf{\Phi}}\right\|_F^2 = \mathrm{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}) + \left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2,$$

donde el problema de minimización $\left\|\mathbf{\Phi} - \widehat{\mathbf{\Phi}}\right\|_F^2$ se puede expresar como la maximización de su complemento $\mathrm{tr}\left(\widehat{\mathbf{\Phi}}^{\mathrm{T}}\mathbf{\Phi}\right)$. Además, recordando la ecuación 6, tenemos que

$$\mathrm{tr}\left(\widehat{\mathbf{\Phi}}^{\mathrm{T}}\mathbf{\Phi}\right) = \mathrm{tr}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}) = \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{W}),$$

y así, el nuevo problema de maximización es

$$\begin{aligned} &\max \operatorname{tr}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}) \\ &\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}_{\mathrm{d}}. \end{aligned} \tag{9}$$

Para resolver el problema anterior, se puede escribir la función de Lagrange de la forma

$$\mathcal{L}(\mathbf{W}|\boldsymbol{\Phi}) = \operatorname{tr}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}) - \operatorname{tr}(\boldsymbol{\Lambda}(\mathbf{W}^{\mathrm{T}}\mathbf{W} - \mathbf{I}_{\mathrm{d}})),$$

donde $\boldsymbol{\Lambda} = Diag(\lambda_1, \dots, \lambda_{D_h})$ son los multiplicadores de Lagrange. Teniendo en cuenta la solución de la primera condición de LaGrange, se puede obtener el siguiente problema dual

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda} \Rightarrow \mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W} = \boldsymbol{\Lambda}. \tag{10}$$

Por lo tanto, una solución factible es cuando $\boldsymbol{\Lambda}$ y $\mathbf{W}$ son el valor propio y el vector propio de la matriz, respectivamente. Además, puesto que este es un problema de maximización, los vectores propios asociados a los valores propios *d* más grandes deben ser seleccionados. Similarmente pre multiplicando la ecuación 10 por $\boldsymbol{\Phi}^{T}$, se obtiene

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}\boldsymbol{\Lambda} \Rightarrow \boldsymbol{K}\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Lambda}, \tag{11}$$

por lo tanto el espacio incrustado $\mathbf{X}$ se puede calcular como los vectores propios de la matriz $\boldsymbol{K}$. Este enfoque es ampliamente explicado en [35].

4.1.2. Métodos Kernel

Hoy en día se ha tenido mayor desarrollo en métodos recientes que tienen por objetivo preservar la topología de los datos. Una topología de este tipo se puede representar en un gráfico basado en datos, construido como no dirigido y ponderado, en el cual los puntos de datos representan los nodos, y una matriz de afinidad y de similitud no negativa que contiene los pesos de las aristas. Esta representación es aprovechada por métodos basados en enfoques espectrales y de divergencia. Por un parte, para el enfoque espectral se puede representar los pesos de las distancias en una matriz de similitud, tal como sucede con el método LE (Laplacian Eigenmaps) [36]. Así también, usando una matriz de similitud no simétrica y enfocándose en la estructura local de los datos, surge el método denominado LLE (Locally Linear Embedding) [37]. Por la otra parte, una vez normalizada, la matriz normalizada también puede representar una distribución de probabilidad, haciendo el método basado en divergencia una incrustación estocástica de vecinos [38]. Dentro de los métodos RD clásicos cuyo objetivo es preservar la variancia o la distancia se tiene el método PCA (Principal Component Analysis) y el método CMDS (Classical Multi-dimensional Scaling) [3].

4.2. MODELO MATEMÁTICO GEOMÉTRICO PROPUESTO

Como se dijo antes, el vínculo directo para acoplar las propiedades de interactividad y controlabilidad con los métodos es el modelo, de esta manera se diseñó un novedoso modelo matemático geométrico (Mat-Geo) él cual tiene por objetivo la combinación de diferentes métodos RD espectrales no supervisados nombrados anteriormente. Dada la versatilidad de que los métodos espectrales se puedan representar por Kernel, la combinación se lleva a cabo teniendo en cuenta las matrices Kernel correspondientes. Este método se basa en un enfoque matemático geométrico que permite realizar la mezcla de los métodos RD de una forma interactiva, así, las matrices Kernel son linealmente combinadas de acuerdo a unos coeficientes respectivos derivados de las coordenadas de un punto dentro de una superficie geométrica. Así, usuarios –inclusive no expertos– pueden fácil e intuitivamente seleccionar un único método o una mezcla de métodos cumpliendo sus necesidades según el punto seleccionado mediante la exploración de la superficie de la forma geométrica. Las etapas desarrolladas se ilustran en la

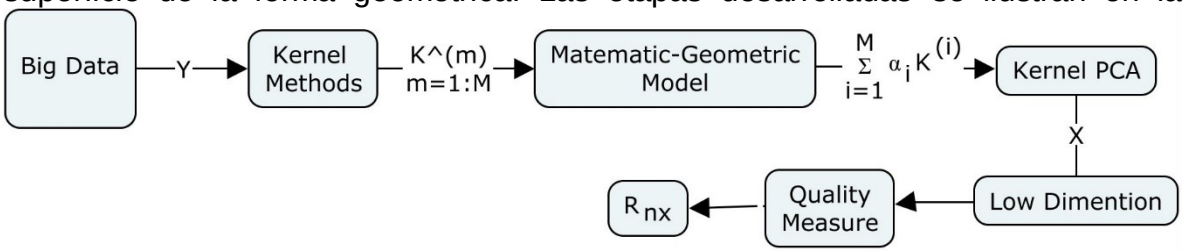


Figura *7*.

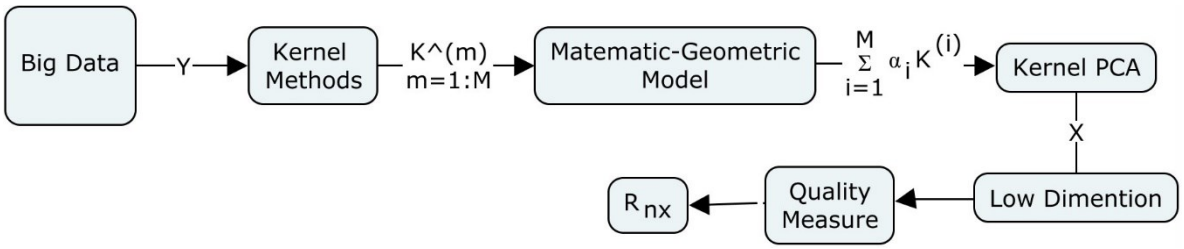
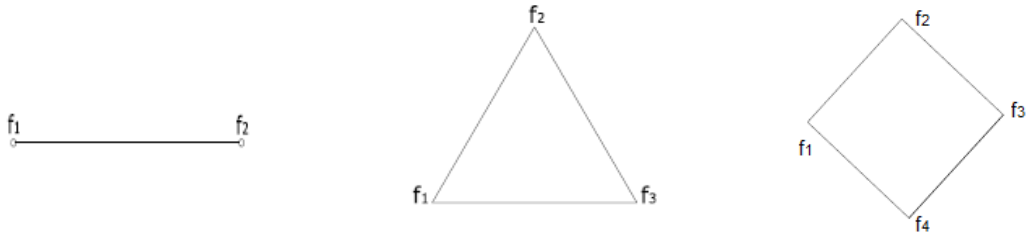


Figura 7. Diagrama ilustrativo de las etapas que comprende la interfaz propuesta

4.2.1. Enfoque poligonal

El enfoque poligonal que se plantea en esta tesis, parte de la idea de que la interactividad de la interfaz permite al usuario realizar una selección de los números de Kernel o métodos RD con los que desea trabajar, por tal motivo se debe generar un polígono que represente cada uno de estos métodos y que le permita bien sea la selección de un método, o realizar una mezcla de métodos. En general, cualquier conjunto de métodos se puede representar como un conjunto de funciones $\{f_1, \ldots, f_M\}$, de tal forma que *M* es el número de métodos considerados por el

usuario, luego nuestra idea radica en que el número de métodos considerados se constituyen en el número de vértices que contiene el polígono geometrico.



(a) *M* = 2 (b) *M* = 3 (c) *M* = 4

Figura 8. Enfoque poligonal para desarrollar la mezcla según el conjunto de funciones

Este modelo puede ser extendido a más de dos métodos dentro de un marco poligonal, en el cual una forma de representar tres funciones seria mediante un triángulo, cuatro funciones como un rombo, cinco funciones como un pentágono, etc. De esta forma, el polígono que visualizara el usuario tendrá *M* vértices que determinaran su forma, tal como se ilustra en la Figura 8.

4.2.2. Creación del polígono geométrico en Matlab

Principalmente se tiene como base que cada arista debe tener una longitud de 1. Según lo anterior se identifica puntos dentro de un plano cartesiano que determinen los vértices de cada poligono, dichos puntos coordenados en (x_i, y_i) con $i = \{2, \dots, M\}$ representan cada uno de los métodos deseados por el usuario, en este sentido para *M* métodos se tendrá *M* vértices.

- **Polígono para *M* = 2**

Empezaremos por describir la construcción del polígono geométrico para *M* = 2, como se puede apreciar en la Figura 8a, este polígono es una línea recta cuyos puntos de inicio y final representan a dos métodos. Su construcción parte de ubicar principalmente el punto de inicio en (0,0) y el punto final en (1,0), de esta forma obtenemos una línea recta de longitud 1 (ver Figura 9a), lo siguiente es centrar la línea en el punto (0,0), para esto solo basta con restar a las coordenadas de x_1 y x_2 un valor de 0.5, ya que de esta manera se obtienen coordenadas para los puntos de inicio y final de (-0.5,0) y (0.5,0), lo que permite obtener una línea recta de longitud 1 centrada en el punto (0,0) (ver Figura 9b).

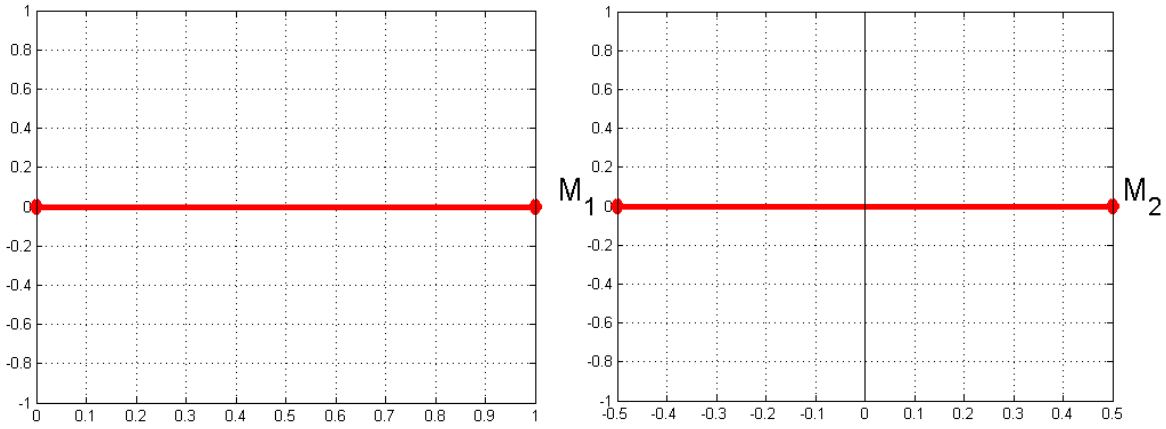


Figura 9. Construcción de la figura poligonal del enfoque geométrico para un numero de métodos $M = 2$

- **Polígono para $M = 3$**

A continuación presentamos la construcción del polígono para una selección de $M = 3$ métodos, como se puede notar en la Figura 10a, el polígono en este caso es un triángulo y se realiza partiendo del punto (0,0) para el primer vértice. Como la longitud de las aristas debe ser 1, entonces el siguiente vértice estará localizado en (1,0).

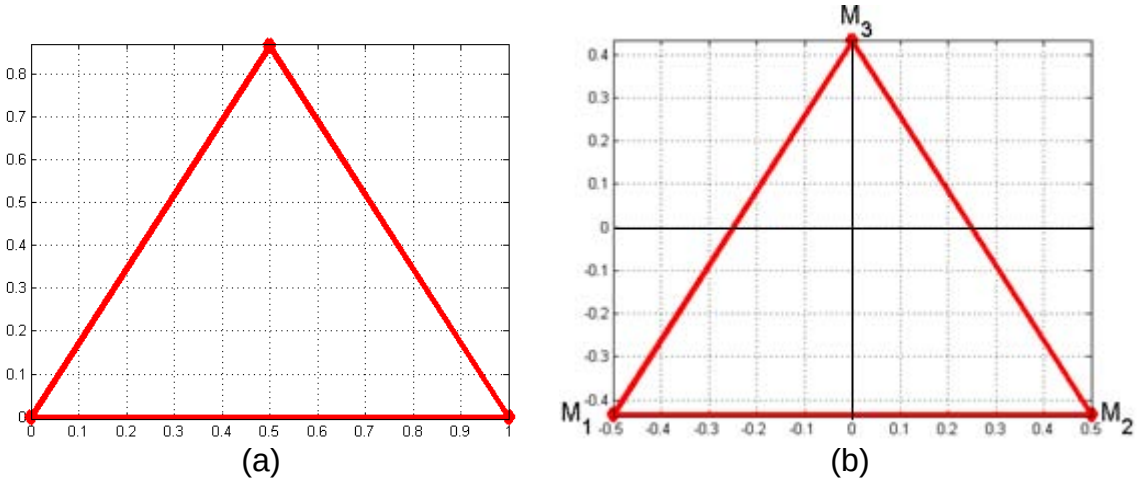


Figura 10. Construcción de la figura poligonal del enfoque geométrico para un numero de métodos $M = 3$

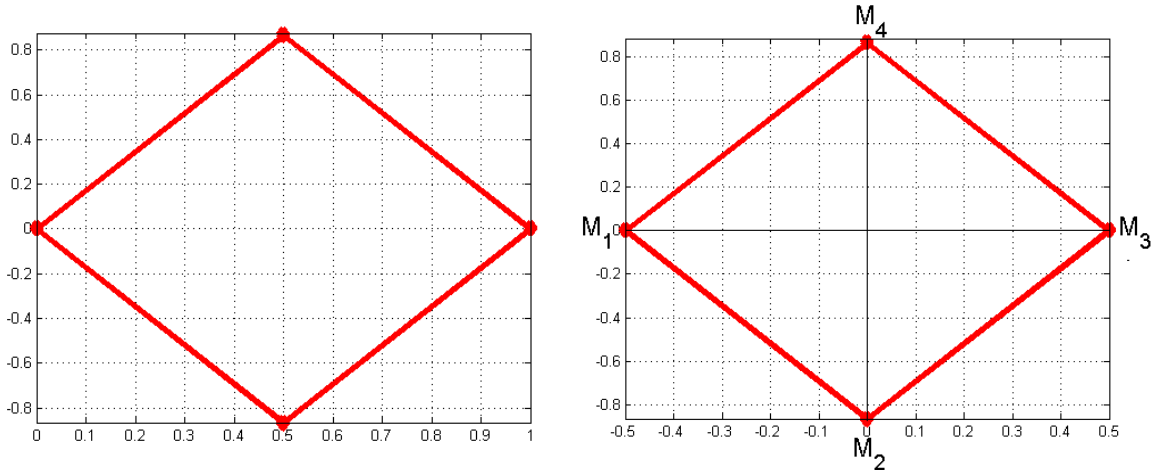
Para la obtención del tercer vértice imaginemos la primera división del triángulo de la Figura 10b como un triángulo rectángulo cuya hipotenusa h es la longitud de la arista comprendida entre el vértice 1 y el vértice 3, ahora es necesario aplicar el

teorema de Pitágoras según la ecuación 12 donde *h* es la longitud de la arista y b es la longitud de la base del triángulo que como es de notar, si la longitud entre el vértice 1 y el vértice 2 tiene también un valor de 1, el valor de b será de 0.5, así:

$$h^2 = a^2 + b^2 \rightarrow \ a^2 = h^2 - b^2 \rightarrow a = \sqrt{1^2 + \left(\frac{1}{2}\right)^2} = \frac{\sqrt{3}}{2} \qquad (12)$$

- **Polígono para *M* = 4**

Para una selección por el usuario de *M* = 4 métodos se debe graficar un polígono de cuatro vértices, para esto aplicaremos lo visto en el polígono anterior y simplemente duplicaremos el triángulo de forma simétrica con el eje x, de esta forma se obtiene el rombo, tal como lo ilustra la Figura 11



(a) (b)

Figura 11. Construcción de la figura poligonal del enfoque geométrico para un numero de métodos *M* = 4

Así, se puede inferir que los vértices de cada figura tendrán una longitud de 1, lo cual era una base principal para la construcción del modelo, por tanto, una vez creado el modelo geométrico se procede a definir la parte matemática. El acoplamiento entre la etapa geométrica y la etapa matemática del modelo es fundamental para la interfaz, ya que esta es la que se encarga de la interactividad y la controlabilidad. El diseño de un modelo matemático eficiente que se ajuste a las condiciones y requerimientos del usuario será abordado en las subsecciones siguientes en las cuales se explica el proceso de determinación de pesos a parir de la selección de un punto específico por el usuario.

4.2.3. Homotopía

Un concepto general de homotopía se refiere a un proceso de mapeo de una función continua en otra. Ahora, sean f_1 y f_2 dos funciones continuas asociadas a los espacios topológicos $\mathcal{X}$ y $\mathcal{Y}$ respectivamente. Una función de homotopía para estos dos espacios topológicos puede ser definida por

$$\begin{aligned} h\colon \mathcal{X} \times [0,1] &\to \mathcal{Y} \\ f_1, f_2 \quad \lambda \quad &\mapsto h(f_1, f_2, \lambda). \end{aligned} \tag{13}$$

Una manera para desarrollar la mezcla es la deformación de una función continua dentro de otra usando homotopía básica [39, 40], un modelo de homotopía simple puede ser escrito como $h\,(f_1, f_2, \lambda) = \lambda\, f_1 + (1 - \lambda)\, f_2$ donde λ es el parámetro de homotopía. En términos de una interfaz interactiva, tal parámetro deriva de la selección de un punto sobre una arista de longitud 1 comprendida entre dos métodos representados por las funciones f_1 y f_2 (ver Figura 8a), de tal forma que $h(f_1, f_2, 0) = f_1$ y $h(f_1, f_2, 1) = f_2$.

4.2.4. Mezcla de métodos RD

En términos de visualización de datos a través de métodos RD, los parámetros a ser combinados son las matrices Kernel, cada matriz corresponde a cada uno de los *M* métodos RD considerados, esto es $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}\}$. Por consiguiente, se obtiene una matriz Kernel $\widehat{\boldsymbol{K}}$ resultante de la mezcla de las *M* matrices Kernel, tal que

$$\widehat{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(m)}, \tag{14}$$

donde α_m es el coeficiente o peso ponderado correspondiente al método m y $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]$ es el vector que contiene los valores de ponderación. Como mencionamos anteriormente, estos coeficientes son asociados a las coordenadas del punto seleccionado por el usuario dentro de la superficie del polígono de *M* vértices.

