Desarrollo de una herramienta interactiva de análisis de datos integrando técnicas de visualización y modelos de interacción

CIELO KATHERINE BASANTE VILLOTA
CARLOS MANUEL ORTEGA CASTILLO

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INENIERÍA ELECTRÓNICA
SAN JUAN DE PASTO
2018

Desarrollo de una herramienta interactiva de análisis de datos integrando técnicas de visualización y modelos de interacción

CIELO KATHERINE BASANTE VILLOTA
CARLOS MANUEL ORTEGA CASTILLO

TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE INGENIERO ELECTRÓNICO

DIRECTOR

DIEGO HERNÁN PELUFFO ORDÓÑEZ

INGENIERO ELECTRÓNICO

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INENIERÍA ELECTRÓNICA
SAN JUAN DE PASTO
2018

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, 12 de junio de 2018

Universidad de Nariño
Facultad de Ingeniería

ACUERDO No. 144
(8 de noviembre de 2018)

CONSIDERANDO

Que, mediante el Acuerdo 133 de Consejo de Facultad, designa como Jurados Evaluadores a los Ingenieros FREDDY ALEXANDER GUASMAYAN y JOHN EVERT BARCO, para que emitan concepto sobre el trabajo de grado "DESARROLLO DE UNA HERRAMIENTA INTERACTIVA DE ANÁLISIS DE DATOS INTEGRANDO TÉCNICAS DE VISUALIZACIÓN Y MODELOS DE INTERACCIÓN" modalidad Investigación, presentado y sustentado por los estudiantes CIELO KATHERINE BASANTE VILLOTA y CARLOS MANUEL ORTEGA CASTILLO, bajo la dirección del Ingeniero DIEGO HERNÁN PELUFFO ORDÓÑEZ,

Que, los profesores designados como jurados externos, evalúan minuciosamente el trabajo de grado "DESARROLLO DE UNA HERRAMIENTA INTERACTIVA DE ANÁLISIS DE DATOS INTEGRANDO TÉCNICAS DE VISUALIZACIÓN Y MODELOS DE INTERACCIÓN"

Que, mediante oficio del 23 de octubre de 2018, el profesor FREDDY ALEXANDER GUASMAYAN de la Universidad Mariana, ratifica que el trabajo de grado "DESARROLLO DE UNA HERRAMIENTA INTERACTIVA DE ANÁLISIS DE DATOS INTEGRANDO TÉCNICAS DE VISUALIZACIÓN Y MODELOS DE INTERACCIÓN" tiene merecimiento de "Laureado" y soporta su concepto en la pertinencia de la investigación para el programa de Ingeniería Electrónica, asimismo, menciona que el trabajo contiene un fundamento teórico robusto y actualizado al contexto investigativo en el lenguaje matemático y de ingeniería,

Que, mediante oficio del 30 de octubre de 2018, el profesor JOHN BARCO JIMÉNEZ de la Institución Universitaria Cesmag, considera que el trabajo de grado "DESARROLLO DE UNA HERRAMIENTA INTERACTIVA DE ANÁLISIS DE DATOS INTEGRANDO TÉCNICAS DE VISUALIZACIÓN Y MODELOS DE INTERACCIÓN" merece la distinción de "Laureado", por cuanto se pudo evidenciar que el tema abordado es de interés actual para la comunidad de investigadores, además, los autores realizaron un análisis comparativo profundo de espacios embebidos y representación con kernell; es de desatacar que el trabajo ya cuenta con una validación externa lograda a través de tres artículos científicos en Publindex de Colciencias.

Que, este organismo acoge favorable los conceptos de los Jurados evaluadores externos FREDDY GUASMAYAN y JOHN BARCO JIMÉNEZ y en consecuencia,

ACUERDA

Art. 1°. Otorgar la distinción de Laureado el trabajo de grado "DESARROLLO DE UNA HERRAMIENTA INTERACTIVA DE ANÁLISIS DE DATOS INTEGRANDO TÉCNICAS DE VISUALIZACIÓN Y MODELOS DE INTERACCIÓN" modalidad Investigación, presentado y sustentado por los estudiantes CIELO KATHERINE BASANTE VILLOTA y CARLOS MANUEL ORTEGA CASTILLO, bajo la dirección del Ingeniero DIEGO PELUFFO.

Art. 2°. Reconocer y exaltar a los Ingenieros CIELO BASANTE y CARLOS ORTEGA, quienes bajo la dirección del Ingeniero DIEGO PELUFFO, lograron la distinción "Laureado" en su trabajo de grado.

Art. 3°. Facultad de Ingeniería, Departamento de Electrónica, anotarán lo de su cargo.

COMUNÍQUESE Y CÚMPLASE

Dada en San juan de Pasto, a los ocho días (8) de noviembre de 2018

DELIO GOMEZ LOPEZ
Presidente
Aprobó: Consejo de Facultad

GLORIA AMPARO THOME P.
Secretaria

AGRADECIMIENTO

“Agradezco a mi madre por darme la vida, quererme mucho, creer en mi y porque siempre me apoya. Mamá gracias por darme una carrera para mi futuro, todo esto te lo debo a ti. A mis compañeros con quienes pasamos muy buenos momentos en nuestra formación profesional, especialmente a Cielo Basante por su paciencia y constancia durante el desarrollo de este trabajo final. A mis profesores por compartir sus conocimientos y experiencias. Especialmente quiero agradecer a mi asesor el profesor Diego Hernán Peluffo Ordoñez por brindarme su confianza, amistad y acompañamiento para hacer este nuevo logro posible, además su paciencia y motivación han sido fundamentales para mi formación como investigador”.

CARLOS MANUEL ORTEGA CASTILLO

“Agradezco a mis padres y familia por el incansable apoyo durante mi carrera, cada uno de ustedes contribuyó de alguna manera para llegar hasta aquí. A mis amigos y compañeros que compartieron conmigo este proceso de enriquecimiento académico, en especial a Carlos Ortega por su ayuda y comprensión en todo momento. A mis profesores por ejercer con amor su profesión. Especialmente quiero agradecer a mi asesor el profesor Diego Peluffo quien más allá de la academia siempre me brindo su amistad y afecto, sin su apoyo y motivación este logro no sería posible, gracias por ampliar mis horizontes hacia el camino de la investigación”.

CIELO KATHERINE BASANTE VILLOTA

DEDICATORIA

“A mi madre Miriam Castillo porque sin su apoyo no sería posible alcanzar esta meta. Su constante compañía me ha dado la fuerza y ánimo para seguir siempre adelante. A mi hermano Héctor Ortega por estar siempre pendiente de mi desarrollo profesional y personal y por alentarme a seguir intentándolo en los momentos difíciles. A mi tía Rubiela Castillo por sus consejos y enseñanzas que siempre me guían por el buen camino. A mi abuelo Alfonso Castillo que siempre buscó mi bienestar y a pesar de que ya no está presente físicamente, seguramente está muy feliz de mi logro”.

CARLOS MANUEL ORTEGA CASTILLO

“A mi padre Eduardo Basante y mi madre Nohora Villota quienes, con su apoyo, esfuerzo, comprensión, pero sobre todo amor, han hecho posible que alcance este escalón en mi vida, siempre me han dado lo mejor de ustedes en aras de mi porvenir y cada esfuerzo se ve reflejado en la persona que soy, gracias por ser los precursores de mis sueños y metas sin importar las adversidades. A mi familia por hacer parte de cada una de las etapas de mi vida como base de amor y unión, especialmente a mis familiares que ya no pueden compartir conmigo, pero que me brindaron todo su afecto y conocimiento para conseguir este logro”.

CIELO KATHERINE BASANTE VILLOTA

RESUMEN

Actualmente, las capacidades humanas de análisis se han visto limitadas ante el inminente crecimiento de las tecnologías para recolectar, comunicar y almacenar grandes volúmenes de información. En general, dichos volúmenes de información se representan en bases de datos de alta dimensión, las cuales no pueden ser directamente interpretadas de forma visual. En este sentido, la reducción de dimensión (RD) se ha convertido en una buena alternativa. Las técnicas de RD extraen información relevante de la base de datos de entrada representada en baja dimensión con el fin de mejorar el desempeño de las subsecuentes tareas de reconocimiento de patrones y minería de datos. Por lo general, la aplicación y la interpretación de procedimientos de RD suelen requerir de personal experto en análisis de datos, y, por tanto, se genera un incremento en tiempo y costo para el desarrollo de las etapas posteriores de análisis. Lo anterior ha motivado el desarrollo de estrategias interactivas que permitan representar de manera gráfica los resultados del análisis de datos de alta dimensión. No obstante, en la comunidad académica, aún se considera un problema abierto el desarrollo de una herramienta -genérica, versátil, de software libre y de dedicación exclusiva- para la visualización interactiva de datos basada en principios de RD.

En este trabajo de grado, se presenta la implementación de una herramienta versátil para el soporte de análisis visual de bases de datos, que permite al usuario obtener interactivamente representaciones gráficas en baja dimensión. Para tal fin, la herramienta incorpora: i) modelos de interacción -uno existente (modelo cromático) y uno propuesto (modelo basado en ángulos)-, ii) una mezcla de métodos espectrales de RD representados en aproximaciones en matrices kernel, iii) técnicas tradicionales de visualización (diagramas de dispersión y diagrama de coordenadas paralelas). Adicionalmente, con el propósito de generar una interacción dinámica (cambios en tiempo real), se implementa el algoritmo de sub-matrices localmente lineales para llevar a cabo el proceso de reducción de dimensión con un menor coste computacional. Es importante resaltar que toda la herramienta se desarrolla propiciando escalabilidad y modularidad.