La relación entre el punto seleccionado dentro de la superficie y los coeficientes ponderados para la combinación lineal está dada por la distancia comprendida entre cada vértice (el cual representa un determinado método) y el punto seleccionado por el usuario, tal como se aprecia en la Figura 12.

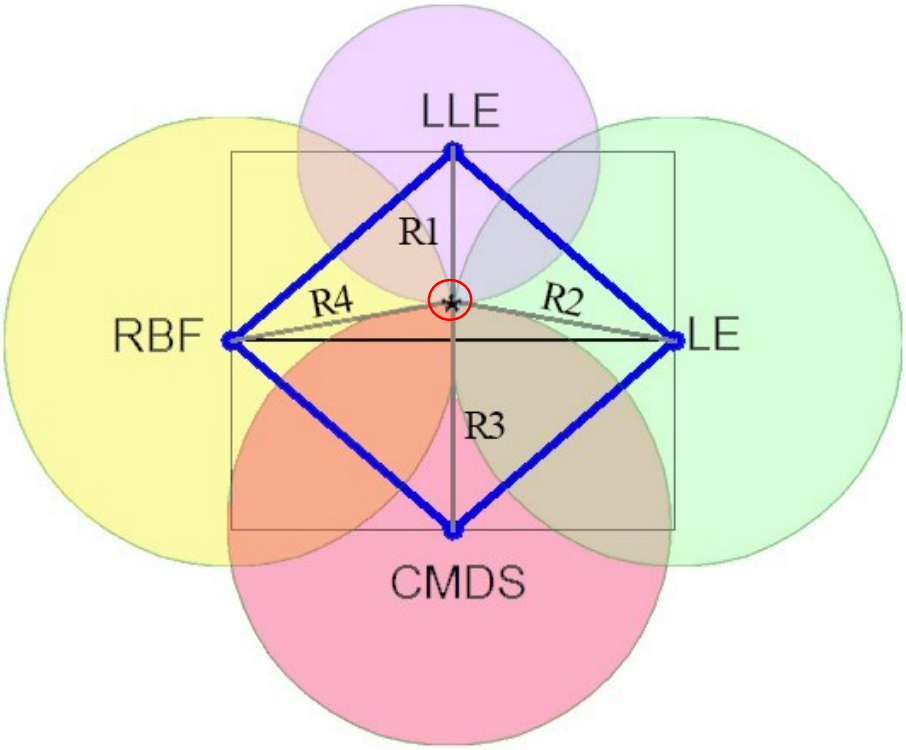


Figura 12. Grafica ilustrativa de la forma en cómo se estiman los coeficientes ponderados para *M* = 4

Tomaremos la Figura 12 a manera de ejemplo. Note que para este caso el usuario ha decidido tomar a consideración 4 métodos (*M* = 4) para mezclar, por lo que el polígono geométrico que representa esas 4 funciones es un rombo. Note también que el usuario ha seleccionado un punto (*) dentro de la superficie poligonal indicado dentro del círculo rojo, los factores de ponderación son determinados mediante la distancia que existe entre cada vértice y el punto seleccionado. Luego, estas distancias serán los radios $\{r_1, \dots, r_M\}$ de *M* círculos con áreas $\{A_1, \dots, A_M\}$ centradas en cada vértice. Por lo tanto, se hace evidente que el área del *m*-ésimo círculo está dada por

$$A_m = \pi r_m^2, \tag{15}$$

para este ejemplo, se tienen 4 círculos que tendrán adjuntas 4 áreas respectivamente.

Luego, las áreas obtenidas son normalizadas para sumar 1, pero como es de notar, si se selecciona un punto cercano a un método determinado la distancia entre ese método y el punto seleccionado será pequeña, por lo tanto su radio y el área del circulo centrado en ese punto serán pequeños también, por tal motivo se utiliza el

valor complementario de la área normalizada a fin de que el método más cercano al punto seleccionado sea el que tiene mayor peso de ponderación haciéndolo más significativo a la hora de mezclar los métodos, lo anteriormente dicho se puede apreciar en la ecuación 16.

$$1 - \frac{A_m}{\sum_{m=1}^{M} A_m}. \tag{16}$$

Sin embargo, no basta solo con aplicar la ecuación 4, ya que si el usuario selecciona un punto muy cercano a un vértice del polígono, si bien el método más cercano a dicho punto tendrá un valor más cercano a 1, pero los otros vértices no tendrán un valor cercano a cero sino que tendrán diferentes valores entre 0 y 1, por tal motivo se agrega un efecto de ecualización de forma que se obtengan coeficientes más acordes a las distancias y ubicación de los métodos para hacer la mezcla más eficiente. Dicho efecto se logra aplicando la función *Sinc*, en este sentido, los valores de $\boldsymbol{\alpha}$ están dados por

$$\alpha_m = \operatorname{sinc}\left(1 - \frac{A_m}{\sum_{m=1}^{M} A_m}\right). \tag{17}$$

4.3. MEDIDAS DE CALIDAD DE LOS MÉTODOS DE REDUCCIÓN.

Para determinar la medida de calidad de los métodos de reducción de dimensión se utiliza un criterio de calidad para valorar las diferentes integraciones mediante la conservación de los *k*-ésimos vecinos desarrollada en [32]. El objetivo de integrar esta medida de calidad a la interfaz propuesta es determinar el desempeño de los métodos RD en cuanto al agrupamiento o la integración se refiere, cabe aclarar que esta medida de calidad no determina el rendimiento computacional de dichos métodos.

La medida de calidad tratada por [32] se acopla a la interfaz como una curva de calidad, cuyo fundamento matemático según [32] parte del siguiente análisis. El rango de $\boldsymbol{\xi}_j$ respecto a $\boldsymbol{\xi}_i$ en el espacio de alta dimensión se denota como $\rho_{ij} = \left|\left\{k: \delta_{ik} < \delta_{ij} \text{ o } (\delta_{ik} = \delta_{ij} \text{ y } 1 \leq k < j \leq N)\right\}\right|$, donde $|A|$ denota la cardinalidad del conjunto A. Similarmente, en [32] definen que el rango de $\boldsymbol{x}_j$ respecto a $\boldsymbol{x}_i$ en el espacio de baja dimensión es $r_{ij} = \left|\left\{k: d_{ik} < d_{ij} \text{ o } (d_{ik} = d_{ij} \text{ y } 1 \leq k < j \leq N)\right\}\right|$. Los *k*-ésimos vecinos de $\boldsymbol{\xi}_i$ y de $\boldsymbol{x}_i$ son los conjuntos definidos por $v_i^K = \left\{j: 1 \leq \rho_{ij} \leq K\right\}$ y $n_i^K = \left\{j: 1 \leq \rho_{ij} \leq K\right\}$, respectivamente. Un primer índice de rendimiento puede ser denotado como

$$Q_{NX}(K) = \sum_{i=1}^{N} \frac{\left|\nu_i^K \cap n_i^K\right|}{KN}. \quad (18)$$

La ecuación 18 resulta en valores comprendidos entre 0 y 1 y mide el promedio normalizado de acuerdo a los *k*-ésimos vecinos correspondientes entre los espacios de alta dimensión y baja dimensión.

Ahora, se define una matriz de co-clasificación como $\boldsymbol{Q} = [q_{kl}]_{1\leq k,j\leq N-1}$ con $q_{kl} = \left|\{(i,j): \rho_{ij} = k \text{ y } r_{ij} = l\right.$. Por lo tanto, $Q_{NX}(K)$ cuenta *k*-por-*k* bloques de $\boldsymbol{Q}$, el rango preservado (en la diagonal principal) y las permutaciones dentro de los vecinos (en cada lado de la diagonal) [32]. En conclusión, la ecuación que representa la curva que se trabaja en [32] y que se integra a la interfaz está dada por

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N-1-K}, \quad (19)$$

para $1 \leq K \leq N-2$.

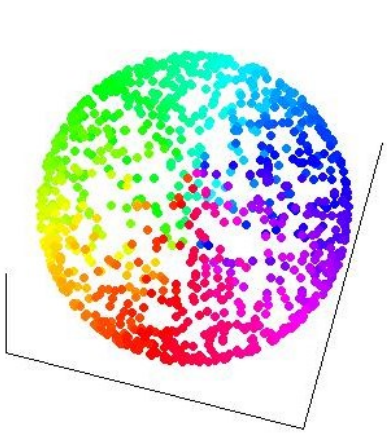
4.4. MARCO EXPERIMENTAL

Para los experimentos, se utilizaran bases de datos disponibles públicamente de la *UCI Machine Learning Repositorio* [41], así como un subconjunto de imágenes de la *Universidad de Columbia Image Library* [42]. Se considera los métodos convencionales de reducción de dimensiones espectral para evaluar el rendimiento de la mezcla del Kernel [36]. En particular, el método que se utiliza para generar los datos embebidos relacionados con la mezcla del Kernel es Kernel PCA (o KPCA). Se utiliza una versión reducida de la tasa media de acuerdo con los *k*-ésimos vecinos para cuantificar la calidad de los datos que se obtiene a partir de la integración [32]. La mezcla proporcionada representa cada enfoque de reducción de dimensiones, así como también ayuda a los usuarios a encontrar una representación adecuada de los datos incorporados en un marco visual e intuitivo.

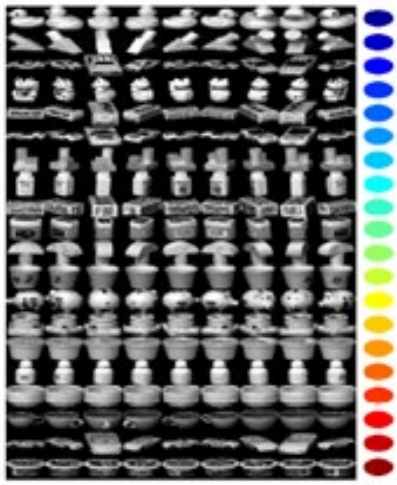
4.4.1. Base de datos

Los experimentos se llevaron a cabo sobre tres conjuntos de datos convencionales. El primer conjunto de datos es una cáscara esférica artificial (*N* = 1500 puntos de datos y *D* = 3). El segundo conjunto de datos es el banco de imágenes COIL-20 [42], que contiene 72 imágenes de nivel de gris que representan 20 objetos diferentes (*N* = 1440 puntos de datos – 20 objetos en 72 poses/ángulos – con D = 1282). El tercer conjunto de datos es un subconjunto seleccionado al azar del banco

de imágenes MNIST [43], que está formado por 6000 imágenes de nivel de gris correspondientes a cada uno de los 10 dígitos (*N* = 1500 puntos de datos – 150 instancias para los 10 dígitos – y *D* = 242). El cuarto conjunto de datos es denominado rollo suizo (*N* = 3000 puntos de datos y *D* = 3). En la Figura 13 se muestra ejemplos de los conjuntos de datos considerados.

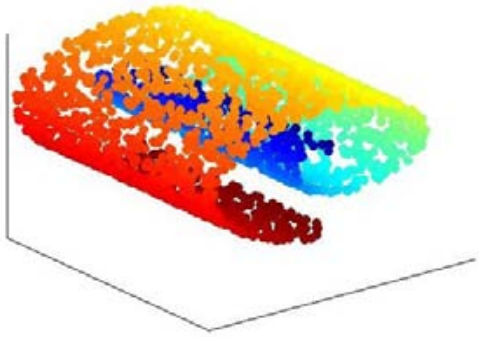


(a) Esfera



(b) COIL-20

(c) MNIST



(d) Rollo suizo

Figura 13. Las cuatro bases de datos experimentales

4.4.2. Controlabilidad de la interfaz propuesta

Las técnicas de reducción de dimensiones resultan significativamente más entendibles y manejables para el usuario cuando se utiliza propiedades de Info Vis que le permitan tener la libertad de seleccionar la mejor forma de representar los

datos, una de estas propiedades es la controlabilidad. Para ello se propone una estrategia geométrica que ajuste los factores de ponderación para realizar la combinación lineal de los métodos de RD a partir de aproximaciones Kernel de métodos convencionales. Dicha combinación es realizada por medio de las coordenadas adscritas a un punto seleccionado por el usuario dentro de la superficie poligonal, las cuales permiten determinar los pesos ponderados de cada método según el modelo matemático geométrico, así, si un resultado no se adecua a los intereses del usuario que manipula la interfaz entonces este puede seleccionar otro punto que satisfaga de mejor manera sus necesidades controlando la forma en cómo se da la mezcla mediante los pesos seleccionados sobre el polígono.

4.4.3. Interactividad de la interfaz propuesta

La comunicación entre el usuario y los sistemas informáticos es de gran importancia para dar significado a la información que se va a transmitir visualmente, por lo tanto es primordial establecer un vínculo entre la interfaz desarrollada y el usuario lo que se realiza mediante la propiedad de Info Vis denotada como interactividad. Para lograr tal objetivo se parte de la obtención de parámetros que permitan una visualización eficiente de grandes volúmenes de información que esté acorde a las necesidades que el usuario requiera, la interfaz desarrollada a través de un enfoque matemático-geométrico debe ser intuitiva permitiendo a personas no expertas obtener resultados según sus exigencias.

5. RESULTADOS Y DISCUSIÓN

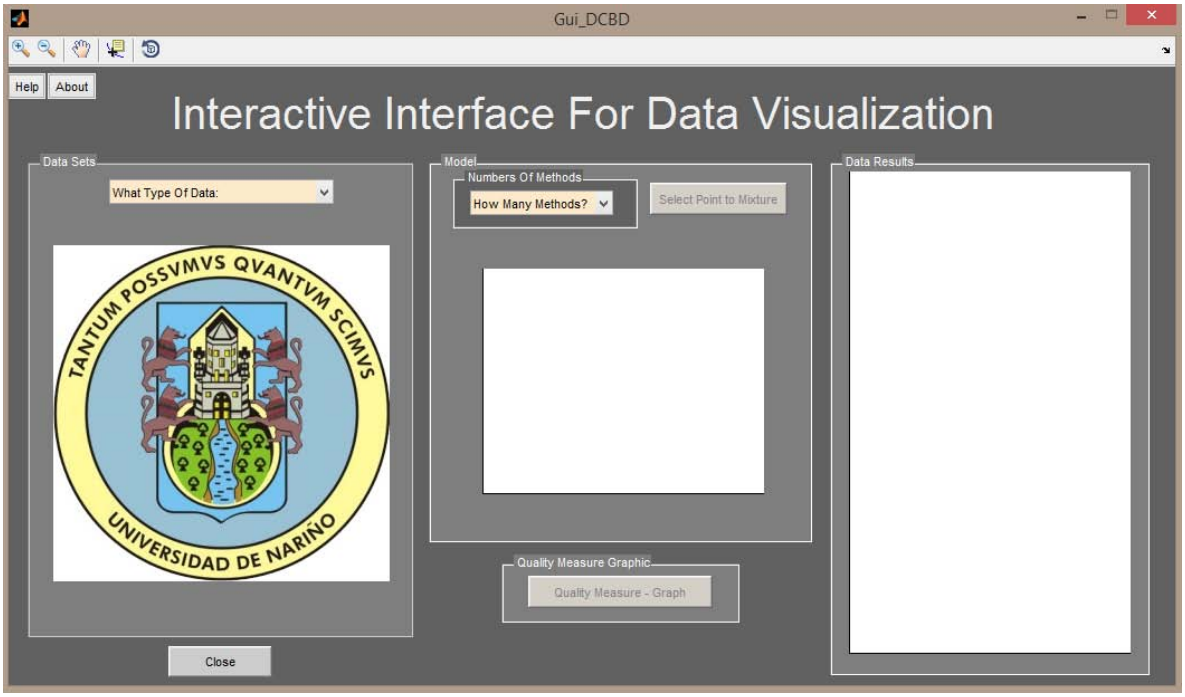
En esta sección se discuten los resultados experimentales mediante los cuales se realizan las pruebas a todos los conjuntos de datos de alta dimensión considerados en la sección 4 y se utiliza $R_{NX}(K)$ como indicador de calidad. Las matrices Kernel resultantes integran un algoritmo Kernel PCA para espacios de datos de salida con 2 dimensiones. Dado que todos los experimentos se llevan a cabo teniendo en cuenta cuatro métodos, las superficies poligonales son entonces una línea, un triángulo y finalmente un rombo, como se muestra en la Figura 8.

5.1. INTERFAZ INTUITIVA E INTERACTIVA

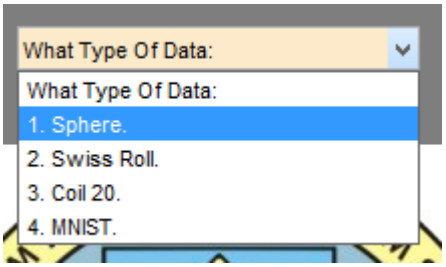
En esta subsección se presenta la interfaz implementada que permite al usuario interactuar de forma intuitiva mediante unos paneles organizados estratégicamente para seguir una secuencia de proceso facilitando la operación de dicha interfaz. La interfaz fue implementada en base al software MatLab mediante un lenguaje de programación de alto nivel, el cual se caracteriza por la forma en que se pueden expresar los algoritmos de manera tal que el ser humano los pueda comprender. La interfaz es de débil acoplamiento con el usuario debido a que este no conoce en si el algoritmo de operación pero si puede interactuar con la maquina según sus requerimientos.

Como se mencionó en la sección 3, una necesidad latente y en la que no se ha explorado ampliamente es lograr desarrollar un trabajo conjunto entre la minería de datos y la visualización de datos, por tal motivo se propone integrar las estas dos tecnologías a partir de las características más relevantes de la Info Vis con la reducción de dimensión mediante una interfaz interactiva e intuitiva. Nuestra interfaz (ver Figura 14a) no solo integra dichas características con las técnicas de reducción de dimensión, sino que también permite al usuario, quien intuitivamente la ópera, tener un control en la forma en cómo se da la mezcla Kernel PCA, lo que le conlleva directamente a controlar la forma en cómo se representan los datos. La interfaz además permite una interactividad en la selección del conjunto de datos (ver sección 4) que se quiere tratar como lo ilustra la Figura 14b, posteriormente el usuario debe seleccionar la cantidad de métodos a utilizar (ver Figura 14c) y a partir de esta cantidad se presenta el polígono geométrico con el cual se estima el valor de los pesos de ponderación para realizar la mezcla como lo evidencia la Figura 14f. El modelo matemático geométrico que hemos desarrollado es el que se encarga de analizar las coordenadas del punto seleccionado por el usuario y mediante un cálculo de distancias y áreas se obtiene un peso ponderado para cada método. Seguido de esto, la interfaz realiza un análisis de componentes principales mediante Kernel PCA lo que permite obtener una representación bidimensional de los datos que están ahora en baja dimensión como se aprecia en la Figura 14g. El objetivo es

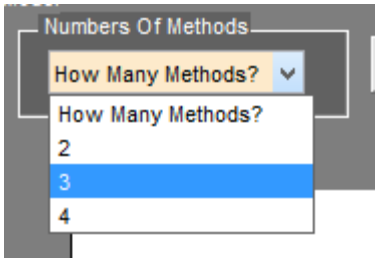
lograr una representación visual de los datos embebidos según los requerimientos del usuario, de no llegar a este resultado, el usuario puede seleccionar una cantidad diferente de métodos y nuevamente seleccionar un punto para realizar la mezcla, una vez más se obtiene una representación bidimensional de los datos. Este proceso se puede realizar las veces que sean necesarias por parte del usuario con el fin de satisfacer sus necesidades.



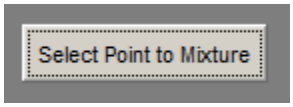
(a) Interfaz desarrollada.



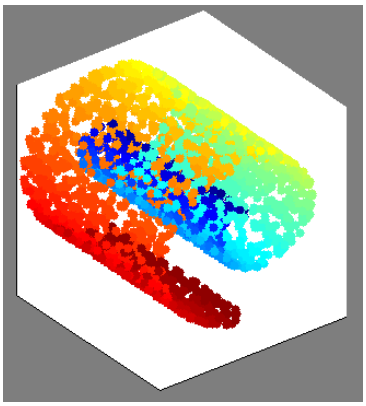
(b) Menú de selección del conjunto de datos.



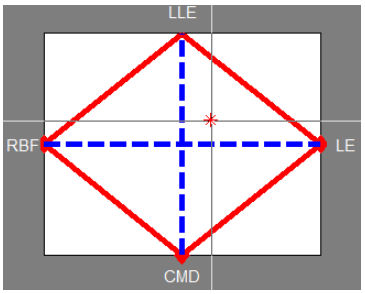
(c) Menú de selección del número de métodos.



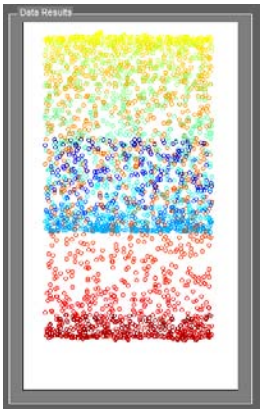
(d) Botón para seleccionar el punto para mezclar dentro de la superficie poligonal.



(e) Representación de los datos seleccionados. (Ejemplo: Rollo suizo)



(f) Polígono geométrico graficado a partir del número de métodos seleccionados. (Ejemplo: 4 métodos)



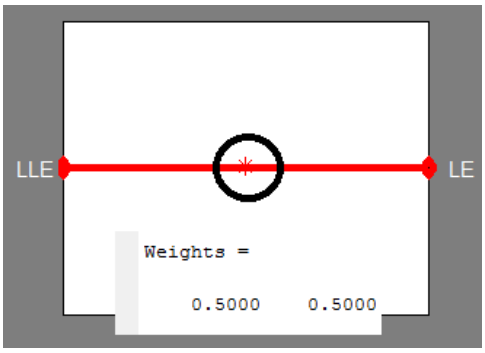
(g) Datos 2D obtenidos del proceso Kernel PCA

Figura 14. La interfaz de usuario interactiva e intuitiva implementada

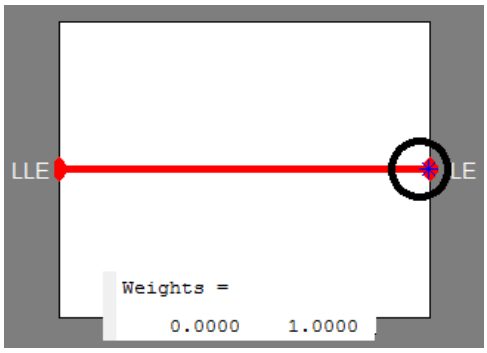
5.2. PRUEBA DE CONTROLABILIDAD DE LA INTERFAZ

Esta prueba consiste en seleccionar puntos dentro de la superficie de un polígono geométrico y determinar los valores de los pesos ponderados denotados como *Weights*. El círculo negro que se aprecia en la Figura 15 resalta el punto seleccionado por el usuario. Cabe resaltar que el usuario controla la forma en cómo se realiza la mezcla a partir de la selección de puntos dentro del polígono geométrico, el usuario no necesariamente sabe los métodos ni la mezcla Kernel PCA que se está dando, pero si el encuentra que el resultado satisface sus necesidades por se dará por concluido el proceso. En conclusión la característica de controlabilidad está enfocada en como el usuario mediante una interacción con el polígono geométrico puede controlar la mezcla realizada a partir de pesos ponderados estimados de la selección de puntos dentro la superficie.

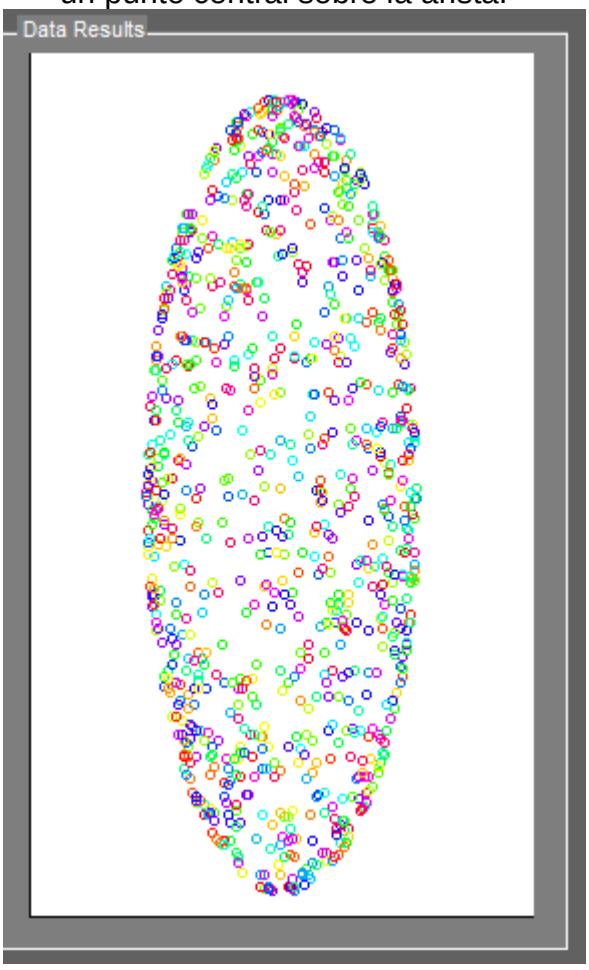
En la Figura 15 se puede apreciar el resultado de los pesos para los puntos seleccionados según el usuario donde se evidencia la variación de las representaciones dada una variación de puntos. Cada punto determina un valor ponderado diferente para cada método, por consiguiente la variación de las representaciones resultan de la variación de los pesos.



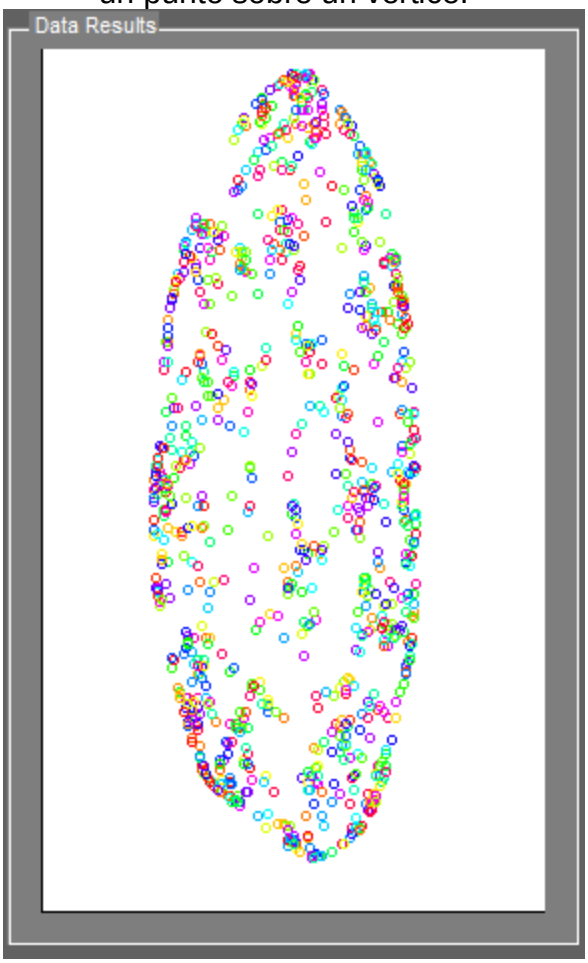
(a) Pesos estimados de la selección de un punto central sobre la arista.



(b) Pesos estimados de la selección de un punto sobre un vértice.



(c) Resultados obtenidos a partir del punto seleccionado



(d) Resultados obtenidos a partir del punto seleccionado

Figura 15. Controlabilidad de la interfaz por parte del usuario

5.3. PRUEBA DE INTERACTIVIDAD DE LA INTERFAZ

Dado que la mezcla que se presenta es una combinación lineal, cuando los coeficientes se seleccionan en el perímetro sólo se consideran dos Kernel. Entonces, el usuario puede apreciar la deformación resultante desplazándose en el borde respectivo desde un vértice a otro, de hecho, en la selección de coeficientes asociados a los vértices, se lleva a cabo el efecto de un único método. Además, al seleccionar puntos interiores se tiene en cuenta los efectos de cada método para calcular el Kernel resultante. Por lo tanto, el enfoque propuesto permite a los usuarios (incluso los que no experto) interactuar con los resultados DR seleccionando intuitivamente puntos de una superficie poligonal. En general, los resultados obtenidos se muestran en las Figura 16, Figura 17, Figura 18 y Figura 19 se aprecia que en la primera columna se representa los diferentes modelos geométricos que se pueden generar y la segunda columna representa los resultados obtenidos a partir de un punto seleccionado sobre la superficie poligonal. La última columna presenta la curva de desempeño de la representación de los datos obtenidos con la mezcla.

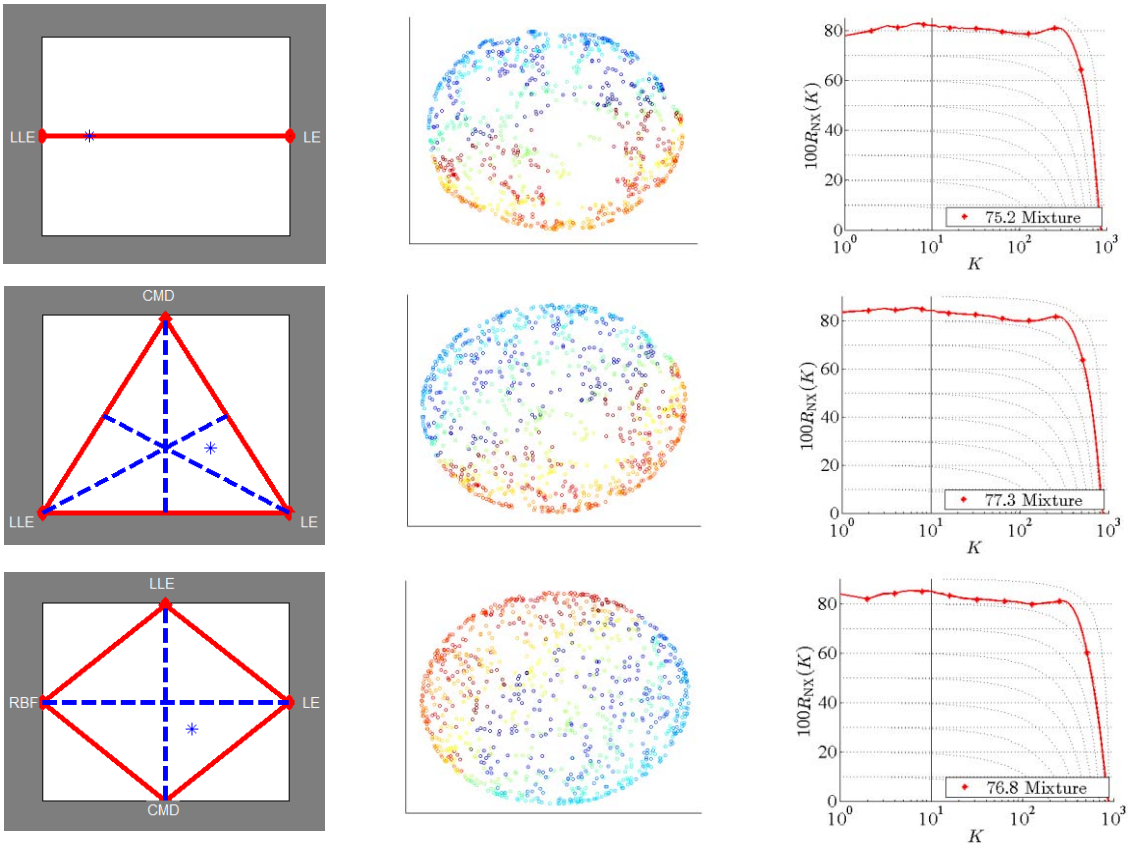


Figura 16. Resultados obtenidos a partir de la interfaz para la esfera

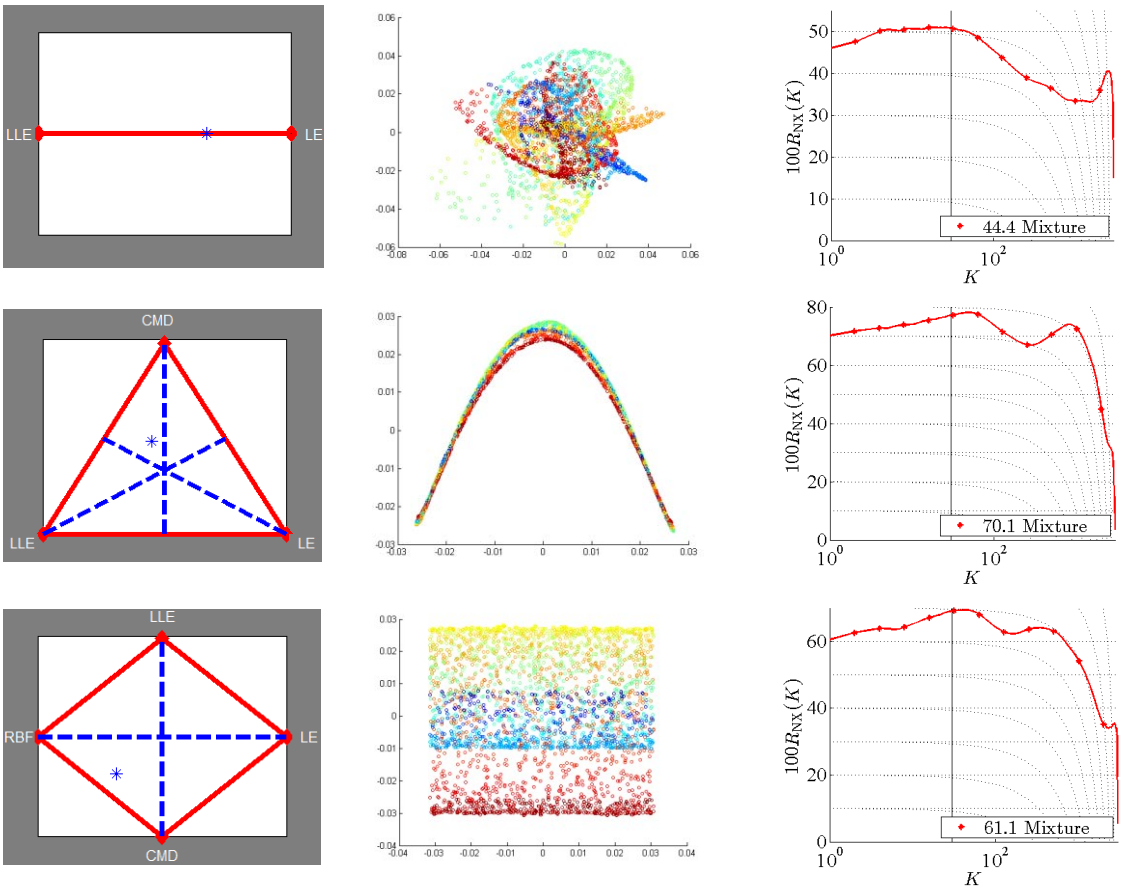
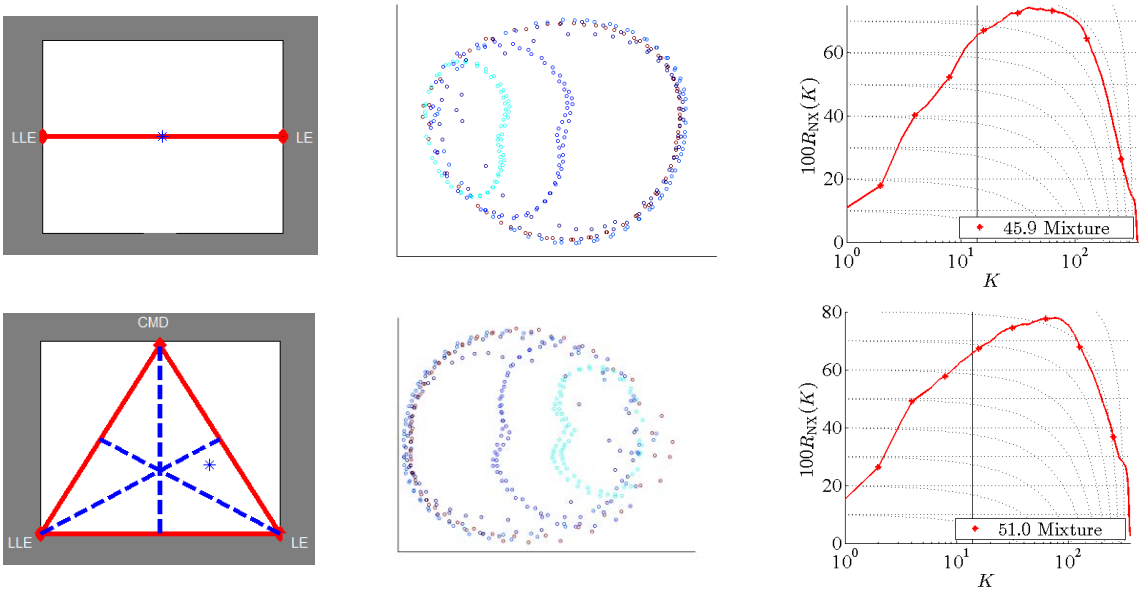