ABSTRACT

Currently, human analysis capabilities are not enough imminent in the face of growth of technologies aiming to collect, communicate and store large volumes of information. Typically, such volumes of information are represented in high-dimensional databases, which can not be directly interpreted visually. That said, the dimensionality reduction (DR) has become to be a good alternative. From the input database, the DR techniques extract relevant information represented in a low-dimensional fashion, so that the performance of the subsequent pattern recognition and data mining tasks is improved. In general, the application and interpretation of RD procedures require expert personnel in data analysis, and, therefore, an increase in time and cost is generated for carrying out the subsequent stages of analysis. This fact has motivated the development of interactive strategies allowing for graphically representing outcomes of the analysis of high-dimensional data. Nonetheless, for the academic community, the development of a - generic, versatile, open source and dedicated - tool for the interactive visualization of data based on DR principles is still considered an open issue.

In this degree work, the implementation of a versatile tool for the support of databases visual analysis is presented, which enables the user to interactively generate low-dimensional graphic representations. For this purpose, the tool incorporates: i) interaction models - an existing one (chromatic model) and another proposed one (model based on angles) -, ii) a mixture of DR spectral methods represented in kernel-matrices-based approximations, iii) technical traditional visualization (scatter plots and parallel coordinates diagram). Additionally, aimed at generating a dynamic interaction (real time changes), the locally linear landmarks algorithm is implemented to perform the DR procedure at a low-computational cost. It is important to highlight that the entire tool is developed under scalability and modularity settings

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

GLOSARIO

Big Data: Big Data, macrodatos o datos masivos es un concepto que hace referencia al almacenamiento de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de dichos datos.

Dimensión: En términos generales, la dimensión de una base de datos es definida como la cantidad de mediciones, características o atributos que tiene cada objeto o muestra.

Reducción de dimensión: Las técnicas de reducción de la dimensión tienen por objetivo final condensar la información de un conjunto de variables en un nuevo conjunto de variables (de menor número que el anterior), con la menor pérdida de información posible.

Espacio embebido: En este trabajo el término “espacio embebido” hace referencia a la representación en baja dimensión resultante cuando un método de reducción de dimensión es aplicado a una base de datos de alta dimensión.

Diagrama de dispersión: Un diagrama de dispersión o gráfica de dispersión o gráfico de dispersión es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos o tres variables para un conjunto de datos.

Coordenadas paralelas: Visualización ideal para comparar varias variables en 2 o 3 dimensiones, cada eje paralelo representa una variable o atributo.

Imagen RGB: Son aquellas imágenes que están representadas a través de los colores primarios rojo, verde y azul. Este tipo de imágenes son las más adecuadas para ser mostradas en monitores y que, finalmente, serán impresas en impresoras de papel fotográfico.

Pixel: El píxel puede definirse como la más pequeña de las unidades homogéneas en color que componen una imagen de tipo digital.

Resolución de intensidad: En escala de grises la resolución de intensidad de una imagen se define como los niveles de grises que una imagen puede tener.

Resolución de espacial: Se define como y el número de filas (el alto) y columnas (el ancho) que tiene una imagen, es decir, el número de pixeles.

INTRODUCCIÓN

Así como el universo físico es muy diverso y amplio, el universo digital abarca todo lo creado o definido por software construido por el hombre, dicho software analiza este universo en constante expansión para encontrar valor oculto y nuevas oportunidades para mejorar el mundo físico. Es imposible calcular la dimensión del universo físico, pero para el universo digital se prevé que para 2020 se alcancen los 44 zettabytes de información -una tasa de crecimiento del 40% por año-, En otras palabras, habrá tantos bits digitales como estrellas en el universo [3]. Sin embargo, el extraer información realmente útil de estos billones de datos que superan cuantiosamente las capacidades humanas de análisis, es un desafío en sí mismo. Surge entonces, la necesidad de proponer o buscar técnicas que permitan representar información exorbitante de una manera más inteligible para el humano, de manera que se aproveche de manera adecuada el contenido de los datos [1], [2], [4]. Una de estas técnicas, es la reducción de dimensión (RD) cuyo objetivo principal es el obtener una nueva representación de los datos originales de alta dimensión, preservando tanto como sea posible su topología, en una dimensión menor. En este sentido, la RD se convierte en una herramienta clave para procesos como el análisis de datos y la aplicación de metodologías de aprendizaje de máquina, dado que produce información más depurada, eliminando errores, redundancias e irrelevancias [7]. En la actualidad existen una gran variedad de herramientas que integran métodos de RD para la representación y visualización de datos. Sin embargo, La mayoría de dichas herramientas son demasiado complejas, abstractas y carecen de interactividad aspectos importantes ante usuarios con conocimientos superficiales en el tema. Además, estas no son de uso versátil, en otras palabras, no le ofrecen al usuario la posibilidad de utilizar los métodos de RD desde diferentes enfoques con el fin de obtener mayor número de representaciones y así acortar la brecha entre el usuario y el conocimiento dentro de los millones de datos [5], [6].

Este trabajo se basa en conceptos y técnicas de reducción de dimensión para el desarrollo de una herramienta versátil de visualización de datos que permite al usuario obtener representaciones interactivas de los datos en baja dimensión. Para este fin, se implementaron tres métodos espectrales de reducción de dimensión en sus aproximaciones de matrices *kernel*: ***Locally Linear Embedding*** (LLE) [9], ***Classical Multidimensional Scaling*** (CMD) [10], [11] y ***Laplacian Eigenmaps*** (LE) [12], quienes podrán ser usados en diferentes enfoques bien sea cada uno por separado o la mezcla ponderada de estos. Para lograr dicha mezcla se integran dos modelos de interacción, el primero es un modelo cromático propuesto en [6] y está basado en el espacio de color RGB, donde cada color primario (rojo(R), verde(G) y azul(B)) representa un método de RD de tal manera que cada color derivado de la combinación de estos colores se verá reflejado como la mezcla de métodos. Se propuso un segundo modelo basado en la medida

de los ángulos de un triángulo donde cada ángulo se identifica como un método RD.

Una vez obtenida la matriz total de la mezcla de las aproximaciones kernel se usa el algoritmo generalizado de análisis de componentes KPCA [6], [8] para obtener el espacio embebido y así pasar a la etapa final de visualización donde se hace uso de dos técnicas tradiciones: una de ellas es el diagrama de dispersión (scatter plot) el cual permite representar datos de hasta 3 dimensiones. Adicionalmente se encuentra la técnica de coordenadas paralelas que permite representaciones de hasta 10 dimensiones, aumentando el espectro de visualización del usuario. Dado que la herramienta es genérica y el usuario puede usar su propia base de datos, se consideró que para bases de datos demasiado grandes los algoritmos implementados podrían tardar demasiado y en efecto la representación ya no sería interactiva. Por esa razón, se decidió implementar el algoritmo de sub-matrices localmente lineales (LLL) [14] para llevar a cabo el proceso de reducción de dimensión con un menor coste computacional, propiciando una representación dinámica (cambios en tiempo real) de los datos. Para probar la versatilidad, interactividad y controlabilidad de la herramienta se realizaron pruebas con 3 bases de datos: una de ellas real (imágenes de dígitos- MNIST) y dos bases de datos artificiales (cascarón esférico en 3D y rollo suizo). El desempeño de los métodos de reducción de dimensión es evaluado mediante una versión escalada de la tasa promedio del acuerdo entre los k-vecinos más cercanos como se explica en [13].

1. DESCRIPCIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La generación, almacenado y procesado de datos (aun en alta dimensión) se han convertido en una necesidad y una actividad cotidiana en diversos sectores. Por ejemplo, grandes y pequeñas empresas están optando por registrar todas las actividades que son ejecutadas por sus clientes y empleados, con el objetivo de tomar decisiones que causen mayor beneficio a la empresa a través del análisis exhaustivo de los datos recolectados. La tecnología ha tenido un avance importante en cuanto al manejo de información, debido a que terabytes de información representados en bases de datos son generados a diario [1], [16]. Aunque el almacenamiento de los datos no ha sido un problema, puesto que la tecnología actualmente permite guardar mayor cantidad de información en espacios físicos reducidos y a una velocidad superior; en contraste se observa que las capacidades de análisis de dicha información no han evolucionado, dejando una brecha entre el usuario y la información [17], [19]. Por consiguiente, es de gran importancia buscar nuevos modos de presentar y/o representar datos de forma comprensible e intuitiva, debido a que en muchas ocasiones los datos no son fácilmente interpretables y, por tanto, se hace necesario la intervención de un experto con conocimientos en análisis de datos [15] [16].

La alta dimensión de los datos es una de las brechas más grandes que existen entre el usuario y la información, debido a que la percepción humana está limitada a tres dimensiones, e inclusive, en tres dimensiones los datos pueden ser ambiguos y no muy claros [16], [18]. En general, la dimensión en una base de datos está definida por el número de mediciones o atributos que posee cada muestra u objeto, en consecuencia, si un registro tiene más de tres mediciones no podrá ser representado de manera tradicional en un diagrama de dispersión en el plano cartesiano, dificultando el proceso de descubrimiento de nuevo conocimiento en una base de datos [15]. La visualización interactiva de datos permite que esa información presente en grandes repositorios (*Big Data*) sea más inteligible, dinámica y provechosa para el ser humano [16], [17], [19]. Por esta razón, la visualización de datos es, en muchos casos, imprescindible, especialmente, en las etapas de análisis donde se hacen hipótesis significativas sobre los datos.

El proceso de extracción de información y detección de patrones es conocido como “Descubrimiento de conocimiento en bases de datos” (DCBD), el cual tiene como finalidad encontrar información de interés o formular predicciones de algún evento en particular. El DCBD ya ha sido ampliamente explorado y

desarrollado en diversos estudios y aplicaciones, tales como: la determinación de perfiles de clientes fraudulentos [22], el descubrimiento de una relación implícita que exista entre síntomas y enfermedades [20], [21] el análisis de mercado, ventas y soluciones a clientes [16], [20], entre otras aplicaciones.

Finalmente, con objeto de facilitar la implementación de un proceso DCBD se han desarrollado herramientas que utilizan técnicas que implican el preprocesamiento, el uso de métodos de minería de datos y/o visualización [25], [26]. Estas herramientas presentan una enorme desventaja, puesto que procesan datos de alta dimensión con un costo computacional alto. Además, presentan resultados difícilmente interpretables, es decir, para analizar los resultados es necesario contar con un conocimiento previo o tener experiencia en análisis de datos [12]. También se considera que estas herramientas no mantienen una apropiada interactividad con el usuario debido a que no es posible variar utilizar bajo diferentes enfoques las técnicas de RD que podrían ser útiles para las necesidades del usuario [23], [24]. Las herramientas de visualización de grandes volúmenes de datos se encargan de aplicar, métodos de minería de datos para la extracción de patrones en formas de reglas o funciones, o métodos de reducción de dimensión orientados a la visualización, pero en cualquier escenario, interpretar las representaciones obtenidas puede llegar a ser algo abstracto [12], [25]. Por lo tanto, se aprecia que existe la necesidad de que el usuario cuente con una herramienta interactiva que permita manipular los métodos de DCBD acorde a sus necesidades para obtener los resultados deseados y que éstos sean fácilmente interpretables sin un conocimiento previo de la metodología aplicada [25], [27].

1.2 JUSTIFICACIÓN

En la actualidad, las tareas de reconocimiento de patrones y la minería de datos involucran generalmente bases de datos de alta dimensión que no pueden ser representados de manera tradicional en el diagrama de dispersión [15], por esta razón la reducción de dimensión se convierte en una herramienta importante para conocer la naturaleza de una base de datos en particular, de esta manera las técnicas RD pueden ayudar a un usuario a elegir metodologías apropiadas para realizar tareas de clasificación, predicción, extracción de nuevo conocimiento, entre otras [27], [28], [29].

La obtención de información relevante de estos enormes volúmenes de datos resulta ineludible en la actualidad, puesto que el descubrimiento de esta nueva información hace la diferencia a la hora de tomar una buena decisión en ámbitos sociales como la economía, el marketing, los negocios, redes de telecomunicación, las leyes, entre otros. Uno de los inconvenientes que se presenta a la hora de analizar las bases de datos es la inexperiencia que tienen las personas en el ámbito de las herramientas DCBD y técnicas RD, puesto que es tarea de un experto aprovechar de la mejor manera la información contenida

en dichas bases de datos.

El desarrollo de este trabajo busca la unión de métodos de interacción y técnicas visualización mediante una interfaz gráfica, la cual será intuitiva, interactiva y podrá exponer al usuario el efecto que tienen las mezclas de métodos espectrales de RD en la representación de su base de datos original en una dimensión menor. Teniendo en cuenta que la herramienta puede ser usada en cualquier área del conocimiento donde se generen datos, cualquier usuario estará en condición de utilizarla adecuadamente, incluso aunque no cuente con conocimientos previos de análisis de datos, es decir, podrán aplicar e interpretar los métodos de RD sin conocer a fondo los aspectos teóricos en los cuales se fundamentan. Adicionalmente, la herramienta cuenta con dos técnicas de visualización que dan continuidad y presentan de forma tangible los resultados de las investigaciones en [6] y [30].

1.3 CONTRIBUCIÓN DE ESTA TESIS

Algunas etapas de la minería de datos se fundamentan en el uso de técnicas de reducción de dimensión como parte de la visualización. Puesto que, los resultados de la aplicación de dichos métodos pueden ser abstractos y se hace necesario la supervisión de un experto para encontrar información útil, surge la necesidad de utilizar metodologías que le brinden al usuario la oportunidad de aplicar técnicas de RD de manera más intuitiva y sencilla. En esta investigación se desarrolla una herramienta de visualización de datos que permita una interacción y control entre el usuario y la máquina, con el fin de que esté pueda manipular e interpretar fácilmente los resultados. Uno de los factores más importantes de la herramienta propuesta es la versatilidad y controlabilidad, dado que permite el usuario hacer uso de los métodos de RD bajo diferentes escenarios, no solo podrán ser empleados de manera independiente, si no también una mezcla pondera de estos.

El usuario tendrá entonces, la facultad de escoger dichos factores de ponderación de manera dinámica, por cualquiera de los dos modelos de interacción, hasta llegar a la representación que más se adecue a sus necesidades. Otro factor importante que cabe mencionar es la generalidad de la herramienta la cual permite que el usuario pueda subir su propia base de datos y que estos sean explorados, dicha base de datos debe contar con datos cuantitativos, además se debe resaltar que para bases de datos demasiado grandes la herramienta podría presentar problemas de procesamientos, se sugiere usar bases de datos que cuenten con un volumen de datos razonable. Teniendo en cuenta que en la actualidad existen grandes volúmenes de datos que pueden hacer que la ejecución del proceso de visualización se dificulte, se implementó un método de optimización, con el objetivo de reducir el costo computacional del cálculo de la descomposición espectral para la RD, llamado sub-matrices localmente lineales (*Locally Linear Landmarks* -LLL-) [14], algoritmo que brinda mayor autonomía al usuario dándole libertad para usar la

metodología propuesta en cualquier repositorio de datos.

La metodología de visualización propuesta aporta soluciones a problemáticas existentes en la representación y visualización de datos puesto que, permitirá que bases de datos puedan ser representadas e interpretadas fácilmente en tiempos de procesamiento razonables. Adicionalmente, se implementaron dos técnicas de visualización: gráficos de dispersión, que es generalmente utilizada en este tipo de herramientas, y coordenadas paralelas, que permiten visualizar los datos en hasta diez dimensiones, y no limitar al usuario a visualizaciones únicamente en dos dimensiones. Además, para que el usuario pueda probar rápidamente la herramienta sin contar con datos externos, ésta incorpora 4 bases de datos de uso común que permiten introducir la herramienta a cualquier usuario.

Otra contribución importante tiene que ver con el aporte científico y la divulgación de resultados a través de la publicación de artículos científicos obtenidos con el desarrollo de este trabajo de grado. Se publicaron tres artículos en total, dos como autores principales en CCC'18 y IWAIPR'18 y uno como coautores IDEAL'18. Las temáticas abordadas en los dos artículos están relacionadas con visualización de datos, además, de formas interactivas y eficientes de aplicar métodos de reducción de dimensión a bases de datos en particular.

1.4 OBJETIVO GENERAL

Desarrollar una herramienta versátil y genérica de visualización interactiva de datos, implementada en lenguaje de alto nivel, basada en reducción de dimensión integrando diferentes modelos de interacción y métodos de visualización.

1.5 OBJETIVOS ESPECÍFICOS

- Establecer los modelos de interacción y métodos de reducción de dimensión con criterios de factibilidad, para ser implementados en leguaje de alto nivel.

- Implementar técnicas de visualización de datos en lenguaje de programación de alto nivel recomendados por la literatura científica.

- Integrar métodos de reducción de dimensión, modelos de interacción y técnicas de visualización en el desarrollo de una herramienta interactiva de visualización de datos

1.6 ORGANIZACIÓN DEL DOCUMENTO

Este trabajo está dividido en 5 secciones principales nombradas de la siguiente

manera: Introducción, descripción del problema y objetivos, marco teórico, metodología, resultados y conclusiones.

En la sección 2, se presenta el planteamiento del problema, la justificación de este trabajo y las contribuciones científicas de esta investigación. Asimismo, el objetivo general y los objetivos específicos planteados al inicio de este trabajo de grado.

En la sección 3, se presenta una revisión bibliográfica que incluyen conceptos básicos de *Big Data*, alta dimensión en bases de datos y las diferentes etapas que conforman el proceso de descubrimiento de nuevo conocimiento en bases de datos como, la reducción de dimensión y visualización de datos. Asimismo, algunas técnicas de visualización interactiva y algoritmos para la optimización del coste computacional en el proceso de RD.

En la sección 4, se describe la metodología de visualización propuesta, así como la implementación de los modelos de interacción, la medida de calidad utilizada, las bases de datos empleadas y el método de optimización de costo computacional implementado LLL. Del mismo modo, Los resultados de los experimentos se discuten en la sección 5.

Finalmente, en la sección 5, se presentan las conclusiones que se obtuvieron a partir de este trabajo, y los trabajos futuros que pueden mejorar la metodología de visualización propuesta y definir nuevas investigaciones.

2. MARCO TEÓRICO

2.1 BIG DATA Y DATOS DE ALTA DIMENSIÓN

Big Data se refiere a la avalancha de datos digitales de muchas fuentes terrestres digitales, incluidos sensores, digitalizadores, escáneres, modelos numéricos, teléfonos móviles, Internet, videos, correos electrónicos y redes sociales. Los tipos de datos incluyen textos, geometrías, imágenes, videos, sonidos y combinaciones de cada uno. Dichos datos pueden estar relacionados directa o indirectamente con la información comercial y social de la comunidad global [24], [31]. Dicha información puede ser estructurada, semiestructurada o no estructurada y puede aportar enorme valor a cualquier entidad. Sin embargo, trabajar con estos grandes volúmenes de información supone un consumo excesivo de recursos humanos e informáticos para su correcta manipulación e interpretación [16], [32].