Figura 17. Resultados obtenidos a partir de la interfaz para el rollo suizo



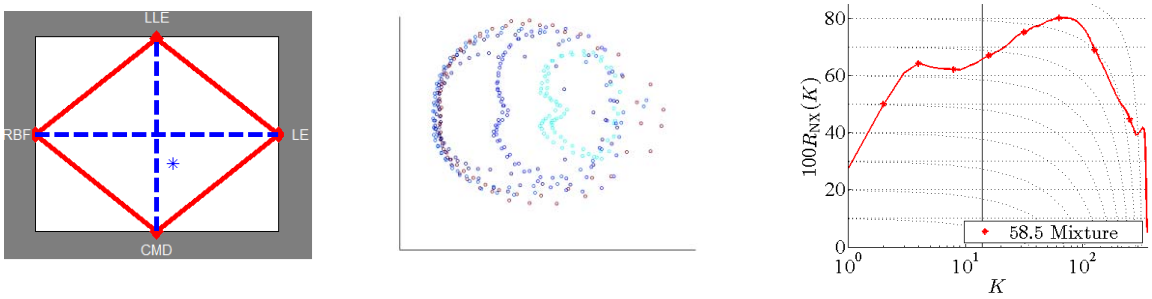


Figura 18. Resultados obtenidos a partir de la interfaz para el Coil 20

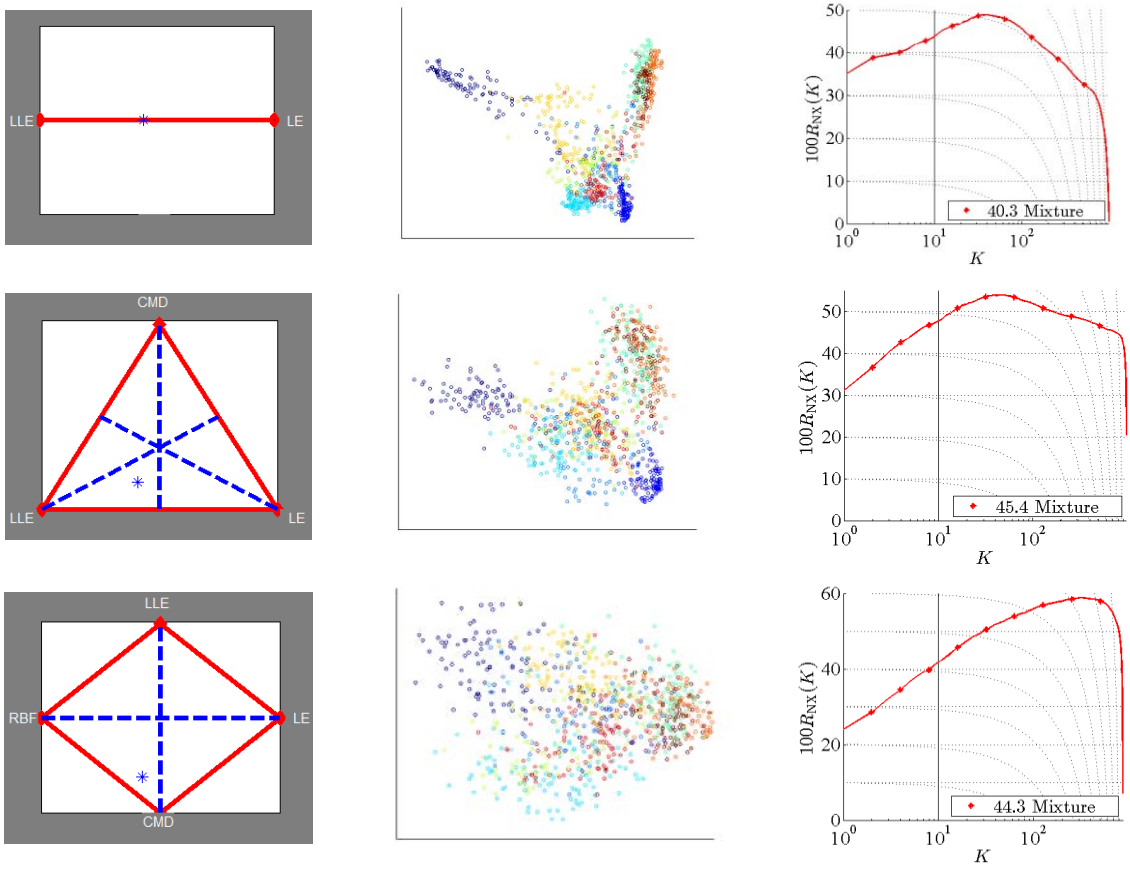


Figura 19. Resultados obtenidos a partir de la interfaz para el MNIST

6. CONCLUSIONES

El modelo matemático geométrico propuesto resulta ser un método novedoso para realizar la mezcla de los Kernel, de las bibliografía consultada muchos artículos realizan mezclas de maneras tradicionales o mediante otras propuestas, pero la propuesta que presentamos es mediante un enfoque geométrico. Este enfoque ha sido bien visto por investigadores en esta área lo que se puede evidenciar con la selección de este trabajo entre los 10 mejores de 200 en concurso en el XX Simposio en procesamiento de señales, imágenes e inteligencia artificial, STSIVA-2015 y esto a su vez cumple con un parámetro de los impactos esperados con este trabajo que era aportar a esta área con nuevos desarrollos tecnológicos.

El algoritmo desarrollado mediante programación secuencial es un método eficiente para el desempeño del sistema propuesto, ya que no es necesario volver a generar los Kernel para un mismo conjunto de datos, sino que se parte de estos mismos Kernel siempre y cuando el usuario quiera seleccionar un punto diferente para representaciones diferentes de los datos seleccionados.

La interfaz propuesta se basa en un enfoque interactivo que permite una visualización de datos embebidos resultantes de los métodos de reducción de dimensión (RD). En particular, el enfoque propuesto se basa en Kernel con una perspectiva geométrica de homotopía que hace referencia a un conjunto de funciones. En este caso, la representación de métodos espectrales, no supervisados de RD está dada por matrices Kernel. En concreto, se utiliza este método para llevar a cabo la combinación lineal de las matrices del Kernel dadas por la relación entre los puntos internos ubicados dentro de una superficie poligonal con los valores de los pesos de ponderación. Haciendo eso, se obtiene una versión interactiva de Kernel PCA. Dado el gráfico y el marco de referencia intuitivo, desde nuestra interfaz un usuario independientemente si es o no experto puede fácilmente seleccionar un método o combinación de métodos escogiendo puntos en una superficie poligonal para cumplir sus necesidades específicas.

Las características de la Info Vis se constituyen en un método clave para proporcionar al usuario un rol de operador en la forma como se realiza la mezcla de las matrices Kernel y de esta forma el usuario puede determinar el resultado más acorde a sus necesidades. Dichas características, que son la controlabilidad y la interactividad, permiten que la interfaz desarrollada trabaje según parámetros de operación determinados por el usuario quien determina la mejor forma de realizar la mezcla a partir de puntos seleccionados sobre la figura geométrica.

Algunas de las herramientas existentes no permiten al usuario interactuar con los datos y otras no permiten controlar la forma como se obtiene la integración de los resultados, con las características de Info Vis acopladas, el usuario no solo realiza una interacción usuario-maquina sino que entra en un ciclo de realimentación siendo

la comunicación visual la que determina la situación de los resultados, y la selección del punto el parámetro de control de cómo se realiza la mezcla.

La integración de técnicas de minería de datos con características de la Info Vis mediante un enfoque geométrico, resulta en una alternativa potencial que permite a usuarios no necesariamente expertos, manipular los grandes conjuntos de datos para obtener datos más acordes a sus necesidades, más inteligibles y naturales para su interpretación sin la supervisión de expertos en interpretación de resultados para poder comprender la información obtenida o extraer verdadero conocimiento.

La reducción de dimensión o RD es una técnica muy importante en el proceso de minería de datos, ya que puede ser destinada a las fases de filtrado de datos y selección de variables, realizando de estas dos fases en la misma técnica facilitando la extracción del conocimiento y permitiendo una forma más natural e inteligible de interpretarlos.

RECOMENDACIONES

En el proceso de minería de datos, es importante tener en cuenta que las representaciones o resultados son destinados a personas que no siempre son expertos en el tema. Es necesario crear sistemas que permitan a los usuarios no expertos obtener representaciones más naturales o inteligibles ya que si bien, es factible recurrir a expertos, esto toma tiempo, esfuerzo y costos monetarios convirtiendo a los sistemas en general no eficientes en estos factores.

Existe la necesidad de seguir explorando y principalmente desarrollar nuevas aproximaciones Kernel que permitan obtener mejores representación de datos embebidos, ya que si bien la importancia de la interfaz radica en el acoplamiento de características de Info Vis y métodos de reducción de dimensión, las representaciones Kernel son las que se encargan de las dos primeras fases de la minería de datos, y si se logra desarrollar mejores aproximaciones Kernel de técnicas RD esto permitirá obtener resultados muchos más precisos e inteligibles.

Se debe tener en cuenta que la interfaz parte un algoritmo base creado en MatLab, este cambio de script a interfaz no resulta una tarea fácil ya que se debe tener en cuenta que para la representación de datos como el MNIST y el Coil se deben realizar mosaicos de las imágenes contenidas y a estas asignarles una paleta de colores, lo que resulta tedioso ya que se debe orientar siempre al objeto (grafica) en el que se están representando los datos. De otra manera persistirá un error en la representación gráfica de cada objeto (grafica) presente en la interfaz.

Existen diferentes técnicas de minería de datos que se pueden trabajar en miras de un acople con las características de Info Vis y realizar una comparación cuantitativa del desempeño de estas técnicas. Si bien muchas de estas responden eficientemente a los problemas de grandes conjuntos de datos, no representan eficientemente estos resultados. Por otra parte, otras técnicas no permiten siquiera aplicarse a datos que no han sido filtrados o preprocesados por lo que intentar acoplarlas con Info Vis sería una tarea tediosa y sin esperar muy buenos resultados.

Igual que en la minería de datos, existen diferentes técnicas de Info Vis que se pueden trabajar para buscar sistemas más eficientes de representación de datos y agrupamiento. Todo en aras de buscar representaciones más precisas e inteligibles hacia el usuario ya que no siempre será un usuario experto.

REFERENCIAS

[1] E. Bertini y D. Lalanne, «"Surveying the complementary role of automatic data analysis and visualization in knowledge discovery" Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration,» *ACM,* 2009.

[2] D. H. Peluffo-Ordoñez, J. A. Lee y M. Verleysen, «Short review of dimensionality reduction methods based on stochastic neighbor embedding". Advances in Self-Organizing Maps and Learning Vector Quantization.,» *Springer International Publishing,* pp. 65-74, 2014.

[3] I. Borg y P. Groenen, Modern multidimensional scaling: Theory and applications., Springer Science & Business Media, 2005.

[4] W. Dai y H. Peng, «"Research on Personalized Behaviors Recommendation System Based on Cloud Computing.",» *TELKOMNIKA Indonesian Journal of Electrical Engineering 12.2,* pp. 1480-1486, 2013.

[5] M. Ward, G. Gristein y D. Keim, «Interactive data visualization: foundations, techniques, and applications,» *AK Peters, Ltd.,* 2010.

[6] P. C. Wong, «Visual Data Mining.,» de *Computer Graphics and Applications.*, 1999, pp. 20-21.

[7] Q. Yang y X. Wu, «10 challenging problems in data mining research,» *International of Information Technology & Decision Making, 5(04),* pp. 597-604, 2006.

[8] H. Geppert, M. Vogt y J. Bajorath, «Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation,» *Journal of chemical information and modeling, 50(2),* pp. 205-216, 2010.

[9] J. C. Riquelme, R. Ruiz y K. Gilbert, «Inteligencia Artificial,» de *Minería de Datos: Conceptos y tendencias*, vol. 10, 2006.

[10] S. J. Vallejos, Minería de Datos, Universidad Nacional del Nordeste, Argentina, 2006.

[11] D. A. Keim, «Visual Database Exploration Techniques.,» de *Visual Techniques for Exploring Databases*, 1997.

[12] D. Keim, F. Mansmann, J. Schneidewind y H. Ziegler, «Challenges in Visual Data Analysis,» de *Proceedings of Information Visualization*, 2006, pp. 9-16.

[13] C. Ahlberg y E. Wistrand, «An Information Visualization and Exploration Environment,» *Int'l Symp, Information Visualization,* pp. 66-73, 1995.

[14] A. Kerren, A. Ebert y J. Meyer, Human-centered visualization environments, 2006.

[15] C. P. Lopez, Minería de datos: técnicas y herramientas, Paraninfo, 2007.

[16] M. Tascón, «Introducción: Big Data. Pasado, presente y futuro,» de *Telos: Cuadernos de comunicación e innovación*, 2013, pp. 47-50.

[17] D. Pimentel, M. Cataldi y G. Muñiz, «De la visualización a la sensorización de información,» *Blucher Design Proceedings,* pp. 129-133, 2013.

[18] J. C. Alvarado-Perez y H. Bolaños Ramires, *Descubrimiento de conocimiento en bases de datos: La perspectiva de la visualización inteligente de la información,* 2014.

[19] D. A. Keim y H. P. Kriegel, «Visualization Techniques for Mining Large Databases: A compraison,» *IEEE Trans. Knowledge and Data Eng, 8(6),* pp. 923-936, 1996.

[20] W. S. Cleveland, «Visualizing Data,» *Horbart Press,* 1993.

[21] A. Inselberg y B. Dimsdale, «Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry,» *IEEE Visualization '90,* pp. 361-375, 1990.

[22] A. Inselberg, «The Plane with Parallel Coordinates,» de *The Visual Computer*, 1985, pp. 69-91.

[23] A. Inselberg, «Multimensional Detective,» *Information Visualization (InfoVis '97),* pp. 100-107, 1997.

[24] P. E. Hoffman, Table Visualizations: A Formal model and Its Applications, University of Massachusetts.

[25] D. A. Keim y H. P. Kriegel, «VisDB: Database Exploration Using Multimensional Visualization,» de *Computer Graphics and Applications*, 1994, pp. 40-49.

[26] C. G. Beshers y S. K. Feiner, «Automated Design of Data Visualizations,» de *Scientific Visualization - Advances and Challenges*, 1994, pp. 88-102.

[27] R. J. Hendley, N. S. Drew, A. M. Wood y R. Beale, «Narcissus: Visualizing Information,» *Int'l Symp. Informaction Visualization (InfoViz '95),* pp. 90-96, 1995.

[28] E. Tufte, «The Visual Display of Quantitative information,» de *Graphics Press*, 1983.

[29] M. Mramor, G. Leban, J. Demsar y B. Zupan, «Visualization-based cancer cicroarray data classification analysis,» de *Bioinformatics*, Oxford, 2007, pp. 2147-2154.

[30] J. A. Ochoa y J. F. Trinidad , «Reconocimeinto de patrones,» 2004.

[31] J. Ruiz-Shulcloper, A. Guzman Arenas y J. F. Martinez-Trinidad, «Seleccion de Variables y Clasificación Supervisada,» de *Enfoque Lógico Combinatorio al Reconocmiento de Patrones I*, 1999.

[32] J. A. Lee, E. Renard, G. Bernard, P. Dupont y M. Verleysen, «Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation,» *Neurocomputing,* 2013.

[33] J. Ham, D. D. Lee, S. Mika y B. Scholkopf, «A kernel view of the dimensionality reduction of manifolds,» *Proceedings of the twenty-first international conference on Machine Learning, ACM,* p. 47, 2004.

[34] J. Cook, I. Sutskever, A. Mnih y G. E. Hinton, «Visualizing similarity data with a mixture of maps,» *International Conference on Artificial Intelligence and Statistics,* pp. 67-74, 2007.

[35] D. Peluffo-Ordóñez, J. Lee y M. Verleysen, «Generalized kernel framework for unsupervised spectral methods od dimensionality reduction,» *IEEE Symposium Series on Computational Intelligence,* 2014.

[36] M. Belkin y P. Niyogi, «Laplacian eigenmpas for dimensionality reduction and data representation,» de *Neural computation*, vol. 15, 2003, pp. 1373-1396.

[37] S. T. Roweis y L. K. Saul, «Nonlinear dimensionality reduction by locally linear embedding,» de *Science*, vol. 290, 2000, pp. 2323-2326.

[38] G. E. Hinton y S. T. Roweis, «Stochastic neigghbor embedding,» de *Advances in neural information processing systems*, 2002, pp. 833-840.

[39] J. Harmann, M. P. Murphy, C. S. Peters y P. C. Staecker, «Homotopy equivalence in graph-like digital topological spaces,» 2014.

[40] D. H. Peluffo-Ordoñez, J. C. Alvarado-Perez, J. A. Lee y M. Verleysen, «Geometrical Homotopy for data visualization,» *Computational Intelligence and Machine Learning,* 2015.

[41] M. Lichman, «UCI Machine learning repository,» 2013. [En línea]. Available: http://archive.ics.uci.edu/ml..

[42] S. A. Nene, S. K. Nayar y H. Murase, «Columbia object image library (coil-20),» 1996. [En línea]. Available: http://www.cs.Columbia.edu/CAVE/coil-20.html.

[43] Y. LeCun, L. Bottou, Y. Bengio y P. Haffner, «Gradient-based learning applied to document recognition,» *Proceedings of the IEEE 86(11),* 1998.

ANEXOS

Esta sección ha sido destinada a los resultados tangibles logrados con el trabajo realizado en esta tesis. Estos anexos contienen una descripción más ampliada de los resultados mencionados en la sección 6 donde exponemos los detalles más relevantes.

ANEXO 1. PSEUDOCODIGO DEL SCRIPT DE PROGRAMACION

Pseudocodigo Info Vis

1. **Inicio**
2. **Escribir** “Ingrese tipo de datos: ”; % 1=Esfera, 2=Rollo, 3=MNIST, 4=Coil.
3. **Leer** opt;
4. **Switch** opt
    a. **Caso 1:** crear esfera;
    b. **Caso 2:** crear rollo suizo;
    c. **Caso 3:** cargar MNIST;
    d. **Caso 4:** cargar Coil;
5. **Graficar** datos;
6. **Escribir** “Ingrese número de métodos: ”;
7. **Leer** Ne;
8. **Switch** Ne
    a. **Caso 1:** crear línea;
    b. **Caso 2:** crear triangulo;
    c. **Caso 3:** cargar rombo;
9. **Hacer** aproximaciones Kernel;
10. **Hacer** asociar Kernel a vértices; % Kernel *i* en el vértice (x(i),y(i)).
11. **Escribir** “Seleccione un punto dentro de la superficie del polígono: ”;
12. **Leer** c1, c2; % c1 = coordenada en x, c2 = coordenada en y
13. **Hacer** r(i) = norma([c1-x(i) c2-y(i)]); % cálculo de distancias entre punto seleccionado y métodos deseados.
14. **Hacer** A(i) = π*r(i)^2; % cálculo de áreas a partir de distancias determinadas.
15. **Hacer** normalizar áreas;
16. **Hacer** Pesos(i) = 1 – Area_normalizada(i);
17. **Hacer** Pesos = 1 - sinc(Pesos.^4);% calcular pesos ponderados;
18. **Hacer** M = {Kernel 1, Kernel 2, Kernel 3, Kernel 4} **%** Matriz embebida con kernels.
19. **Para** i=1 hasta i=Ne **Hacer** Ksum = Ksum + Pesos(j)*M{j};
20. **Hacer** [Xsum,~] = eigs(Ksum); % Calcular eigenvectores
21. **Graficar** Xsum; % Graficar matriz de baja dimensión.