Los conceptos fundamentales que son agrupados en el área del Big Data son: volumen, visualización, variabilidad y velocidad; todos suponen un reto al momento de procesar y almacenar la información. La variedad de su origen y la rapidez con la que se incrementa su volumen (**Figura 1**), son algunos de los factores que dan lugar a esta área emergente de investigación. El avance de la tecnología ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos que tomaría demasiado tiempo y sería muy costoso cargarlos a una base de datos relacional para su análisis. Por ejemplo, uno de los conceptos innovadores en los últimos años ha sido internet de las cosas (IOT), el cual consiste en tener dispositivos avanzados y sus sensores conectados a un sistema ciber-físico para medir ubicación, movimiento, vibraciones, temperatura, precipitación, humedad, cambios químicos, entre otros. IOT genera continuamente flujos de datos de todo el mundo, generalmente desde dispositivos móviles interconectados como computadoras personales, celulares, cámaras, etc.

El almacenamiento de *Big Data* en el almacenamiento físico tradicional es problemático ya que las unidades de disco duro (HDD) a menudo fallan y los mecanismos tradicionales de protección de datos no son eficientes con el almacenamiento a gran escala. Además, la velocidad de Big Data requiere que los sistemas de almacenamiento pueden escalar rápidamente, lo que es difícil de lograr con los sistemas de almacenamiento tradicionales. Actualmente los servicios de almacenamiento en la nube ofrecen un almacenamiento prácticamente ilimitado con alta tolerancia a fallas, lo que brinda posibles soluciones para abordar los desafíos de almacenamiento de Big Data. Sin embargo, transferir y alojar *Big Data* en la nube es costoso dado el tamaño del volumen de datos. Los principios y algoritmos deben desarrollarse para determinar el valor informativo de los datos y sus conjuntos de datos de preservación al equilibrar el costo del almacenaje y transmisión de datos con la rápida acumulación de información.

Una ventaja importante del *Big Data* es la información adicional que puede ser obtenida de grandes bases de datos en lugar de un análisis de bases de datos pequeñas y separadas entre sí. La visualización de Big Data descubre patrones ocultos y descubre correlaciones desconocidas para mejorar la toma de decisiones. Dado que *Big Data* a menudo es heterogéneo en cuanto a tipo, estructura y semántica, la visualización es fundamental para darle sentido a esa información. Pero es difícil proporcionar visualización en tiempo real e interacción humana para explorar y analizar. Según [33], existen cinco funcionalidades clave para la visualización de *Big Data* de la siguiente manera: gráficos altamente interactivos que incorporan las mejores prácticas de visualización de datos; análisis visual integrado, intuitivo y accesible; interfaces interactivas basadas en web para previsualizar, filtrar o muestrear datos antes de las visualizaciones; procesamiento en memoria; y respuestas y perspectivas fácilmente distribuidas a través de dispositivos móviles y portales *web*. La metodología de estas aplicaciones depende en gran medida de la eficacia de procesamiento y extracción de patrones significativos de los conjuntos de datos y la precisión de la búsqueda [7], [23].

Figura 1. Según [3] el universo digital tiene una tasa de crecimiento de 40% por año, lo que permite decir que para 2020 se tendrán más de 44 zettabytes de información, tantos bytes de información como estrellas en la galaxia **Fuente:** [3]

Tradicionalmente la estructura de un conjunto de datos se presenta como una matriz, de tantas filas como variables medidas en cada unidad (individuo, empresa, inmueble, calle de una gran ciudad, procedimiento judicial, etc.) y, tantas

columnas como datos [16], [20]. De este modo los volúmenes de datos generados actualmente (*Big Data*) crecen en cantidad y dimensión, lo que hace, que un registro pueda fácilmente tener docenas de atributos y el dominio de cada atributo un rango bastante amplio, por este motivo, es importante obtener representaciones fácilmente interpretables en bases de datos de alta dimensión [26], [34]. Sin embargo, el problema de los datos de alta dimensión es a menudo abordado por el usuario delimitando el análisis sólo a unas cuantas características o mediciones. En consecuencia, la definición de un subespacio que sólo contenga algunas de las mediciones efectuadas en cada registro puede ser propenso a errores [27], [32], por consiguiente, necesario encontrar un espacio de características reducido que posea la información más relevante del espacio original.

2.2 DESCUBRIMIENTO DE CONOCIMIENTO DE BASES DE DATOS

La emergencia de métodos de manejo de grandes volúmenes de datos producidos a diario en diferentes áreas ha nombrado esta época como la "era de la información". Los millones de datos alrededor del mundo presentes en nuestra vida parece crecer sin un fin a la vista. Además, herramientas computacionales facilitan que terabytes de información sean almacenados, pero aprovechar todo el potencial de información oculto o no explícito, también el procesar dichas bases de datos es cada vez una tarea más difícil, por lo que actualmente existen un sin número de herramientas que tienen como objetivo el darles sentido a los datos, así como descubrir patrones, identificar oportunidades, predecir nuevas situaciones, etc. [7], [8].
Surge la necesidad de desarrollar y aplicar algoritmos y herramientas que permitan al usuario analizar, entender, visualizar y tomar decisiones de los enormes almacenamientos de datos [9]. El reto fundamental para las aplicaciones de *Big data* es el explorar una base de datos y extraer descubrimiento, esta tarea o área de investigación es principalmente abordada por el área de **Descubrimiento de conocimiento de bases de datos** (DCBD o KDD -por sus siglas en inglés-) proceso que envuelve todo el procedimiento que conlleva el transformar los datos de un nivel bajo a un nivel alto de conocimiento en otras palabras DCBD es una tarea no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles en los datos [10].

El proceso DCBD puede ser dividido en las siguientes etapas (**Figura 2**)**: Creación de un conjunto de datos objetivo,** en esta etapa se elige un subconjunto de datos y variables o atributos que serán usados para la tarea de descubrimiento, luego se procede a realizar una **limpieza de los datos**, etapa que soluciona problemas como, datos incompletos (donde hay atributos o valores de atributos perdidos), ruido (valores incorrectos o inesperados) y datos inconsistentes (conteniendo valores y atributos con nombres diferentes). Los datos

“sucios” en algunos casos deben ser eliminados ya que pueden contribuir a un análisis inexacto y resultados incorrectos. **Integración de los datos**, combina datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos. **Selección de tarea de minería de datos**, consiste en buscar el objetivo y las herramientas del proceso de minería, identificando los datos que han de ser extraídos, buscando los atributos apropiados de entrada y la información de salida para representar una tarea como clasificación, regresión, *clustering* etc. **Transformación de los datos** en esta etapa los datos son transformados y consolidados en formas apropiadas para la minería de datos, algunas veces la transformación y la consolidación de los datos son desarrollados antes del proceso de selección de datos, la reducción de los datos puede también ser desarrollada en esta etapa para obtener una representación más acorde a una tarea en particular. **Minería de datos**, en este proceso se aplican métodos inteligentes para la extracción de patrones inmersos en un conjunto de datos. **Interpretación del patrón**, se identifican los verdaderos patrones de interés que representan conocimiento. Por último, se tiene una etapa denominada **representación del conocimiento** en donde técnicas de visualización y representación son usadas para presentar el conocimiento en forma de patrones al usuario [1], [9], [10].

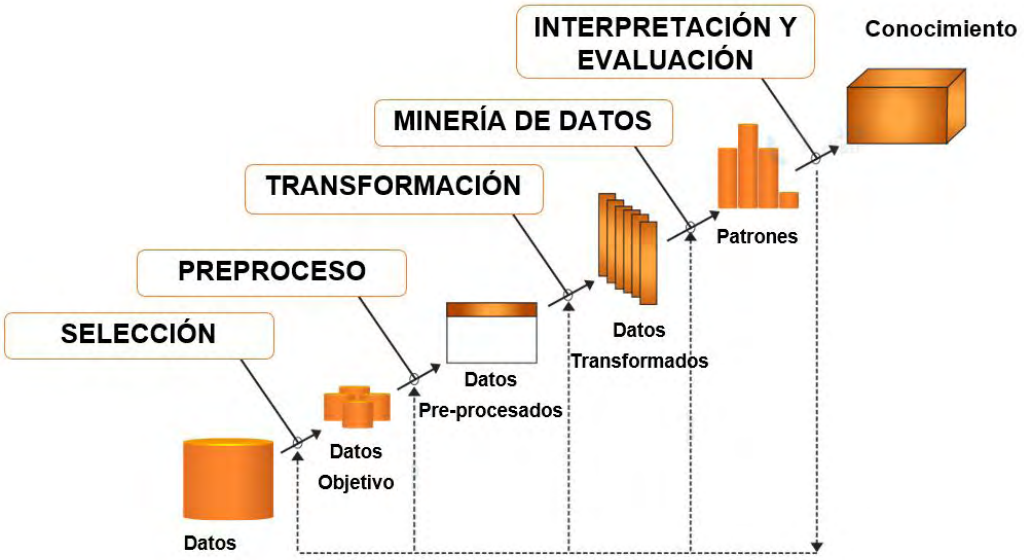


Figura 2. Esquema general de las etapas presentes en el proceso de descubrimiento de conocimiento en bases de datos (DCBD): integración de los datos, selección de los datos, transformación de los datos, minería de datos, evaluación del patrón, representación del conocimiento y por último interpretación y evaluación para encontrar nuevo conocimiento. **Fuente:** Realizado en esta investigación, basado en [9].