22. **Escribir** “Desea graficar curva de calidad: “
23. **Leer** resp;
24. **Si** resp = si **Graficar** Rnx; % Grafica de la curva de calidad.
25. **Fin**

ANEXO 2. CODIGO DEL PROGRAMA PARA ANALISIS VISUAL

```
clear, clc, close all;
% Reemplazar en opt un numero que indique el tipo de datos a analizar
opt = 1; % 1 for Sphere, 2 for Swiss Roll, 3 for MNIST, 4 for coil 20

%% Funcion que contiene los diferentes sets de datos.

% 1. Esfera.
% 2. Rollo suizo.
% 3. MNIST
% 4. Coil 20

switch opt

    case 1
        spd = 3;
        nbr = 1000;
        X = randn(nbr,3);
        X = bsxfun(@rdivide, X, sqrt(sum(X.^2,2)));

        if spd>3
            X = [X,0.5*randn(nbr,spd-3)./sqrt(spd-3)];
        end

        L = 32+64/180*atan2(X(:,1),X(:,2));
        str = ['Sph',num2str(spd)];
        save(['data_',str],'X','L','str');
        colmap = hsv(64);

        figure(1);
        subplot(131);
        scatter3(X(:,1),X(:,2),X(:,3),30,L,'o','filled');
        view(15,75);
        axis equal;
        colormap(colmap);
        rotate3d on;
        Y = X;

        set(gca,'Fontname','Times','Fontsize',12);
        set(gca,'Fontname','Times','Fontsize',20);
        set(gca,'xtick',[])
        set(gca,'xticklabel',[])
        set(gca,'ytick',[])
```

```
    set(gca,'yticklabel',[])
    set(gca,'ztick',[])
    set(gca,'zticklabel',[])

case 2

    % Rollo Suizo
    s = RandStream('mcg16807','Seed',29);
    RandStream.setGlobalStream(s);
    str = 'swissroll';
    N = 3000;
    t = (3*pi*(rand(N,1).^0.65)+pi/2);
    height = 100*rand(N,1);
    Y = [t.*cos(t) height  t.*sin(t)];
    L = t;

    figure(1);
    subplot(131);
    scatter3(Y(:,1),Y(:,2),Y(:,3),50,t,'o','filled')
    set(gca,'Fontname','Times','Fontsize',20);
    set(gca,'xtick',[])
    set(gca,'xticklabel',[])
    set(gca,'ytick',[])
    set(gca,'yticklabel',[])
    set(gca,'ztick',[])
    set(gca,'zticklabel',[])

case 3

    % MNIST digits
    nbr = 1000;
    load 'mnist_train_1.mat'
    sel = randperm(60000);
    sel = sel(1:nbr);
    X = train_X(sel,:);
    L = train_labels(sel);
    str = 'MNIST';
    save(['data_',str],'X','L','str');
    colmap = jet(64);
    colmap = colmap(5:6:64,:);
    nr = 10;
    nc = 10;
    Xs = X(1:100,:);

    for i = 1:10
        tmp = X(L==i,:);
        Xs(nc*(i-1)+(1:nc),:) = tmp(1:nc,:);
    end

    [r,c] = meshgrid(1:nr,1:nc);

    figure(1);
    subplot(131);
```

```
        mosaic1([c(:),r(:)],Xs(1:(nr*nc),:),28,28,nc,nr,true);
        scatter((nc+0.5)*ones(nr,1),(1.5:(nr-2)/(nr-1):...
            nr-0.5)',450,colmap,'o','filled')

        Y = X;
        clear X
        knn = 30;
        kop = 1;
        t = L;
        N = nbr;
        d = 2;
        l = 200;
        opts.maxit = 100; opts.runtime = 15; opts.tol = 1e-3;

        set(gca,'xtick',[])
        set(gca,'xticklabel',[])
        set(gca,'ytick',[])
        set(gca,'yticklabel',[])
        set(gca,'Fontname','Times','Fontsize',20);

    case 4

        s = RandStream('mcg16807','Seed',29);
        RandStream.setGlobalStream(s);
        load 'coil_1440.mat'
        L = reshape(bsxfun(@plus, (1:20), zeros(72,1)), [1440,1]);
        stp = 1;
        X = X(1:stp:end,:);
        L = L(1:stp:end);
        colmap = jet(20);
        str = 'Coil20';
        nr = 20;
        nc = 9;
        [r,c] = meshgrid(1:nr,1:nc);
        Y = X;

        figure(1);
        subplot(131);
        mosaic1([c(:),r(:)],X(1:8/stp:1440/stp,:),128,128,9,20);
        scatter((nc+0.5)*ones(nr,1),(1.5:(nr-2)/(nr-1):...
        nr-0.5)',150,colmap,'o','filled')
        set(gca,'xtick',[])
        set(gca,'xticklabel',[])
        set(gca,'ytick',[])
        set(gca,'yticklabel',[])
        set(gca,'Fontname','Times','Fontsize',20);

end

%% Graficas de los datos
figure(1)
subplot(131);
set(gca,'xtick',[])
```

```
set(gca,'xticklabel',[])
set(gca,'ytick',[])
set(gca,'yticklabel',[])
title('Data Set','Color','red','FontSize',17);
subplot(232);
set(gca,'xtick',[])
set(gca,'xticklabel',[])
set(gca,'ytick',[])
set(gca,'yticklabel',[])
title('Model','Color','red','FontSize',17);
subplot(235);
title('Quality Measure','Color','red','FontSize',17);
subplot(133);
set(gca,'xtick',[])
set(gca,'xticklabel',[])
set(gca,'ytick',[])
set(gca,'yticklabel',[])
title('Data Result','Color','red','FontSize',17);

%% Kernels

% Kernel 1: LLE method

d = 2;
knn = round(size(Y,1)/100);

[X_LLE, ~,M,conn_comp] = lle(Y, d, knn);
L   = L(conn_comp);
Y   = Y(conn_comp,:);

nbr       = length(conn_comp);
lamb      = eig(M);
lamb      = max(lamb);
K_LLE     = lamb*eye(nbr) - M;

K         = K_LLE;
K         = 0.5*(K + K');
kS        = sum(K,1)./nbr;
K1        = K - bsxfun(@plus, kS, kS') + sum(kS)/nbr;
K1        = K1/max(max(abs(K1)));

%% Kernel 2: CMDS method

DX = pairwisedistances(Y);
DXL = DX;

S0 = DX.^2;
sS = sum(S0,1)./nbr;
S0 = -1/2*(S0 - bsxfun(@plus, sS, sS') + sum(sS)/nbr);

K_CMDS = -0.5*(eye(nbr) - ones(nbr))*DX.^2*(eye(nbr) - ones(nbr));
kS     = sum(K_CMDS,1)./nbr;
```

```
K_CMDS = K_CMDS - bsxfun(@plus, kS, kS') + sum(kS)/nbr;
K      = K_CMDS;
K2     = 0.5*(K + K');
K2      = K2/max(max(abs(K2)));

%% Kernel 3: LE method

rng(1);

nl = 0; d = 2;
[Wp,beta] = x2p(Y',knn); Wp = (Wp+Wp')/2;
tic; [XLE, LL] = lapeig(d,Wp,nl); t1 = toc;

try
       K_LE   = pinvs(LL);
catch
       K_LE   = pinv(LL);
end

K      = K_LE;
K      = 0.5*(K + K');
kS     = sum(K,1)./nbr;
K3     = K - bsxfun(@plus, kS, kS') + sum(kS)/nbr;
K3     = K3/max(max(abs(K3)));

%% Kernel 4: RBF

K4 = gaussaff(Y,{'K',10},0.5);

%% Normalización

K1 = K1./repmat(max(K1),size(K1,1),1);
K2 = K2./repmat(max(K2),size(K2,1),1);
K3 = K3./repmat(max(K3),size(K3,1),1);
K4 = K4./repmat(max(K4),size(K4,1),1);

%% Modelo Matematico-Geometrico

M = {K1, K2, K3, K4};

Ne = length(M);

switch Ne

    case 2

        x = [0 2^.5];
        y = [0 0];

    case 3
```

```
        x = [0 2^.5/2 2.^.5 0];
        y = [2^.5 0 2^.5 2.^.5];

    case 4

        x = [0 (2^.5)/2 2.^.5 (2^.5)/2 0];
        y = [2^.5 0 2^.5 2*2^.5 2.^.5];

end

x = x - (2^.5)/2;
y = y - 2^.5;
y = (y*sqrt(2)/2)/max(abs(y));

disptext = {'*','LLE','CMDS','LE','RBF'};

figure(1)
subplot(232)
cla;
plot(x,y,'-o','LineWidth',4);
axis([min(x), max(x), min(y), max(y)])
hold on
plot([0, 0],[min(y) max(y)],'k','LineWidth',2)
plot([min(x), max(x)],[0 0],'k','LineWidth',2)
set(gca,'xtick',[])
set(gca,'xticklabel',[])
set(gca,'ytick',[])
set(gca,'yticklabel',[])
grid on

p1 = [0 2^.5/2];
p2 = [0 -2^.5/2];
p3 = [2^.5/2 0];
p4 = [-2^.5/2 0];
text(p1(1)-0.1,p1(2)+0.15,disptext{2},'FontSize',20);
text(p2(1)-0.2,p2(2)-0.2,disptext{3},'FontSize',20);
text(p3(1)+0.04,p3(2)+0.02,disptext{4},'FontSize',20);
text(p4(1)-0.38,p4(2)+0.018,disptext{5},'FontSize',20);

Mixture = {'Mixture','LLE','LE','CMDS','RBF'};
Color1 = {'r*','b+','go','mv','cd'};

i = 1;
[x,y] = ginput(1);
text(x,y,disptext{1},'FontSize',20);

% Calculo de la distancia, radios de los circulos.
r1 = norm([x-p1(1) y-p1(2)]);
r2 = norm([x-p2(1) y-p2(2)]);
r3 = norm([x-p3(1) y-p3(2)]);
r4 = norm([x-p4(1) y-p4(2)]);
```

```
% Calculo de la area.
A1 = pi*r1^2;
A2 = pi*r2^2;
A3 = pi*r3^2;
A4 = pi*r4^2;

% Areas normalizadas.
At = A1+A2+A3+A4;
A1 = A1/At;
A2 = A2/At;
A3 = A3/At;
A4 = A4/At;

% Pesos

W1 = 1 - A1;
W2 = 1 - A2;
W3 = 1 - A3;
W4 = 1 - A4;

weights = [W1 W2 W3 W4];
weights = 1 - sinc(weights.^4);

Ksum = 0;
for j = 1:length(M)
    Ksum = Ksum + weights(j)*M{j};
end
[Xsum,~] = eigs(Ksum);
colmap = jet(64);

subplot(133)
scatter(Xsum(:,1),Xsum(:,2),20,L);
set(gca,'xtick',[])
set(gca,'xticklabel',[])
set(gca,'ytick',[])
set(gca,'yticklabel',[])


%% Curva Rnx - Medida de calidad

kop = 1;
id_meth = i;
Ya{1,id_meth} = Xsum./repmat(max(abs(Xsum)),size(Xsum,1),1);
Ya{2,id_meth} = Color1{i};
Ya{3,id_meth} = Mixture{i};

subplot(235)
cla;
nx_scores([-nbr,knn,kop],'r',Y,Ya);
set(gca,'Fontname','Times','Fontsize',20);
set(gca,'Fontname','Times','Fontsize',20);
```

ANEXO 3. ARTICULO DE CONFERENCIA INTERNACIONAL

Este anexo contiene el artículo con el cual fuimos seleccionados entre los 25 mejores temas de investigación en la categoría D y por consiguiente fuimos seleccionados para sustentación en modalidad ponente, cabe resaltar que el artículo fue seleccionado entre los 10 mejores temas de investigación y por lo tanto se recibió una invitación a publicar una versión extendida de este en la revista Ingeniería y Universidad de la Pontificia Universidad Javeriana.

Interactive interface for efficient data visualization via a geometric approach

J. A. Salazar-Castro, Y. C. Rosas-Narváez, A. D. Pantoja

Facutly of Engineering
Universidad de Nariño
alejo26st@udenar.edu.co
ad_pantoja@udenar.edu.co

Juan C. Alvarado-Pérez

Facutly of Engineering
Universidad de Salamanca
jcalvarado@usal.es

Diego H. Peluffo-Ordóñez

Facutly of Engineering
Universidad Cooperativa de Colombia sede Pasto
diego.peluffo@campusucc.edu.co

Abstract

Dimensionality reduction (DR) methods represent a suitable alternative to visualizing data. Nonetheless, most of them still lack the properties of interactivity and controllability. In this work, we propose a data visualization interface that allows for user interaction within an interactive framework. Specifically, our interface is based on a mathematic geometric model, which combines DR methods through a weighted sum. Interactivity is provided in the sense that weighting factors are given by the user via the selection of points inside a geometric surface. Then, (even non-expert) users can intuitively either select a concrete DR method or carry out a mixture of methods. Experimental results are obtained using artificial and real datasets, demonstrating the usability and applicability of our interface in DR-based data visualization.

1. Introduction

Dimensionality reduction (DR) is a key stage for designing pattern recognition and data mining systems when dealing with high-dimensional data sets [1]. The aim of DR methods is to extract lower dimensional, relevant information (called embedded data) from high-dimensional input data, so that both the performance of a pattern recognition system can be improved and data representation becomes more intelligible [2]. Since DR methods are often developed under determined design parameters and pre-established optimization criteria, they still lack the properties of user interaction and controllability, which are characteristic of information visualization procedures [3]. The field of information visualization (Info Vis) is aimed at developing graphical ways of representing data so that information can be more usable and intelligible for the user. Then, one can intuit that DR can be improved by importing some properties of Info Vis methods. This is in fact the premise on which this research is based [4].

This paper presents an attempt to link the field of dimensionality reduction with that of information visualization, in order to harness the special properties of the latter within DR frameworks. In particular, the properties of controllability and interactivity are of interest, which should make the DR outcomes significantly more understandable and

tractable for the (no-necessarily-expert) user [5]. These two properties allow the user to have freedom to select the best way for representing data. Specifically, we propose a geometrical strategy to set the weighting factors for linearly combining DR methods. This is done from kernel approximations [6, 7] of conventional methods (Classical Multidimensional Scaling - CMDS [3], Laplacian Eigenmaps – LE, and Locally Linear Embedding - LLE), which are combined to reach a mixture of kernels. To involve the user in the selection of a method, we use a polygonal approach so the points inside the polygon surface defines the degree or level that a kernel is used, that is, the set of weighting factors. Such polygon has as many edges as the number of considered kernels. This approach allows to evaluating visually the behavior of the embedding data regarding the kernel mixture.

For experiments, we use publicly available databases from the UCI Machine Learning Repository [8] as well as a subset of images from Columbia University Image Library [9]. To assess the performance of the kernel mixture, we consider conventional methods of spectral dimensionality reduction such as multidimensional scaling, locally linear embedding and laplacian eigenmaps [10]. The quality of obtained embedded data is quantified by a scaled version of the average agreement rate between *K*-ary neighborhoods [19]. Provided mixture represents every single dimensionality reduction approach as well as it helps users to find a suitable representation of embedded data within a visual and intuitive framework.

The remaining of the paper is organized as follows: In section 2, data visualization using DR is outlined. Section 3 introduces a novel mathematical geometric model based on a polygonal approach aimed at performing customized DR tasks. Experimental setup and results are shown in Sections 4 and 5, respectively. Finally, Section 6 gathers some final remarks as conclusions and future work.

2. Data visualization

An intuitive way of visualizing numerical data is via a 2D or 3D scatter plot, which is a natural and intelligible visualization fashion for human beings. Therefore, it entails that the initial data should be represented into a lower-dimensional space. In this sense, dimensionality reduction takes places, being an important stage within both the pattern recognition and data visualization systems. Correspondingly, DR is aiming at reaching a low-dimensional data representation, upon which both the classification task performance is improved in terms of accuracy, as well as the intrinsic nature of data is properly represented [1]. So, a more realistic and intelligible visualization for the user is obtained [2]. In other words, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$, such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq i\leq N}$, being $\boldsymbol{x}_i \in \mathbb{R}^d$, where $d < D$. Figure 1 depicts an instance where a manifold, so-called Swiss roll, is embedded into a 2D representation, which resembles to an unfolded version of the original manifold.

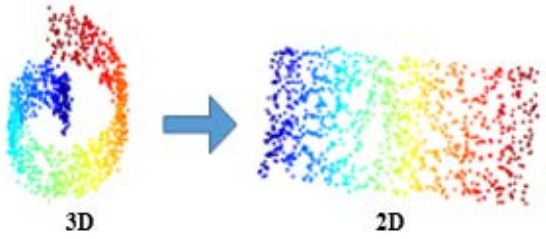


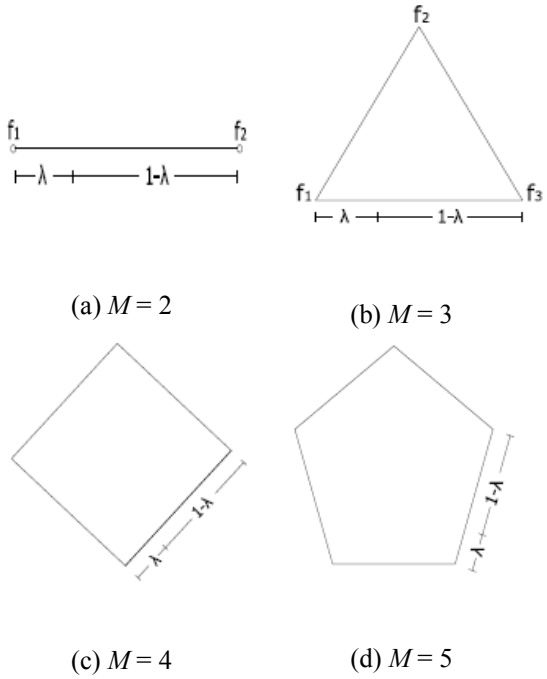
Figure 1. Dimensionality reduction effect over a Swiss roll manifold. Resultant embedded data is an attempt to unfolding the original data.

Classical DR approaches aims to preserve variance (principal component analysis - PCA) or distance (classical multidimensional scaling - CMDS) [3]. Nowadays, more developed, recent methods are aiming at preserving the data topology. Such a topology can be represented by a data-driven graph, built as a non-directed and weighted one, in which data points represent the nodes, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. This representation is exploited by both spectral and divergence-based methods. On one hand, for spectral approaches, similarity matrix can represent the weighting factor for pairwise distances as happens in Laplacian eigenmaps (LE) [10]. As well, using a non-symmetric similarity matrix and focusing on data local structure, the Locally Linear Embedding (LLE) method arose [18]. On the other hand, once normalized, similarity matrix can also represent probability distributions, as do the methods based on divergences such as stochastic neighbor embedding [11].