Muchas bases de datos que son usadas en el proceso DCBD cuentan con un número bastante amplio de muestras haciendo referencia al término B*ig data*,

teniendo en cuenta esto en la actualidad se han desarrollado herramientas que permitan facilitar el desarrollo de las etapas DCBD sin que se entorpezca debido a los grandes volúmenes de información. Las técnicas de reducción de dimensión surgen tras la necesidad de tratar eficazmente estos datos de alta dimensión siendo reducido el número de muestras por ejemplo 8000 a una cantidad manejable o entendible como de 4 o 5. Sin embargo el usar técnicas de reducción de dimensión implica la intervención de un experto que pueda interpretar el tipo de muestras que se obtienen que muchas veces pueden ser abstractas [7], [21], [35].

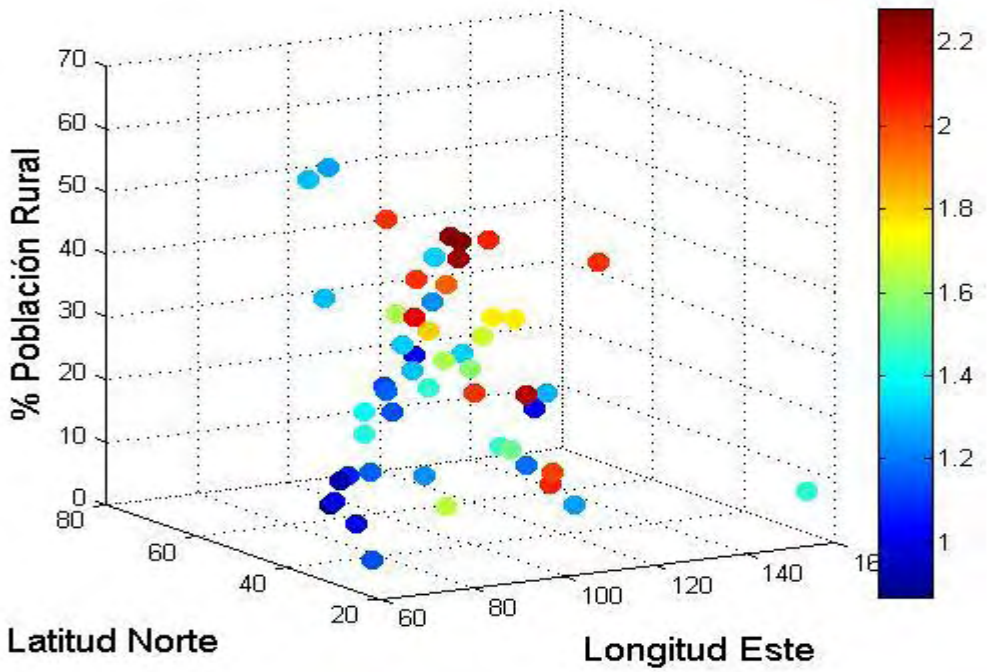


Figura 3. Diagrama de dispersión (*scatter plot*) de la base de datos disponible en Matlab que representa las muertes en avenidas de estados unidos como función de la longitud, latitud y si la locación es rural o urbana (representados en los 3 ejes). Los colores representan el número de muertes. **Fuente**: Esta investigación.

La reducción de dimensión como una parte importante del proceso DCBD está inmerso en diferentes etapas. En partes como preprocesamiento y limpieza de los datos el proceso de RD es bastante útil sobre todo para comprimir o reducir el número de datos para luego aplicar técnicas de aprendizaje de máquina y minería de datos de una manera más eficiente. En la etapa de aplicar algoritmos de minería de datos, RD puede ser considerada como una técnica para el descubrimiento de patrones propios de los datos, dichos patrones permiten generar modelos de clasificación, reglas de asociación y agrupamientos. Finalmente, en la etapa de visualización, RD permite representar datos de alta dimensión en una dimensión perceptible e inteligible para los humanos.

Como el objetivo de este trabajo es realizar una visualización de datos interactiva

para cualquier usuario, por lo que en las etapas de DCBD en las que se enfoca son la transformación de datos y la visualización, de manera que la RD permita reducir la dimensión de los datos y poder representar correctamente la naturaleza intrínseca de estos. Este proceso tiene como función mapear datos en un espacio de características menos bajo específicos criterios que conserven características esenciales para la aplicación del usuario, como producto final de este proceso se puede representar los datos reducidos en un plano cartesiano como un diagrama de dispersión (*scatter plot*) **Figura 3**.

2.3 REDUCCIÓN DE DIMENSIÓN

En muchas bases de datos, los vectores de datos medidos son de alta dimensión, pero generalmente es posible que la información se encuentra cerca de una variedad de dimensión inferior. En otras palabras, es posible que los datos de alta dimensión son mediciones múltiples, indirectas de una fuente subyacente, que típicamente no se puede medir directamente. Usando un método de reducción de dimensión adecuado es posible encontrar información relevante en un espacio de dimensión inferior, ignorando la información repetitiva y poco relevante existente en la base de datos original. Teniendo en cuenta lo anterior, el proceso de reducción de dimensión generalmente se aplica como parte de una etapa de preprocesamiento dentro del proceso DCBD debido a que puede mapear o proyectar los datos a un espacio en donde los datos originales son representados con menos atributos o características, con el fin, de mejorar tareas como la minería de datos y el reconocimiento de patrones. La capacidad de representar la base de datos original con dos o tres mediciones ayuda en gran medida a la representación de los datos con el fin de que sean fácilmente interpretables por parte del usuario [27], [29], [32].

La reducción de dimensión se basa en el concepto de extracción de características, el cual consiste en transformar los datos desde un espacio de alta dimensión hacia un espacio de menor dimensión, de tal manera que el espacio de características quede óptimamente reducido de acuerdo con un criterio de evaluación, cuyo fin es distinguir el subconjunto que representa mejor la base de datos original. La aplicación de métodos de reducción de dimensión puede generar una representación de los datos de alta dimensión originales en un espacio de dimensión menor, el cual es formado por una combinación lineal o no lineal de unos atributos dados.

Los datos iniciales corresponden a muestras u objetos representados en características o variables. La inclusión de un gran número de variables dentro del proceso de exploración de los datos puede incrementar costos y tiempo de procesado e incluso puede generar datos con información ruidosa e irrelevante que pueden ser eliminadas con un método de RD [29]. Como se ha mencionado antes, una forma intuitiva de visualizar datos es mediante gráficos 2-D o 3D, lo que resulta en una visualización natural e inteligible para los seres humanos. En este

sentido, la reducción de dimensión toma lugar, siendo una etapa importante en los sistemas de visualización de datos [16], [24]. Técnicamente, la reducción de dimensión (RD) tiene por objetivo alcanzar una representación de los datos dentro de un espacio de baja dimensión, sobre el cual, se puede mejorar el desempeño de las tareas de minería de datos, y a la vez hace que la representación de los datos, considerando su naturaleza intrínseca, sea más adecuada e inteligible para el ser humano [6].