3. Mathematical geometric approach

In this Section, we introduce a novel method for interactive data visualization within a white box environment, which consists of the combination of

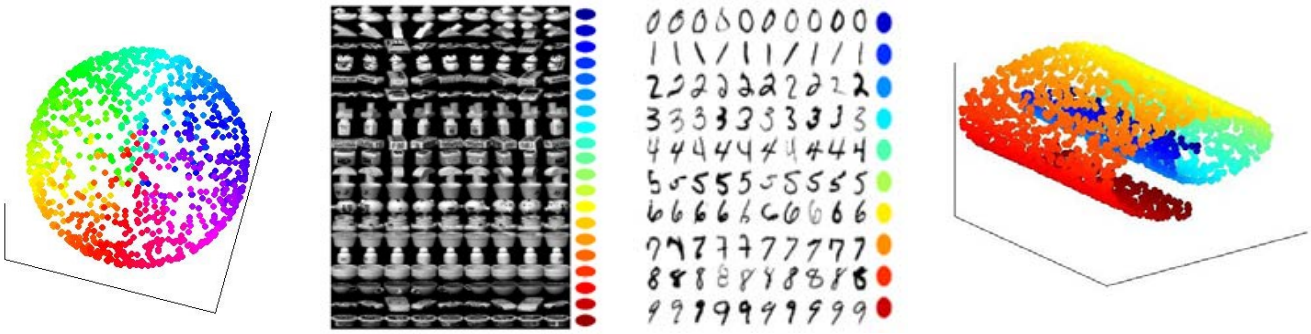
different, spectral unsupervised DR methods. For the sake of versatility and since spectral methods are susceptible to be represented by kernels, combination is carried out taking into account the corresponding kernel matrices. Our method is based on a mathematical geometric approach that allows for performing the mixture of DR methods in an interactive fashion, so that kernel matrices are linearly combined and respective coefficients are related to geometric coordinates inside a geometric surface. So, users –even non-expert ones – might easily and intuitively select a single method or combine methods fulfilling their needs by just de exploring the geometric shape and picking up points from the surface thereof. Figure 2 shows graphically a possible mathematic-geometric model regarding a polygonal approach. In general, any set of methods can be represented by a collection of functions $\{f_1, .., f_M\}$, where M is the number of considered functions. A fashion to perform a pair-wise mixture is the continuous deformation of one function onto another using homotopy basic [12, 13]. Then, a simple model of homotopy can be written as $h(f_1, f_2, \lambda) = \lambda f_1 + (1-\lambda) f_2$, where λ is the homotopy parameter. In terms of an interactive interface, such a parameter would become a sliding bar. So far, this approach can be graphically represented by a line with length 1 drawn between two points representing the two functions, as seen in Figure 2(a).



(a) $M = 2$ (b) $M = 3$

(c) $M = 4$ (d) $M = 5$

Figure 2. Polygonal approach to perform the mixture of a set of functions $\{f_1, \ldots, f_M\}$. Parameter λ controls the pairwise mixture.



(a) *sphere* (b) *COIL-20* (c) *MNIST* (d) *Swiss roll*

Figure 4. The four considered datasets.

Notwithstanding, this model can be naturally extended to more than two methods within a polygonal framework, in such a way that three functions are to be represented with a triangle (Figure 2(b)), four functions with a rhombus (Figure 2(c)), and so forth.

For data visualization purposes through DR methods, the terms to be combined are the kernel matrices corresponding to the considered DR methods. Therefore, we obtain a resultant kernel matrix $\widehat{\boldsymbol{K}}$ as the mixture of M kernel matrices $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}\}$ so:

$$\widehat{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(m)}, \qquad (1)$$

where α_m is the coefficient or weighting factor corresponding to method m and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]$ is weighting vector. As mentioned above, these

coefficients are to be associated with geometric coordinates of points inside the surface.

In this work, the relationship between the points inside the surface and the coefficients of linear combination is given by the distance from every single vertex (representing considering methods) to the selected point, as can be appreciated in Figure 3. Then, these distances as the ratios $\{r_1, \dots, r_M\}$ of circles with areas $\{A_1, \dots, A_M\}$ centered at each vertex. Therefore, it is evident that the area of the m-th circle is $A_m = \pi r_m^2$.

Then, once such areas are normalized to sum to 1, the complement value of them becomes a proper estimation of the weighting factors. In this connection, the values of $\boldsymbol{\alpha}$ are given by:

$$\alpha_m = \operatorname{sinc}\left(1 - \frac{A_m}{\sum_{m=1}^{M} A_m}\right). \qquad (2)$$

Additionally, in order to assigning a higher value to that weighting factor corresponding to the vertex closer to the selected point, function $\operatorname{sinc}(\cdot)$ is used, which also adds an equalization effect over the values.

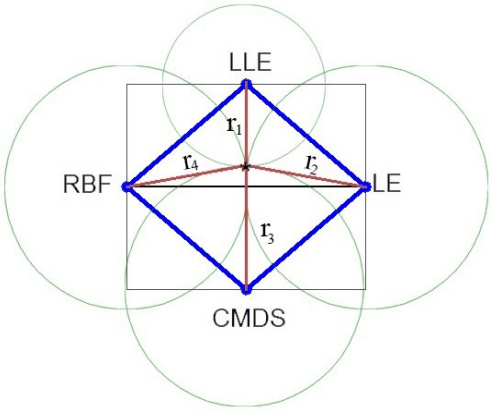


Figure 3. Graphical explanation of the way to estimate weighting factors.

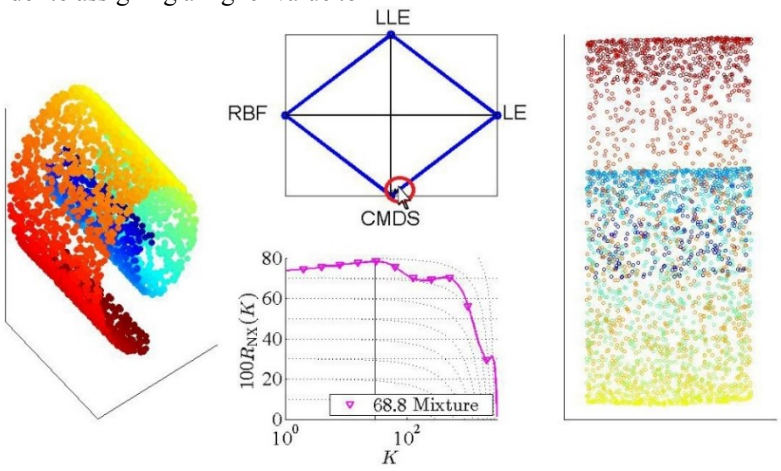


Figure 5. A view of the proposed interface. This is an example for Swiss roll dataset. As the position indicator or cursor is moved around and by clicking on a point inside the surface, a new embedding space (right hand) is reached. Its corresponding $R_{NX}(K)$ curve is also shown.

4. Experimental setup

Experiments are carried out over three conventional data sets. The first data set is an artificial spherical shell (N = 1500 data points and D = 3). The second data set is the COIL-20 image bank [14], which contains 72 gray-level images representing 20 different objects (N = 1440 data points –20 objects in 72 poses/angles–with D = 1282). The third data set is a randomly selected subset of the MNIST image bank [15], which is formed by 6000 gray-level images of each of the 10 digits (N = 1500 data points –150 instances for all 10 digits– and D = 242). The fourth data set is a toy set here called Swiss roll (N = 3000 data points and D = 3). Figure 4 depicts examples of the considered data sets.

Three kernel approximations for spectral DR methods [6] are considered. Namely, Classical Multidimensional Scalling (CMDS), locally linear embedding (LLE), and graph Laplacian Eigenmaps (LE). CMDS kernel is the double centered distance matrix D ∈ RN×N so:

$$\boldsymbol{K}^{(1)} = \boldsymbol{K}_{\text{CMDS}} = -\frac{1}{2}\left(\boldsymbol{I}_N - \boldsymbol{1}_N \boldsymbol{1}_N{}^T\right)\boldsymbol{D}\left(\boldsymbol{I}_N - \boldsymbol{1}_N \boldsymbol{1}_N{}^T\right), \quad (3)$$

where the ij entry of $\boldsymbol{D}$ is given by $d_{ij} = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2$ and $\|\cdot\|^2$ stands for Euclidean norm. A kernel for LLE can be approximated from a quadratic form in terms of the matrix $\boldsymbol{W}$ holding linear coefficients that sum to 1 and optimally reconstruct observed data Define a matrix $\boldsymbol{M} \in R^{N\times N}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{W})(\boldsymbol{I}_N - \boldsymbol{W}^T)$ and λ_{max} as the largest eigenvalue of M. Kernel matrix for LLE is in the form:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \quad (4)$$

Since kernel PCA is a maximization problem of the covariance of the high dimensional data represented by a kernel, LE can be expressed as the pseudo-inverse of the graph Laplacian L:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LE} = \boldsymbol{L}^T, \quad (5)$$

where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, $\boldsymbol{S}$ is a similarity matrix and $\boldsymbol{D} = \text{Diag}(\boldsymbol{S}\boldsymbol{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [6]. The similarity matrix S is formed in such a way that the relative bandwidth parameter is estimated keeping the entropy over neighbor distribution as roughly $\log(K)$ where K is the given number of neighbors as explained in [16]. The number of neighbors is established as $K = 30$.

As well, a RBF kernel is also considered: $\boldsymbol{K}^{(4)} = \boldsymbol{K}_{RBF}$ whose ij entries are given by $exp(-0.5||yi - yj||/\sigma^2)$ with σ = 0.1. For all methods, input data is embedded into a 2-dimensional space ($d = 2$).

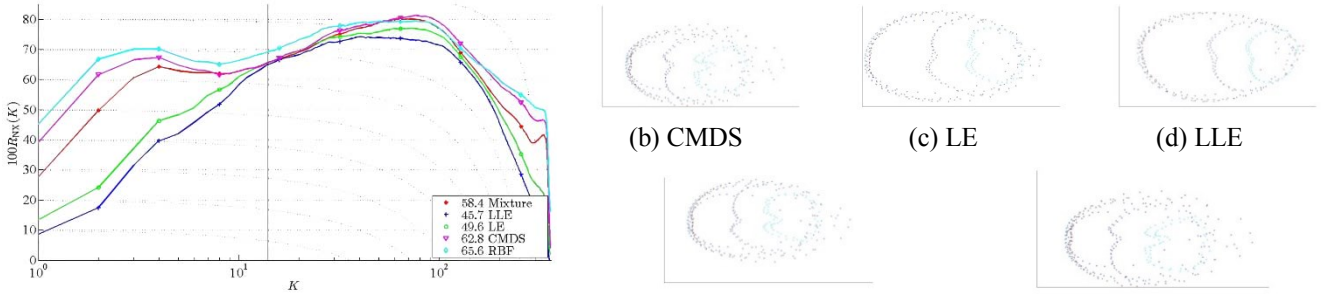
Accordingly, our approach is performed considering $M = 4$ kernels. The resultant kernel provided $\widehat{\boldsymbol{K}}$ here as well as the individual kernels $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}\}$ are tested by obtaining embedded data from kernel PCA, as explained in [17].

To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [19] is used, which is ranged within the interval [0, 1]. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N-1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.

5. Results and discussion

Following are presented some experimental results aimed at testing all the considered datasets regarding the embedded data using $R_{NX}(K)$ as a quality indicator. Resultant kernel matrices feed a kernel PCA algorithm to output 2D-dimensional data spaces. Since all the experiments are carried out considering four methods, the polygonal surface is then a rhombus as shown in Figure 5.

Given that the mixture presented here is a linear combination, when coefficients are selected from the perimeter only two kernels are considered. Then, user can appreciate the deformation of the resulting embedding from a method onto that from another method by moving from one vertex to another on the respective edge.



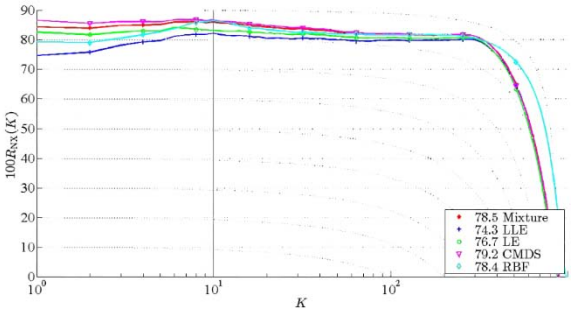
(a) $R_{NX}(K)$ for all considered methods.

(b) CMDS (c) LE (d) LLE

(e) Mixture of Kernels (f) RBF

Figure 6. Results for COIL dataset. Results are shown regarding the quality measure $RNX(K)$. The curves and their AUC (a) for all considered methods are depicted, as well as the embedding data (b)-(f). Individual

embedding data spaces are obtained by selecting the points rightly on the vertexes meanwhile the mixture is done with the coefficients associated to the central point.



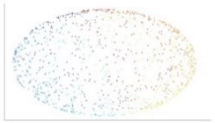
(a) $R_{NX}(K)$ for all considered methods.

(b) CMDS

(c) LE

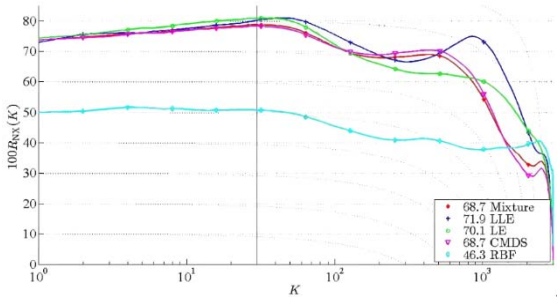
(d) LLE

(e) Mixture of Kernels

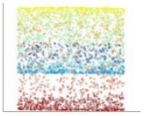


(f) RBF

Figure 7. Results for the spherical shell dataset.

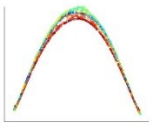


(a) $R_{NX}(K)$ for all considered methods.

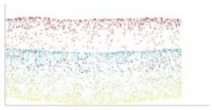


(b) CMDS

(c) LE



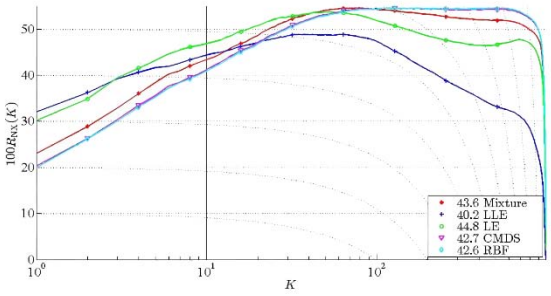
(d) LLE



(e) Mixture of Kernels

(f) RBF

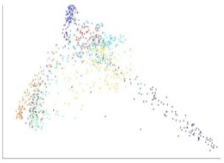
Figure 8. Results for the Swiss roll dataset.



(a) $R_{NX}(K)$ for all considered methods.

(b) CMDS

(c) LE



(d) LLE

(e) Mixture of Kernels

(f) RBF

Figure 9. Results for MNIST dataset.

Indeed, selecting coefficients associated with the vertexes, the effect of a single method is performed. In addition, when selecting inner points the effects of every method is taken into account to calculate the resultant kernel. Therefore, the proposed approach enable users (even those not expert) to interact with the DR outcomes by intuitively selecting points from a polygonal surface. Overall obtained results are shown in Figures 6 to 9.

6. Conclusion and future work

The proposed interface represents an interactive approach to visualize the embedded data resulting from dimensionality reduction (DR) methods. This approach is based on a geometric perspective of homotopy allowing for dealing with more than two

functions. In this case, kernel matrices representing spectral, unsupervised methods of DR. In particular, our approach relates the inner points location of a polygonal surface with the values of weighting factors used to carry out the linear combination of kernel matrices. Given this graphic and intuitive framework, even non-expert users might easily select a method or combination of methods by picking up points from a polygonal surface fulfill their specific needs.

As a future work, more developed and interactive models are to be explored. As well, ways to optimize and speed up algorithm routines are to be studied and developed.

Acknowledgments

This work is supported by the *"Grupo de Investigación en Ingeniería Eléctrica y Electrónica – GIIEE"* from Universidad de Nariño, as well as ESLINGA Research Group from Universidad Cooperativa de Colombia, sede Pasto.

As well, authors acknowledgment the research project "*Diseño e implementación de un prototipo electrónico de bajo costo para terapias de biofeedback en tratamientos de trastornos psicofisiológicos*" funded by Fundación CEIBA and Gobernación de Nariño.

References

[1] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery" Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. ACM, 2009.

[2] D. H. Peluffo-Ordónez, J. A. Lee and M. Verleysen. "Short review of dimensionality reduction methods based on stochastic neighbor embedding". Advances in Self-Organizing Maps and Learning Vector Quantization. Springer International Publishing, 2014. 65-74.

[3] I. Borg and J. Patrick. Modern multidimensional scaling: Theory and applications. Springer Science & Business Media, 2005.

[4] W. Dai, and P. Hu. "Research on Personalized Behaviors Recommendation System Based on Cloud Computing." TELKOMNIKA Indonesian Journal of Electrical Engineering 12.2 (2013): 1480-1486.

[5] M. Ward, G. Grinstein, and D. Keim, Interactive data visualization: foundations, techniques, and applications. AK Peters, Ltd., 2010.

[6] J. Ham, D. D. Lee, S. Mika, and B. Scholkopf. "A kernel view of the dimensionality reduction of manifolds." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004, p-47.

[7] D. H. Peluffo-Ordónez, J. A. Lee, and M. Verleysen, "Generalized kernel framework for unsupervised spectral methods of dimensionality reduction," in Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, Dec 2014, pp. 171–177.

[8] M. Lichman, "UCI Machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[9] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Dept. Comput. Sci., Columbia Univ., New York. [Online] http://www.cs.Columbia.edu/CAVE/coil-20.html, vol. 62, 1996.

[10] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", Neural computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[11] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In Advances in neural information processing systems, pages 833–840, 2002.

[12] J. Harmann, M. P. Murphy, C. S. Peters, and P. C. Staecker, "Homotopy equivalence in graph-like digital topological spaces," arXiv preprint arXiv: 1408.2584, 2014.

[13] D. H. Peluffo-Ordóñez, J. C. Alvarado-Pérez, J. A. Lee and M. Verleysen. Geometrical Homotopy for data visualization. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

[14] S. A. Nene, S. K. Nayar, H. Murase: Columbia object image library (coil-20). Dept. Compute. Sci., Columbia Univ., New York, 62 (1996), http://www.cs.columbia.edu/CAVE/coil-20.htm.l

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11) (1998) 22782324.

[16] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton: Visualizing similarity data with a mixture of maps. In: International Conference on Artificial Intelligence and Statistics. (2007) 67–74

[17] D. Peluffo-Ordóñez, J. Lee, and M. Verleysen: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: IEEE Symposium Series on Computational Intelligence. (2014).

[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, pp. 2323–2326. (2000).

[19] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing. (2013).

ANEXO 4. ARTICULO PRESENTADO A REVISTA INDEXADA

En esta subsección se muestra el articulo presentado a la revista Ingeniería y Universidad de la Pontificia Universidad Javeriana, resultado de haber sido escogidos dentro de los 10 primeros mejores artículos en la categoría D modalidad ponentes en el *The Twentieth Symposium on Signal Processing, Images and Computer Vision*, STSIVA-2015.