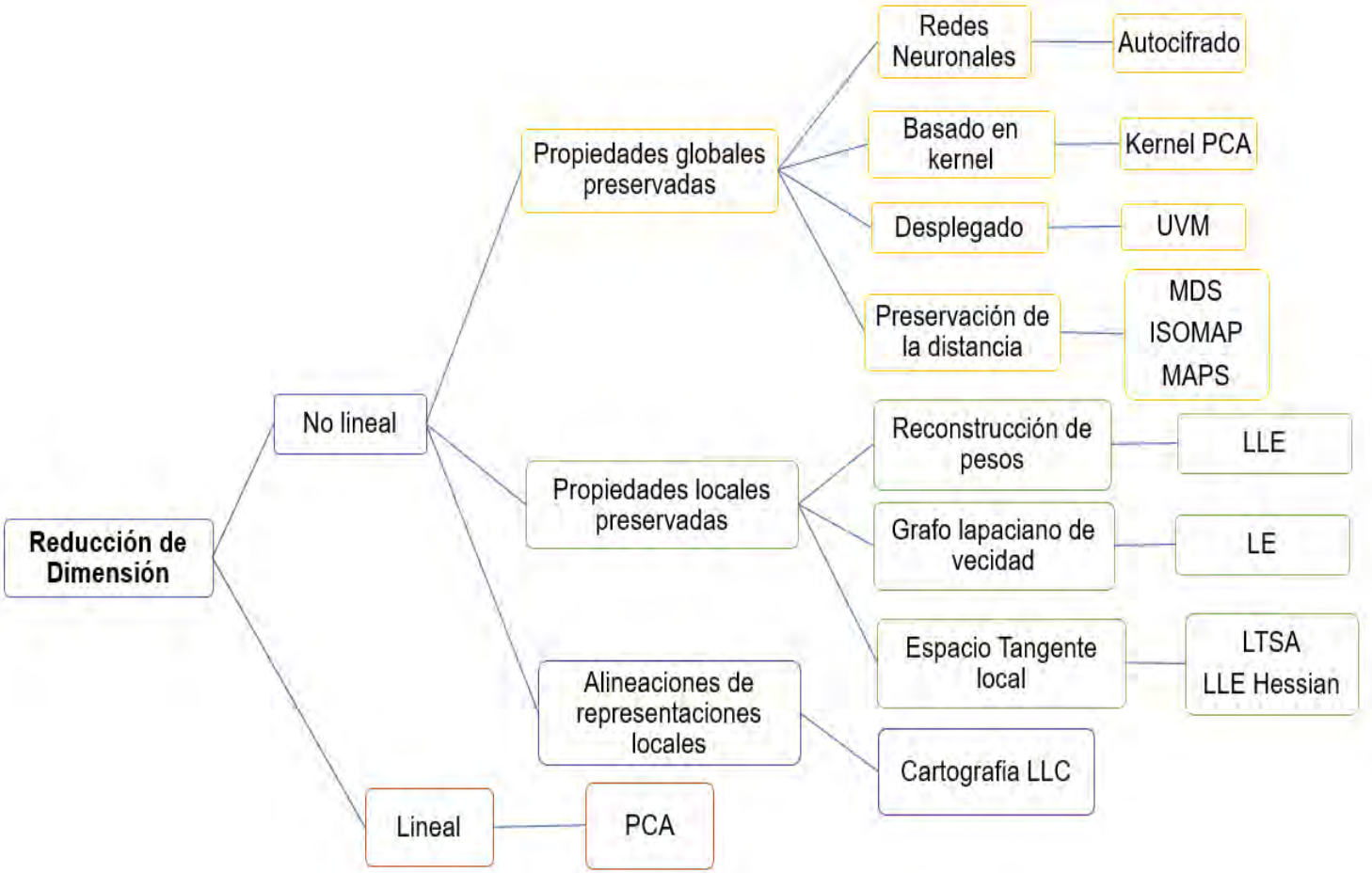


Figura 4. Clasificación de los diferentes métodos de reducción de dimensión existentes. En grandes rasgos como se indica en la figura los métodos de RD pueden clasificarse como métodos lineales y no lineales. **Fuente:** [32] adecuado en esta investigación.

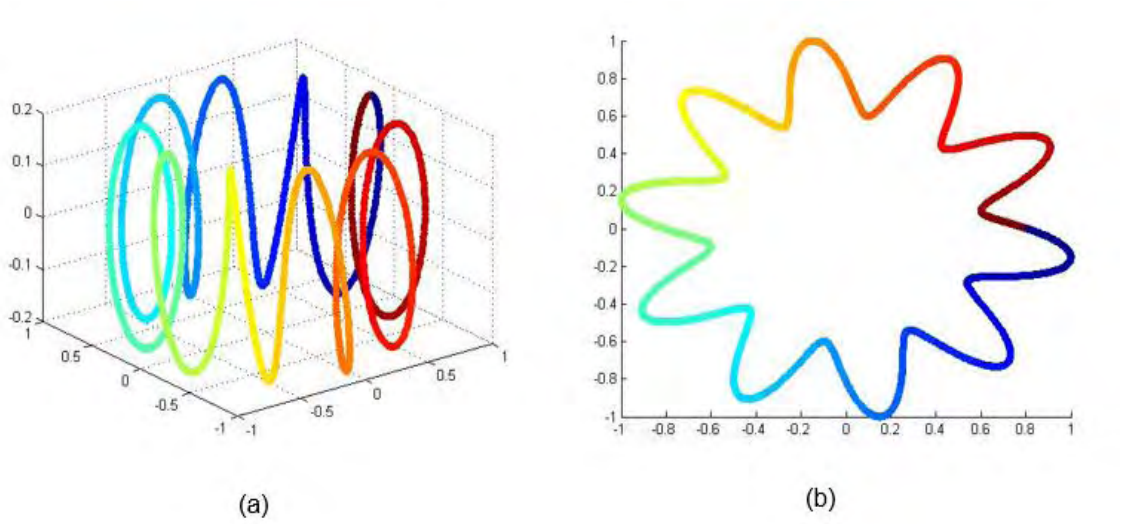


Figura 5. Reducción de dimensión de una toroide espiral (a) en 3D. en (b) se muestra el espacio embebido obtenido al aplicar el método LLE. **Fuente:** Esta investigación.

Por esta razón las técnicas RD se convierten en herramientas importantes de análisis que permiten aprovechar todo el conjunto de mediciones realizadas por cada registro. En la **Figura 4** se puede encontrar una clasificación general de los métodos

Dos formas populares de reducción de dimensión son los métodos análisis de componentes principales (PCA) [36] y escalamiento multidimensional (MDS) [37]. Tanto PCA como MDS son métodos de vectores propios diseñados para modelar variabilidades lineales en datos de alta dimensión. En PCA, se calcula las proyecciones lineales de mayor varianza de los vectores propios superiores de la matriz de covarianza de datos. En MDS clásico (o métrico), se calcula la incrustación de baja dimensión que mejor conserva las distancias de pares entre los puntos de datos. Las distancias corresponden a distancias euclidianas, los resultados de MDS métricos son equivalentes a PCA. Análisis de componentes principales se puede emplear de una manera no lineal por medio del truco del *kernel*. La técnica resultante es capaz de construir asignaciones no lineales que maximizan la varianza en los datos. La técnica resultante se titula *kernel* PCA.

Otras técnicas no lineales prominentes son *Locally Linear Embedding* (LLE) [38] (**Figura 5**) y *Laplacian Eigenmaps* (LE) [39]. LLE reduce la dimensión preservando las vecindades de cada muestra perteneciente a la base de datos en alta dimensión. LLE intenta descubrir estructura no lineal en datos de alta dimensión mediante la explotación de simetrías locales de reconstrucciones lineales. En particular, LLE mapea los datos dentro de un solo sistema de coordenadas global de menor dimensionalidad. Por otro lado, LE construye un grafo a partir de la

información de las vecindades que componen el conjunto de datos. Cada punto de datos sirve como un nodo en el grafo y la conectividad entre los nodos se rige por la proximidad de puntos vecinos (usando, por ejemplo, el algoritmo vecino *k*-más cercano). El grafo generado se puede considerar como una aproximación discreta del espacio de baja dimensión en el espacio de alta dimensión. La minimización de una función de costo basada en el grafo asegura que los puntos cercanos entre sí en el conjunto de datos original se mapean uno cerca del otro en el espacio de baja dimensión, preservando las distancias locales.

2.4 VISUALIZACIÓN INTERACTIVA DE DATOS

Los seres humanos son criaturas altamente visuales, y es que a pesar de que sus capacidades para detectar específicos patrones dentro de millones de datos no son las suficientes, un humano incluso en su temprana edad puede interpretar representaciones en un gráfico de barras o una representación cartesiana sencilla [11]. La visualización provee una descripción rápida, efectiva y clara de los datos, uno de sus principales objetivos es el representar de manera gráfica los datos para hacerlos más cercanos al usuario, de manera que este pueda interpretar algunas estructuras, patrones o anomalías presentes en ellos [40]. De lo anterior, se puede inferir que la visualización está estrechamente ligada a la minería de datos dada su capacidad para representar los patrones encontrados en una base de datos.

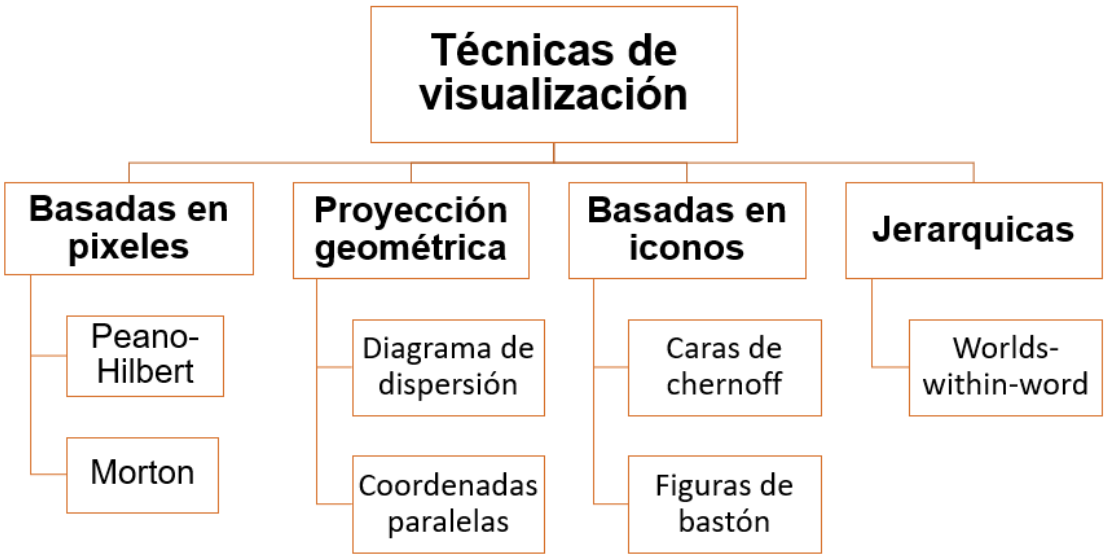


Figura 6. Clasificación de las principales técnicas de visualización clasificadas según el principio en el que se basan y al grupo al que perteneces. **Fuente:** esta investigación.

Actualmente existen muchas técnicas de visualización (**Figura 6**) quienes se basan en diferentes criterios y principios: a) **Técnicas basadas en píxeles**, mapean cada dimensión haciendo uso de los píxeles donde el nivel de intensidad

de este refleja el valor de la dimensión, por lo que esta técnica es netamente visual, esta técnica hace uso de curvas como la de *peano-hilbert* o *Morton* [42] para la representación de datos multidimensionales. b) **Técnicas de proyección geométrica,** estas no pueden representar datos de alta dimensión sin la ayuda de herramienta como la RD. Sin embargo, ayudan al usuario a encontrar agrupamientos o proyecciones interesantes dentro de datos multidimensionales [41], [43].

Un ejemplo de estas técnicas son el diagrama de dispersión (*scatter plot*) y las coordenadas paralelas. c) **Técnicas de visualización basadas en iconos,** mapean cada dato multidimensional con un pequeño icono. Dos ejemplos populares de estas técnicas son, las caras de *Chernoff* y las figuras de bastón [7]. d) **Técnicas de visualización Jerárquicas,** dividen el espacio de n-dimensiones en subconjuntos (subespacios) de tal forma que los subespacios puedan ser visualizados de manera jerárquica. Uno de los métodos más representativos de visualización jerárquico es el conocido como "*Worlds- within-Worlds*," o *n-Vision* [44], [35].

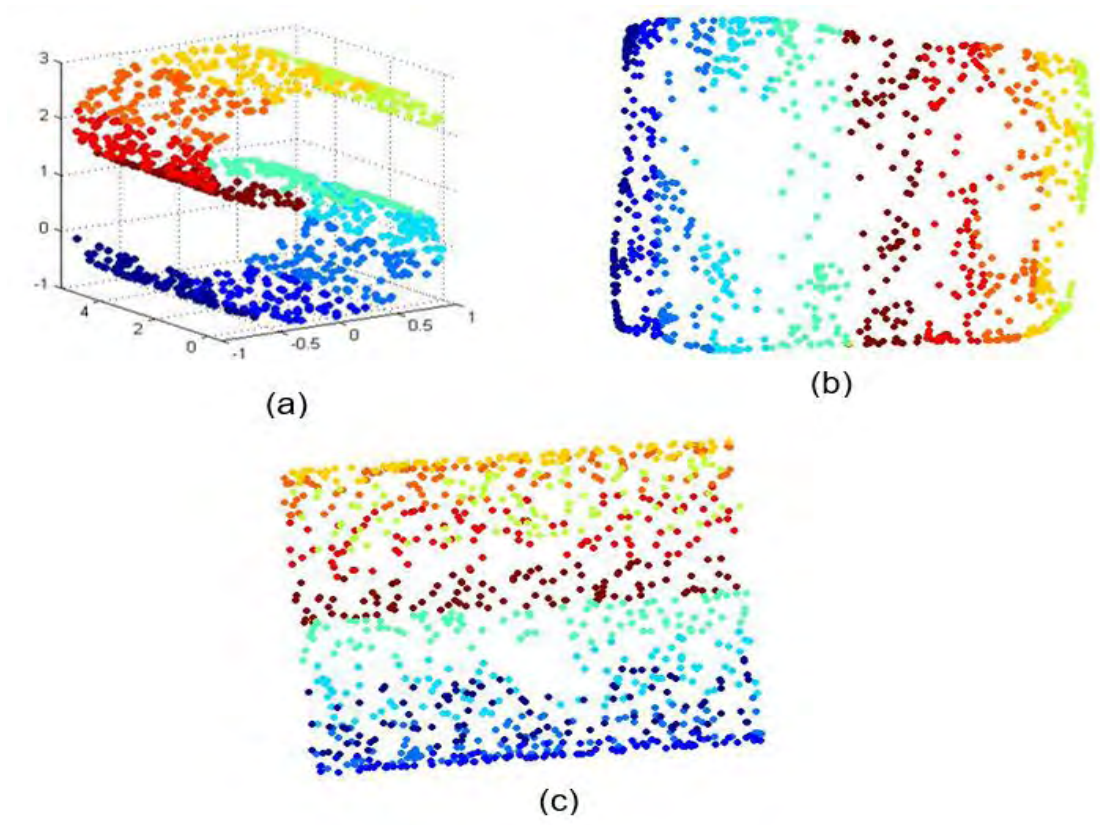


Figura 7. Representación obtenida luego de aplicar dos métodos de reducción uno de ellos conocido como análisis de componentes principales (PCA) (c) y el segundo *Locally linear embedding* (LLE) (b), aplicado a la base de datos de la letra S en 3D. En este ejemplo se puede apreciar como un espacio se redujo a dos

dimensiones, en donde es más fácil analizar e interpretar la base de datos **Fuente:** Esta investigación

Cada vez la visualización de datos toma mayor importancia en las diferentes disciplinas de la vida diaria y es por eso que, se han desarrollado un amplio número de técnicas para extraer estructuras o información más compleja de los datos. Sin embargo, emergen dos problemáticas en la representación de grandes volúmenes de datos: la primera de ellas es el tiempo y recursos computacionales que deben ser invertidos en la manipulación de los datos, lo cual no es tan conveniente si se plantea un análisis interactivo de grandes conjuntos de datos. El segundo, radica en que los sistemas actuales de representación carecen de herramientas efectivas para representar estos datos de manera que se evite la superposición de estos en un gráfico, es decir, en algunos casos existen demasiados registros para graficar, haciendo la visualización demasiado densa para ser útil para el usuario [9], [15], [16]. De lo anterior, surge la reducción de dimensión como técnica de visualización que permite representar bases de datos de alta dimensión, con menor número de atributos que reúna en gran medida la información de los datos originales [45].

Como bien se sabe datos en alta dimensión no pueden ser representado con herramientas que se limitan a espacios más reducidos como lo es el diagrama de dispersión (**figura 3**) que básicamente representa hasta 4 dimensiones incluyendo colores o figuras agregadas. Si bien, las técnicas nombradas anteriormente pueden ser usadas para representar mayor número de atributos, los resultados gráficos que se obtienen pueden no ser lo suficientemente entendibles para un usuario inexperto. En consecuencia, se podrían presentar ambigüedades al momento de la interpretación de una base de datos [8], [45]. De este modo, la reducción de dimensión como técnica de visualización se convierte en una gran herramienta que puede sacar provecho de todas las mediciones (características) que son realizadas a los registros, eliminando los datos redundantes y poco trascendentes con el fin de que puedan ser representados en un plano cartesiano bidimensional o tridimensional que representa de manera concreta la naturaleza de los datos y de esta manera posibles tendencias o patrones (**Figura 7**) [8].

2.5 TÉCNICAS DE OPTIMIZACIÓN PARA REDUCIR EL COSTE COMPUTACIONAL

Muchos métodos importantes en informática científica, incluyendo aplicaciones y técnicas relacionadas con el aprendizaje automático, se basan en resolver descomposiciones espectrales, lo que significa encontrar valores propios y vectores propios de matrices. Uno de los problemas clave con estos enfoques es que no existe una solución de forma cerrada para identificar valores propios y vectores propios. En su lugar, debemos confiar en los algoritmos, la mayoría de los cuales terminan explotando aproximaciones numéricas a través de esquemas

iterativos. Dado que hay una tendencia en el campo de aprendizaje automático para trabajar con conjuntos de datos cada vez más grandes, que se caracteriza por un alto número de características, muchos eigen-algoritmos iterativos se convierten en víctimas de inestabilidades numéricas. Cálculos numéricos inestables pueden conducir a tasas de convergencia más lentas, lo que puede significar que el procedimiento iterativo podría tomar más tiempo para converger a una solución numérica óptima. En el peor de los casos, tal inestabilidad produce una solución que simplemente es incorrecta. Otro problema clave para resolver grandes eigen-sistemas, es que intrínsecamente trabajan con matrices grandes y los recursos computacionales necesarios escalan exponencialmente con las dimensiones de las matrices. En otras palabras, cuanto más grande es la matriz, más alta será la complejidad espacial y temporal resultante.
Las bases de datos que se pueden procesar son muy grandes, por lo tanto, es necesario implementar un método de optimización que permita reducir el tiempo de procesamiento. Para alcanzar este objetivo se estudiaron cuatro (4) métodos. Dos (2) métodos que se basan en el paralelismo computacional y dos (2) métodos matemáticos que dan una solución aproximada a la descomposición espectral.

2.5.1 Métodos basados en paralelismo computacional

La computación paralela permite que muchas instrucciones se ejecutan simultáneamente. Se basa en el principio de que los problemas grandes se pueden dividir en partes más pequeñas que pueden resolverse de forma concurrente (“en paralelo”). Actualmente hay una tendencia por el uso de GPU para el procesamiento de grandes volúmenes de información, ya que tienen buenos resultados en cuanto a la computación paralela. Las GPU están diseñadas originalmente para renderizar gráficos por ordenador a altas velocidades de cuadros, lo que en última instancia significa que están optimizados para máximo rendimiento paralelo. Esto resulta ser muy útil para calcular algoritmos que se basan en operaciones algebraicas tales como productos vectoriales y matriciales, que se pueden calcular de manera muy eficiente en un esquema paralelo.

Es importante destacar que muchos eigen-algoritmos se basan en el cálculo de dichos productos vectoriales y matriciales. Sin embargo, implementar algoritmos en una GPU es un proceso mucho más complicado en comparación con trabajar con una CPU normal. Para la implementación de un eigen-algoritmo en una GPU, se estudió el Algoritmo simétrico QR con permutaciones [31], el cual aborda el problema relacionado con la inestabilidad numérica mediante el reordenamiento óptimo de filas y columnas de matrices en cada iteración del clásico algoritmo QR, explicado más claramente en [46], [47], [48].