Visualización interactiva de datos usando métodos de reducción de dimensión: Un enfoque matemático-geométrico[1]

Interactive data visualization using dimensionality reduction: A mathematical geometric approach[2]

Jose Salazar-Castro[3]

Yesid Rosas-Narváez[4]

[1] Este artículo se deriva de un proyecto de investigación denominado Implementación de una interfaz de visualización de datos eficiente e interactiva a partir de una perspectiva geométrica, Código o Número de registro (*En caso de tener*). Desarrollado por el grupo de investigación Grupo de Investigación en Ingeniería Eléctrica y Electrónica de la Universidad de Nariño y por el grupo Eslinga de la Universidad Cooperativa de Colombia, Pasto, Colombia.

[2]

[3] Ingeniero Electrónico, Universidad de Nariño. Auxiliar de Investigación, Universidad de Nariño. Pasto, Colombia. Correo electrónico: alejo26st@udenar.edu.co.

[4] Ingeniero Electrónico de la Universidad de Nariño. Auxiliar de Investigación de la Universidad de Nariño. Pasto, Colombia. Correo electrónico: Yesid_ronar@hotmail.com.

Andrés Pantoja[5]

Juan Alvarado-Pérez[6]

Diego Peluffo-Ordoñez[7]

Resumen

Los métodos de reducción de dimensión (RD) representan una alternativa adecuada para visualizar datos. Sin embargo, muchos de ellos aún carecen de propiedades de interactividad y controlabilidad. En este trabajo, se propone una interfaz de visualización de datos que permita la interacción del usuario dentro de un framework interactivo. Específicamente, la interfaz propuesta está basada en un modelo matemático-geométrico, el cual combina métodos de RD a través de una suma ponderada. La interactividad es proporcionada en el sentido que los factores ponderados son dados por el usuario mediante la selección de puntos dentro de una superficie geométrica. Por tanto, los usuarios (incluso aquellos que no son expertos) intuitivamente podrían seleccionar un método RD en particular, o realizar una mezcla de métodos. Los resultados experimentales obtenidos con datos artificiales y reales demuestran la usabilidad y aplicabilidad de la interfaz en visualización de datos basada en RD.

Palabras clave: Interfaz interactiva, reducción de dimensión, visualización de datos.

Abstract

Dimensionality reduction (DR) has become a natural suitable alternative to visualizing high-dimensional data (i.e. representing data in an intelligible way: a 2D or 3D representation). Due to their design conditions, most of DR methods lack the properties of interactivity and controllability. This work introduces a the design of a data visualization interface allowsing for user controllability within an interactive framework. Specifically, our interface is based on a mathematic geometric model, which combines DR methods through a weighted sum. Interactivity is provided in the sense that weighting factors are given by the user via the selection of points inside a geometric surface. Then, (even non-expert) users can intuitively either select a concrete DR method or carry out a mixture of methods. Experimental results are obtained using artificial and real datasets, demonstrating the usability and applicability of our interface in DR-based data visualization.

[5] Ingeniero electrónico, universidad Nacional de Colombia, sede Manizales. Magister en Ingeniería Electrónica y de Computadores, Universidad de los Andes. Doctor en Ingeniería, Universidad de los Andes. Docente tiempo completo y director del grupo de investigación GIIEE, Universidad de Nariño. Pasto, Colombia. Correo electrónico: ad_pantoja@udenar.edu.co.

6 Ingeniero de Sistemas, Universidad de Nariño, Máster en Sistemas Inteligentes, Universidad de Salamanca, Estudiante de doctorado, Universidad de Salamanca. Pasto, Colombia. Correo electrónico. jcalvarado@usal.edu.co.

[7] Ingeniero electrónico, Universidad Nacional de Colombia. Maestría en Ingeniería - Automatización industrial. Doctorado en Ingeniería - Línea de automatización Industrial. Universidad Nacional de Colombia, sede Manizales. Docente tiempo completo y director del grupo de investigación Eslinga, Universidad Cooperativa de Colombia, sede Pasto. Pasto, Colombia. Correo electrónico: diego.peluffo@campusucc.edu.co.

Keywords: Data visualization, dimensional reduction, Interactive interface.

1. Introduction

The field of information visualization (Info Vis) is aimed at developing graphical ways of representing data so that information can be more usable and intelligible for the user. In addition, to designing pattern recognition or data mining systems for dealing with high-dimensional data sets, dimensionality reduction (DR) becomes a determinant stage [1]. DR methods are aiming at extracting lower dimensional, relevant information (called embedded data) from high-dimensional input data, in order that both the performance of a pattern recognition system can be improved and data representation becomes more intelligible [2]. Classical DR approaches aims to preserve variance (principal component analysis - PCA) or distance (classical multidimensional scaling - CMDS) [3]. Nowadays, more developed, recent methods are aiming at preserving the data topology. Such a topology can be represented by a data-driven graph, built as a non-directed and weighted one, in which data points represent the nodes, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. This representation is exploited by both spectral and divergence-based methods. On one hand, for spectral approaches, similarity matrix can represent the weighting factor for pairwise distances as happens in Laplacian eigenmaps (LE) [10]. As well, using a non-symmetric similarity matrix and focusing on data local structure, the Locally Linear Embedding (LLE) method arose [18]. On the other hand, once normalized, similarity matrix can also represent probability distributions, as do the methods based on divergences such as stochastic neighbor embedding [11].

Nonetheless, given that DR methods are often developed under determined design parameters and pre-established optimization criteria, they still lack the properties of user interaction and controllability, which are characteristic of Info Vis procedures [3]. Thus, we can intuit that DR can be improved by importing some properties of Info Vis methods. This is in fact the premise on which this research is based [4].

In this work, we present an initial approach to link the field of dimensionality reduction with that of information visualization, in order to harness the special properties of the latter within DR frameworks. Particularly, the DR outcomes significantly make data more understandable and tractable for the (no-necessarily-expert) user due to the properties of controllability and interactivity, for this reason, both properties are of interest [5]. These two properties allow the user to have freedom to select the best way for representing data. Specifically, a geometrical strategy is proposed to set the weighting factors for linearly combining DR methods. To do so, we take advantages of kernel approximations [6, 7] of conventional methods (CMDS, LLE, and LLE), which are combined to reach a mixture of kernels. We use a polygonal approach to involve the user in the selection of a method, so that the degree or level of a kernel usage (the set of weighting factors) is defined for the point inside the polygon surface. The edges of such polygon depends of the number of considered kernels. This approach allows to evaluating visually the behavior of the embedding data regarding the kernel mixture.

For experiments, publicly available databases from the UCI Machine Learning Repository are used [8], as well as a subset of images from Columbia University Image Library [9]. We

consider conventional methods of spectral dimensionality reduction to assess the performance of the kernel mixture [10]. Particularly, we use Kernel PCA to generate the embedded data related to the kernel mixture. We use a scaled version of the average agreement rate between K-ary neighborhoods to quantifying the quality of obtained embedded data [19]. Provided mixture represents every single dimensionality reduction approach as well as it helps users to find a suitable representation of embedded data within a visual and intuitive framework.
The remaining of the paper is organized as follows: In section 2, data visualization using kernel PCA is outlined. Section 3 introduces a novel mathematical geometric model based on a polygonal approach aimed at performing customized DR tasks. Experimental setup and results are shown in Sections 4 and 5, respectively. Finally, Section 6 gathers conclusions and future work as some final remarks.

2. Data visualization using Kernel PCA

In formal terms, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq i\leq N}$, such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x_i}]_{1\leq i\leq N}$, being $\boldsymbol{x}_i \in \mathbb{R}^d$, where $d < D$. As well, from another point of view, observed data matrix is conformed by D variables such that $\boldsymbol{y}^{(l)} \in \mathbb{R}^N$ with $l \in \{1,\ldots,D\}$, meanwhile latent data matrix by d variables denoted as $\boldsymbol{x}^{(\ell)} \in \mathbb{R}^N$ with $\ell \in \{1,\ldots,d\}$.
Suppose that there exists an unknown high dimensional representation space $\boldsymbol{\Phi} \in \mathbb{R}^{\boldsymbol{D}_h \times N}$ such that $\boldsymbol{D}_h \gg D$, in wich calculating the inner product should improve the representation and visualization of resultant embedded data in contrast to that obtained directly from the observed data. Hence, the need of a kernel representation arises to calculate the dot product in the unknown high dimensional space. Let $\Phi(\cdot)$ be a function that maps data from the original dimension to a higher one, such that:

$$\begin{aligned} \Phi(\cdot):\mathbb{R}^D &\rightarrow \mathbb{R}^{\boldsymbol{D}_h} \\ \boldsymbol{y}_i &\mapsto \Phi(\boldsymbol{y}_i). \end{aligned}$$

Therefore, the i-th column vector of matrix $\boldsymbol{\Phi}$ is given by $\boldsymbol{\Phi}_i = \Phi(\boldsymbol{y}_i)$. By the Mercer's condition or kernel trick takes place, a kernel function $k(\cdot,\cdot)$ allows for estimating the dot product $\Phi(\boldsymbol{y}_i)^{\mathrm{T}}\, \Phi(\boldsymbol{y}_j) = k(\Phi(\boldsymbol{y}_i), \Phi(\boldsymbol{y}_j))$. Arranging all the possible dot products in a matrix $\mathbf{K} = [k_{ij}]$, we get a kernel matrix:

$$\mathbf{K} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}, \tag{1}$$

where $k_{ij} = k(\boldsymbol{y}_i, \boldsymbol{y}_j)$.

The formulation of Kernel PCA is done under the assumption that $\boldsymbol{\Phi}$ is centered. This condition can be ensured by algebraically modifying the calculation of the dot product as it will be explained further. To project data, a linear combination with a d-dimensional base is used. Such a base can be arranged in an orthonormal rotation matrix $\mathbf{W} \in \mathbb{R}^{\boldsymbol{D}_h \times N}$, such that $\mathbf{W} = [\mathrm{w}^{(1)},\ldots, \mathrm{w}^{(d)}]$ and $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}_d$, where $\mathbf{w}^{(\ell)} \in \mathbb{R}^{\boldsymbol{D}_h}$ and $\mathbf{I}_d$ is a d-dimensional identity matrix. Then, projected data matrix $\mathbf{X} \in \mathbb{R}^{\boldsymbol{D}_h \times N}$ can be calculated as:

$$\mathbf{K} = \mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}. \tag{2}$$

Generally, the projection is performed over a lower dimensional space, which means that data are projected with a low-rank representation of rotation matrix ($\mathrm{d} < \mathrm{D}$). Nonetheless, data can be fully projected by using a whole base setting $\mathrm{d} = \mathrm{D}$. Furthermore, from equation (2) a lower-rank data matrix $\widehat{\boldsymbol{\Phi}} \in \mathbb{R}^{D_h \times N}$ can be obtained when $\mathrm{d} < \mathrm{D}$ by

$$\widehat{\boldsymbol{\Phi}} = \mathbf{W}\mathbf{X}. \tag{3}$$

Then, we can also write that $\widehat{\boldsymbol{\Phi}} = \boldsymbol{W}\boldsymbol{W}^{T}\boldsymbol{\Phi}$. The variance criterion can be expressed as $E_{\boldsymbol{\Phi}}\{\|\boldsymbol{\Phi}_i - \boldsymbol{W}\boldsymbol{W}^{T}\boldsymbol{\Phi}_i\|_2^2\}$ where $\|\cdot\|_2$and $E_{\boldsymbol{\Phi}}\{\cdot\}$denotes Euclidean norm, and expected value operator regarding $\boldsymbol{\Phi}$, respectively. Considering $E_{\boldsymbol{\Phi}}$ as the simple average, the mean-square-error-based objective function can be written as:

$$\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\|\boldsymbol{\Phi}_{\mathrm{i}} - \mathbf{W}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}_{\mathbf{i}}\|_2^2 = \frac{1}{\mathrm{N}}\left\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}\right\|_{\mathrm{F}}^2, \tag{4}$$

where $\|\cdot\|_F$ stands for Frobenius norm. Following we explain how to solve the optimization problem and calculate the embedded space.

A feasible optimal solution of the problem

$$\begin{gathered}\min\left\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}\right\|_{\mathrm{F}}^2 \\ \mathrm{s.t}\,\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}_{\mathrm{d}}, \mathrm{d} < \mathrm{D} \\ \mathbf{X} = \mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi},\end{gathered}$$

is selecting $\mathbf{W}$ and $\mathbf{X}$ as the eigenvectors associated to the d largest eigenvalues of $\boldsymbol{\Phi}\boldsymbol{\Phi}^{T}$ and the kernel matrix $\mathbf{K} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$, respectively.

The objective function can be extended as:

$$\left\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}\right\|_{\mathrm{F}}^2 = \mathrm{tr}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}) - 2\mathrm{tr}\left(\widehat{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{\Phi}\right) + \mathrm{tr}\left(\widehat{\boldsymbol{\Phi}}^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}\right). \tag{5}$$

Since term $\mathrm{tr}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}) = \left\|\widehat{\boldsymbol{\Phi}}\right\|_F^2$ is constant and $\mathrm{tr}\left(\widehat{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{\Phi}\right) = \mathrm{tr}\left(\widehat{\boldsymbol{\Phi}}^{\mathrm{T}}\widehat{\boldsymbol{\Phi}}\right)$, the following duality takes place:

$$\left\|\widehat{\boldsymbol{\Phi}}\right\|_F^2 = \mathrm{tr}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}) + \left\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}\right\|_F^2,$$

where the problem of minimizing $\left\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}\right\|_F^2$ can be expressed as maximizing its complement $\mathrm{tr}\left(\widehat{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{\Phi}\right)$. In addition, recalling equation (3), we have that

$$\mathrm{tr}\left(\widehat{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{\Phi}\right) = \mathrm{tr}(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}) = \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}),$$

and, thus the new optimization problem is:

$$\begin{aligned} &\max \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}) \\ &\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}_{\mathrm{d}}. \end{aligned} \tag{6}$$

To solve the previous problem, we can write a Lagragian in the form:

$$\mathcal{L}(\mathbf{W}|\boldsymbol{\Phi}) = \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}) - \mathrm{tr}(\boldsymbol{\Lambda}(\mathbf{W}^{\mathrm{T}}\mathbf{W} - \mathbf{I}_{\mathrm{d}})),$$

where $\boldsymbol{\Lambda} = Diag(\lambda_1, \dots, \lambda_{D_h})$ are the Lagrange multipliers. By solving the first order condition on the Lagragian, we get the following dual problem:

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda} \Rightarrow \mathbf{W}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W} = \boldsymbol{\Lambda}. \tag{7}$$

Therefore, a feasible solution is when $\boldsymbol{\Lambda}$ and $\mathbf{W}$ are the eigen- value and eigenvector matrix, respectively. Furthermore, since this is a maximization problem, the eigenvectors associated to the *d* largest eigenvalues must be selected. Similarly, pre-multiplying equation (7) by $\boldsymbol{\Phi}^{T}$, we get:

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{W}\boldsymbol{\Lambda} \Rightarrow \mathbf{K}\mathbf{X}^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Lambda}, \tag{8}$$

and therefore embedded space $\mathbf{X}$ can be calculated as the eigenvectors of kernel matrix $\mathbf{K}$. This approach is widely explained in [17].

3. Mathematical geometric approach

In this Section, we introduce a novel method for interactive data visualization within a white box environment, which consists of the combination of different, spectral unsupervised DR methods. For the sake of versatility and since spectral methods are susceptible to be represented by kernels, combination is carried out taking into account the corresponding kernel matrices. Our method is based on a mathematical geometric approach that allows for performing the mixture of DR methods in an interactive fashion, so that kernel matrices are linearly combined and respective coefficients are related to geometric coordinates inside a geometric surface. So, users –even non-expert ones – might easily and intuitively select a single method or combine methods fulfilling their needs by just de exploring the geometric shape and picking up points from the surface thereof. Figure 2 shows graphically a possible mathematic-geometric model regarding a polygonal approach. In general, any set of methods can be represented by a collection of functions$\{f_1, \dots, f_M\}$, where M is the number of considered functions. A fashion to perform a pair-wise mixture is the continuous deformation of one function onto another using homotopy basic [12, 13]. Then, a simple model of homotopy can be written as $h\,(f_1, f_2, \lambda) = \lambda\, f_1 + (1 - \lambda)\, f_2$, where λ is the homotopy parameter. In terms of an interactive interface, such a parameter would become a sliding bar. So far, this approach can be graphically represented by a line with length 1 drawn between

two points representing the two functions, as seen in Figure 1(a). Notwithstanding, this model can be naturally extended to more than two methods within a polygonal framework, in such a way that three functions are to be represented with a triangle (Figure 1(b)), four functions with a rhombus (Figure 1(c)), and so forth.
For data visualization purposes through DR methods, the terms to be combined are the kernel matrices corresponding to the considered DR methods. Therefore, we obtain a resultant kernel matrix K as the mixture of M kernel matrices $\{\boldsymbol{K}^{(1)}, \dots, \boldsymbol{K}^{(M)}\}$ so:

$$\widehat{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(m)}, \tag{9}$$

where α_m is the coefficient or weighting factor corresponding to method m and $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_M]$ is weighting vector. As mentioned above, these coefficients are to be associated with geometric coordinates of points inside the surface.
In this work, the relationship between the points inside the surface and the coefficients of linear combination is given by the distance from every single vertex (representing considering methods) to the selected point, as can be appreciated in Figure 2. Then, these distances as the ratios $\{r_1, \dots, r_M\}$ of circles with areas $\{A_1, \dots, A_M\}$ centered at each vertex. Therefore, it is evident that the area of the m-th circle is $A_m = \pi r_m^2$. Then, once such areas are normalized to sum to 1, the complement value of them becomes a proper estimation of the weighting factors. In this connection, the values of α are given by:

$$\alpha_m = \operatorname{sinc}\left(1 - \frac{A_m}{\sum_{m=1}^{M} A_m}\right). \tag{10}$$

Additionally, in order to assigning a higher value to that weighting factor corresponding to the vertex closer to the selected point, function sinc(·) is used, which also adds an equalization effect over the values.

4. Experimental setup

Experiments are carried out over three conventional data sets. The first data set is an artificial spherical shell (N = 1500 data points and D = 3). The second data set is the COIL-20 image bank [14], which contains 72 gray-level images representing 20 different objects (N = 1440 data points –20 objects in 72 poses/angles–with D = 1282). The third data set is a randomly selected subset of the MNIST image bank [15], which is formed by 6000 gray-level images of each of the 10 digits (N = 1500 data points –150 instances for all 10 digits– and D = 242). The fourth data set is a toy set here called Swiss roll (N = 3000 data points and D = 3). Figure 3 depicts examples of the considered data sets.

Three kernel approximations for spectral DR methods [6] are considered. Namely, Classical Multidimensional Scalling (CMDS), locally linear embedding (LLE), and graph Laplacian Eigenmaps (LE). CMDS kernel is the double centered distance matrix $\boldsymbol{D} \in R^{N \times N}$ so:

$$\boldsymbol{K}^{(1)} = \boldsymbol{K}_{\text{CMDS}} = -\frac{1}{2}\left(\boldsymbol{I}_N - \mathbf{1}_N \mathbf{1}_N{}^T\right)\boldsymbol{D}\left(\boldsymbol{I}_N - \mathbf{1}_N \mathbf{1}_N{}^T\right), \qquad (11)$$

where the ij entry of D is given by $d_{ij} = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2$ and $||\cdot||^2$ stands for Euclidean norm. A kernel for LLE can be approximated from a quadratic form in terms of the matrix W holding linear coefficients that sum to 1 and optimally reconstruct observed data Define a matrix $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{W})(\boldsymbol{I}_N - \boldsymbol{W}^T)$ and λ_{max} as the largest eigenvalue of M. Kernel matrix for LLE is in the form:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \qquad (12)$$

Since kernel PCA is a maximization problem of the covariance of the high dimensional data represented by a kernel, LE can be expressed as the pseudo-inverse of the graph Laplacian L:

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LE} = \boldsymbol{L}^T, \qquad (12)$$

where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, S is a similarity matrix and $\boldsymbol{D} = \text{Diag}(\boldsymbol{S}\mathbf{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [6]. The similarity matrix S is formed in such a way that the relative bandwidth parameter is estimated keeping the entropy over neighbor distribution as roughly log(K) where K is the given number of neighbors as explained in [16]. The number of neighbors is established as $K = 30$.

As well, a RBF kernel is also considered: $\boldsymbol{K}^{(4)} = \boldsymbol{K}_{RBF}$ whose ij entries are given by $exp(-0.5||yi - yj||/\sigma^2)$ with σ = 0.1. For all methods, input data is embedded into a 2-dimensional space ($d = 2$).

Accordingly, our approach is performed considering $M = 4$ kernels. The resultant kernel provided $\widehat{\boldsymbol{K}}$ here as well as the individual kernels $\{\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}\}$ are tested by obtaining embedded data from kernel PCA, as described in Section 2.2. To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [19] is used, which is ranged within the interval [0, 1]. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N - 1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.

5. Results and discussion

Following are presented some experimental results aimed at testing all the considered datasets regarding the embedded data using $R_{NX}(K)$ as a quality indicator. Resultant kernel matrices feed a kernel PCA algorithm to output 2D-dimensional data spaces. Since all the experiments are carried out considering four methods, the polygonal surface is then a rhombus as shown in Figure 4.

Given that the mixture presented here is a linear combination, when coefficients are selected from the perimeter only two kernels are considered. Then, user can appreciate the deformation of the resulting embedding from a method onto that from another method by moving from one vertex to another on the respective edge. Indeed, selecting coefficients

associated with the vertexes, the effect of a single method is performed. In addition, when selecting inner points the effects of every method is taken into account to calculate the resultant kernel. Therefore, the proposed approach enable users (even those not expert) to interact with the DR outcomes by intuitively selecting points from a polygonal surface. Overall obtained results are shown in Figures 5 to 8.

6. Conclusions and future work

The here proposed interface is based on an interactive approach to visualize the embedded data resulting from dimensionality reduction (DR) methods. Particularly, proposed approach relies on kernels with a geometric perspective of homotopy allowing for dealing with a set of functions. In this case, the representation of spectral, unsupervised methods of DR is given by kernel matrices. Specifically, we use this approach to carry out the linear combination of kernel matrices given by the relation among the inner points location of a polygonal surface with the values of weighting factors. Doing so, we obtain an interactive kernelized version of PCA. Given the graphic and intuitive framework, from our interface an user -irrespective of whether is a non-expert user- might easily select a method or combination of function by picking up points from a polygonal surface fulfilling their specific needs without regard the mixture of methods.

As a future work, more developed and interactive models are to be explored. As well, ways to optimize and speed up algorithm routines are to be studied and developed. In adittion, we will explore more kernel properties to design the best methods for mixing.

References

[1] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery" Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. ACM, 2009.

[2] D. H. Peluffo-Ordónez, J. A. Lee and M. Verleysen. "Short review of dimensionality reduction methods based on stochastic neighbor embedding". Advances in Self-Organizing Maps and Learning Vector Quantization. Springer International Publishing, 2014. 65-74.

[3] I. Borg and J. Patrick. Modern multidimensional scaling: Theory and applications. Springer Science & Business Media, 2005.

[4] W. Dai, and P. Hu. "Research on Personalized Behaviors Recommendation System Based on Cloud Computing." TELKOMNIKA Indonesian Journal of Electrical Engineering 12.2 (2013): 1480-1486.

[5] M. Ward, G. Grinstein, and D. Keim, Interactive data visualization: foundations, techniques, and applications. AK Peters, Ltd., 2010.

[6] J. Ham, D. D. Lee, S. Mika, and B. Scholkopf. "A kernel view of the dimensionality reduction of manifolds." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004, p-47.

[7] D. H. Peluffo-Ordónez, J. A. Lee, and M. Verleysen, "Generalized kernel framework for unsupervised spectral methods of dimensionality reduction," in Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, Dec 2014, pp. 171–177.

[8] M. Lichman, "UCI Machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml.

[9] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-20)," Dept. Comput. Sci., Columbia Univ., New York. [Online] http://www.cs.Columbia.edu/CAVE/coil-20.html, vol. 62, 1996.

[10] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", Neural computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[11] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In Advances in neural information processing systems, pages 833–840, 2002.

[12] J. Harmann, M. P. Murphy, C. S. Peters, and P. C. Staecker, "Homotopy equivalence in graph-like digital topological spaces," arXiv preprint arXiv: 1408.2584, 2014.

[13] D. H. Peluffo-Ordóñez, J. C. Alvarado-Pérez, J. A. Lee and M. Verleysen. Geometrical Homotopy for data visualization. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2015.

[14] S. A. Nene, S. K. Nayar, H. Murase: Columbia object image library (coil-20). Dept. Compute. Sci., Columbia Univ., New York, 62 (1996), http://www.cs.columbia.edu/CAVE/coil-20.htm.l

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11) (1998) 22782324.

[16] J. Cook, I. Sutskever, A. Mnih, and G. E. Hinton: Visualizing similarity data with a mixture of maps. In: International Conference on Artificial Intelligence and Statistics. (2007) 67–74

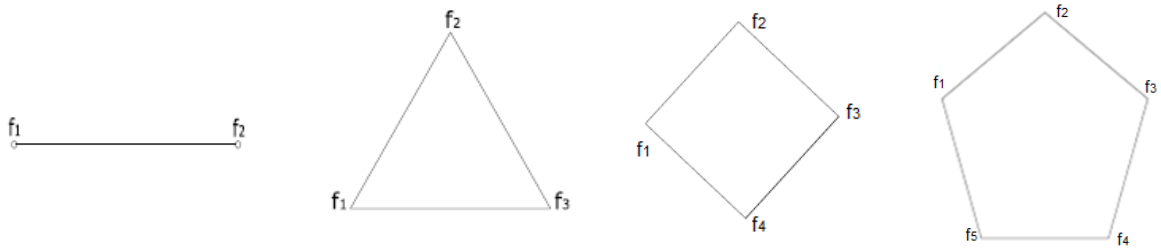
[17] D. Peluffo-Ordóñez, J. Lee, and M. Verleysen: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: IEEE Symposium Series on Computational Intelligence. (2014).

[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, pp. 2323–2326. (2000).

[19] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of kullback-leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing. (2013).

Figures:

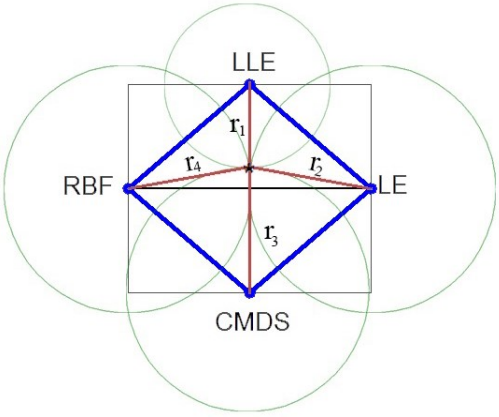
Figure 1. Polygonal approach to perform the mixture of a set of functions $\{f_1, \ldots, f_M\}$. Parameter λ controls the pairwise mixture.



(a) $M = 2$ (b) $M = 3$ (c) $M = 4$ (d) $M = 5$

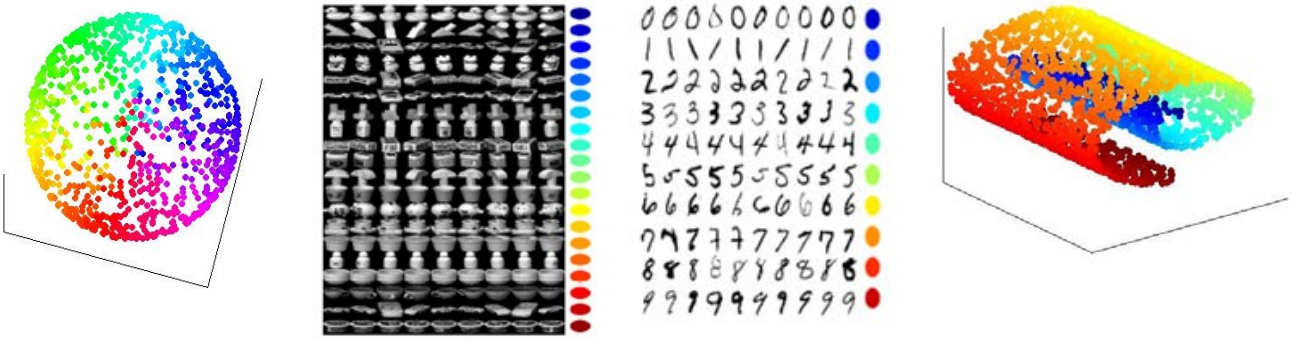
Fuente: presentación propia de los autores.

Figure 2. Graphical explanation of the way to estimate weighting factors.



Fuente: presentación propia de los autores.

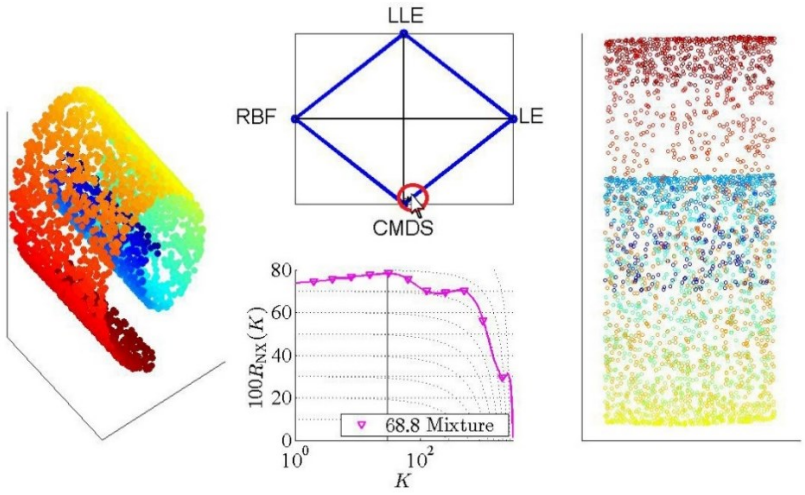
Figure 3. The four considered datasets.



(a) Sphere (b) COIL-20 (c) MNIST (d) Swiss roll

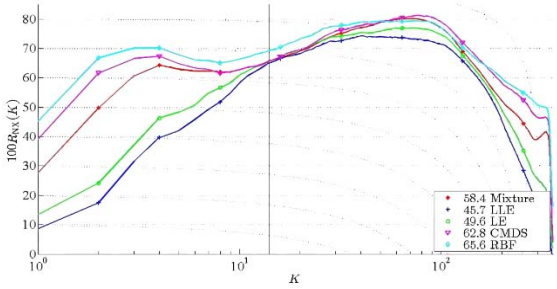
Fuente: presentación propia de los autores.

Figure 4. A view of the proposed interface. This is an example for Swiss roll dataset. As the position indicator or cursor is moved around and by clicking on a point inside the surface, a new embedding space (right hand) is reached. Its corresponding $R_{NX}(K)$ curve is also shown.



Fuente: presentación propia de los autores.

Figure 5. Results for COIL dataset. Results are shown regarding the quality measure $R_{NX}(K)$. The curves and their AUC (a) for all considered methods are depicted, as well as the embedding data (b)-(f). Individual embedding data spaces are obtained by selecting the points rightly on the vertexes meanwhile the mixture is done with the coefficients associated to the central point.



(a) R_{NX} (K) for all considered methods.

(b) CMDS

(c) LE

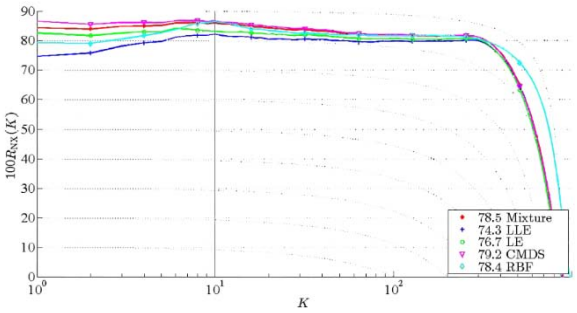
(d) LLE

(e) Mixture of Kernels

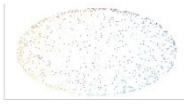
(f) RBF

Fuente: presentación propia de los autores.

Figure 6. Results for the spherical shell dataset.



(a) R_{NX} (K) for all considered methods.

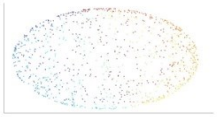


(b) CMDS

(c) LE

(d) LLE

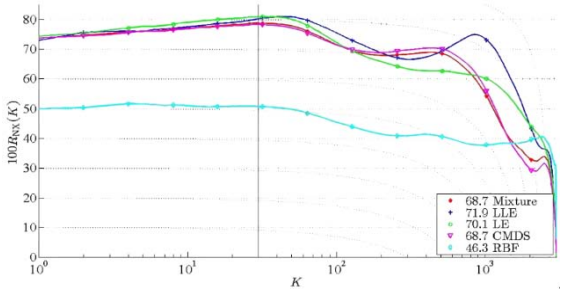
(e) Mixture of Kernels



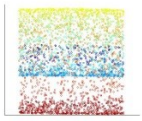
(f) RBF

Fuente: presentación propia de los autores.

Figure 7. Results for the Swiss roll dataset.

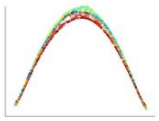


(a) R_{NX} (K) for all considered methods.

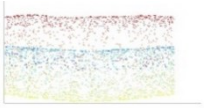


(b) CMDS

(c) LE



(d) LLE

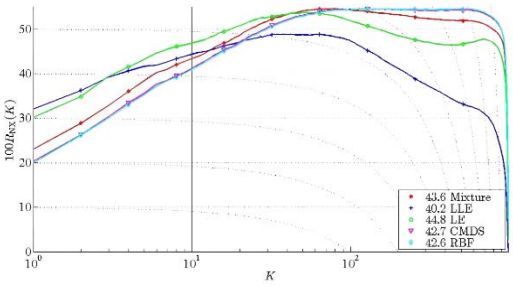


(e) Mixture of Kernels

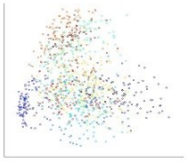
(f) RBF

Fuente: presentación propia de los autores.

Figure 8. Results for MNIST dataset.

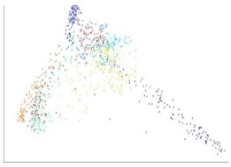


(a) R_{NX} (K) for all considered methods.



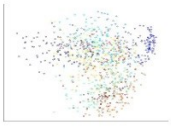
(b) CMDS

(c) LE



(d) LLE

(e) Mixture of Kernels



(f) RBF

Fuente: presentación propia de los autores.

ANEXO 5. MANUAL DE LA INTERFAZ DE USUARIO

MANUAL DE USUARIO

- **Selección de datos para analizar**

Para seleccionar el conjunto de datos deseados abra el menú desplegable en el panel denotado como *data sets* y proceda a seleccionar el conjunto de datos comprendido entre el rollo suizo, la esfera, MNIST o Coil 20. Si seleccionó un conjunto de datos diferente al deseado, repita el proceso anterior para seleccionar los datos para análisis.

- **Selección de la figura geométrica**

Abra el menú desplegable del panel *model* y seleccione la cantidad de métodos con los que quiere realizar la mezcla. Escoja de 2 a 4 para trabajar con métodos entre LE, LLE, CMDS y RBF respectivamente. Proceda a seleccionar la cantidad de métodos deseados para crear un polígono geométrico que de acuerdo a su selección tendrá forma de línea recta, triangulo o rombo, una vez seleccione el número de métodos se habilitara el botón para seleccionar un punto para mezclar. Usted es libre de seleccionar el número de métodos comprendido en ese rango pero si su selección fue errónea, puede volver a desplegar el menú y seleccionar un valor diferente, en cualquier caso una figura geométrica será presentada.

- **Representación de resultados**

Para obtener una representación resultante siga los siguientes pasos, primero presione el botón *Select point to mixture* dentro del panel *model*, un cursor de identificación será presentado. Ubique el cursor sobre el punto que quiere tomar para realizar la mezcla y presione el clic izquierdo del mouse. Tenga en cuenta que entre más cercano este el punto seleccionado a un método, este recibirá un mayor peso de ponderación. Una vez seleccione un punto el sistema realizara automáticamente la mezcla de métodos y arrojara un resultado. El resultado será presentado dentro de la gráfica en blanco situada dentro del panel *data results*.

Si el resultado obtenido no satisface sus necesidades, usted es el libre de seleccionar otro punto nuevamente presionando el botón *Select point to mixture*. En cualquier caso, seleccionar otra cantidad diferente de métodos produce otra figura en la que puede explorar seleccionando puntos para suplir sus exigencias.

- **Valorar desempeño de los métodos RD**

Si usted desea, usted puede generar una curva de calidad la que indica el desempeño de los métodos RD según la agrupación de *k* vecinos. En esta curva usted puede observar relaciones entre el número de vecinos y la eficiencia del

desempeño de los métodos cuantitativamente en un valor de porcentaje. Para esto, presione el botón denotado como *Quality Measure*.

Si necesita ayuda, puede presionar el botón *Help* para reproducir un video ilustrativo de cómo operar la interfaz, si esto no despeja sus dudas, en el botón *About*, puede encontrar información sobre los autores así como sus correos de contacto. No dude en escribir.

ANEXO 6. PAGINA WEB

Dentro del desarrollo de este proyecto, uno de los productos esperados era una diseñar una página web en la que se pueda subir información relacionada a la interfaz, algoritmos, ejecutables, datos y otros productos adicionales como los artículos y videos. La página web fue creada en google sites y se puede acceder mediante el siguiente vínculo:

https://sites.google.com/site/degreethesisdiegopeluffo/interactive-interface-for-data-vis