Por otro lado, se encuentra un método más clásico que permite paralelizar la eigen-descomposición, como lo es el método Jacobi. La base del método consiste en construir una sucesión convergente definida iterativamente. El límite de esta

sucesión es precisamente la solución del sistema. A efectos prácticos si el algoritmo se detiene después de un número finito de pasos se llega a una aproximación al valor de la solución del eigen-sistema, lo cual se explica más claramente en [49].

2.5.2 Métodos matemáticos con soluciones aproximada

Un novedoso método que optimiza la descomposición espectral multi-escala es MSEIGS, explicado en profundidad en [50], el cual emplea la agrupación gráfica para dividir el gráfico en varios clusters que son manejables en tamaño y permiten un rápido cálculo de la eigen-descomposición por métodos estándar. Los vectores propios obtenidos de cada subespacio se combinan para inicializar el algoritmo Lanczos, el cual se utiliza para calcular la eigen-descomposición del espacio original [51]. Además, este método permite calcularse de manera paralela y muestra mejoras significativas cuando se ejecuta de este modo.

Por otro lado, se estudió el método sub-matrices localmente lineales (-LLL- por sus siglas en inglés-*Locally Linear Landmarks*-) explicado en [14], el cual aproxima la solución de la descomposición espectral resolviendo el sistema para un grafo más pequeño. El subconjunto de puntos (puntos de referencia) se toman de la base de datos original. Los puntos de referencia se aplican a la fórmula de Nystrom [52], [53] para estimar los vectores propios sobre todos los puntos de la base de datos. Esto tiene el problema de que las afinidades entre los puntos de referencia no se benefician de los puntos restantes y pueden representar superficialmente los datos si se usan muy pocos puntos de referencia. Por lo tanto, se utiliza todos los puntos de datos al restringir la proyección latente de cada punto, para ser una función lineal local de las proyecciones latentes de los puntos de referencia. Esto construye una nueva matriz de afinidad entre puntos de referencia que conserva una estructura múltiple incluso con pocos puntos de referencia, permite reducir el tamaño del eigen-problema y define un mapeo rápido, no lineal fuera de la muestra.

3. METODOLOGÍA

A lo largo de este trabajo se ha hablado sobre la importancia de reducción de dimensión como parte de la visualización interactiva de datos siendo estas etapas del proceso DCBD y la minería de datos. Existe la necesidad latente de acoplar de brindar al usuario herramientas que le ayuden a explorar cualquier repositorio de datos, además que este haga parte del proceso de visualización y por ende reducción de dimensión sin que él tenga conocimiento amplio de este tipo de procedimientos. La reducción de dimensión parte de una base de datos $Y \in \mathbb{R}^{D*N}$ donde D hace referencia al número de atributos con los que se cuenta y N el número de registros queriendo mapear estos en un espacio de baja dimensión $X \in \mathbb{R}^{d*N}$ donde $d \ll D$ de modo que el espacio embebido obtenido y los datos originales están compuestos por el mismo número de puntos o registros, lo que quiere decir que una base de datos de alta dimensión puede ser proyectada en una menor atributos sin que el número de muestras originales se vea afectado.

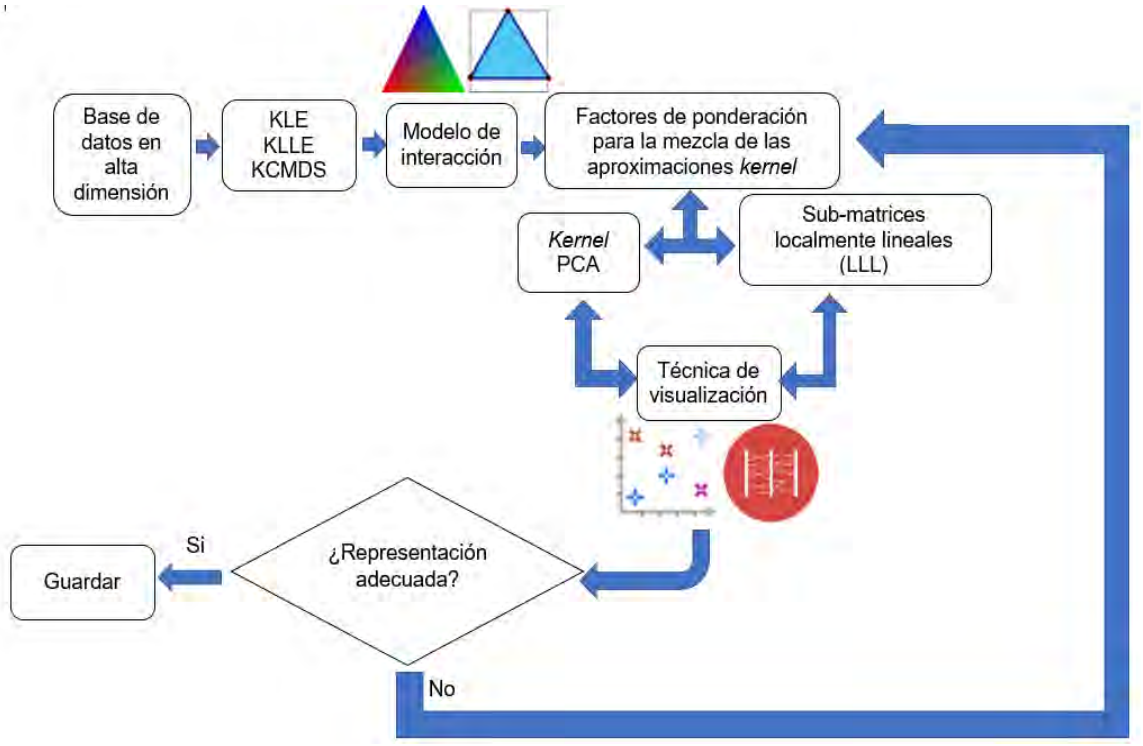


Figura 8. Esquema general de la metodología utilizada para el proceso de visualización interactiva donde se fusionan métodos de interacción y técnicas de visualización **Fuente:** Esta investigación.

En este trabajo se brinda un amplio rango de visualización de espacios embebidos ya que se pueden representar hasta 10 atributos o dimensiones, según la necesidad del usuario, estos se representarán bien sea en un diagrama de

dispersión o con coordenadas paralelas. Como se explicará en las secciones siguientes esta investigación hace uso de representaciones kernel de métodos espectrales para la reducción de dimensión y dos modelos de interacción uno cromático y el segundo basado en la medida de los ángulos de un triángulo, estos permitirán que el usuario utilice los métodos según las necesidades. Para el cálculo del espacio embebido resultante el usuario puede hacer uso de dos métodos que serán explicados con mayor profundidad, uno de ellos es la representación kernel del método de análisis de componentes principales (PCA) y el segundo usa una porción de los datos originales dando una aproximación del espacio embebido obtenido, pero en un tiempo de ejecución menos que kernel PCA. En la **figura 8** se muestra el esquema general de la metodología propuesta para la herramienta de visualización.

3.1. MÉTODOS DE REDUCCIÓN DE DIMENSIÓN CON APROXIMACIONES KERNEL

Los métodos de RD espectrales han sido ampliamente utilizados [54], [55], puesto que presentan la versatilidad de ser representados mediante matrices *kernel* que representan funciones de distancia asociadas a un método de reducción de dimensión en particular. Para el desarrollo de la metodología de visualización propuesta se tienen en cuenta tres métodos espectrales de reducción de dimensión llamados: *Classical Multidimensional Scalling* (CMDS), *Locally linear Embedding* (LLE), y *Laplacian Eigenmaps* (LE), los cuales son ampliamente explicados en [31], [56].

La representación *kernel* para el método de reducción CMDS se define como la matriz de distancia $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ doblemente centrada, es decir haciendo que la media de las filas y las columnas sea cero, así:

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

donde la ij-esima entrada de $\boldsymbol{D}$ es dada por la distancia euclidiana $d_{ij} = \|y_i - y_j\|_2^2$.

El *kernel* para el método LLE puede ser aproximado a partir de la forma cuadrática en términos de la matriz $\boldsymbol{\mathcal{W}}$ con coeficientes lineales que suman 1 y pueden de manera óptima reconstruir los datos originales. Sea una matriz $\mathbf{M} \in \mathbb{R}^{N\times N}$ definida por la expresión $\boldsymbol{M} = (\boldsymbol{I}_N\text{-}\boldsymbol{\mathcal{W}})\ (\boldsymbol{I}_N\text{-}\boldsymbol{\mathcal{W}}^T)$ y λ_{max} como el valor propio más grande de $\mathbf{M}$:

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

Debido a que *kernel* PCA es un problema de maximización de la covarianza de alta dimensión representada por un *kernel*, LE puede ser representado como la matriz pseudo-inversa del grafo $\boldsymbol{L}$, como se muestra en la siguiente expresión.

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

Donde $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S}$, tal que $\boldsymbol{S}$ es una matriz de disimilitud y $\boldsymbol{D}$ =*Diag* $(\boldsymbol{S} * \mathbf{1}_N)$ la matriz del grado de $\boldsymbol{S}$. La matriz de similitud $\boldsymbol{S}$ está formada de tal manera que el

parámetro del ancho relativo se estima manteniendo la entropía de la distribución con el vecino más cercano con aproximadamente $\log(k)$, donde k es el número dado de vecinos como se explica en [57]. El número de vecinos se establece como el entero más cercano al 15% de la cantidad de datos.

3.2. KERNEL PCA

Como se ha dicho anteriormente, los métodos de RD tienen como objetivo encontrar a partir de una matriz $\mathbf{Y} \in \mathbb{R}^{D\times N}$, un espacio embebido $\mathbf{X} \in \mathbb{R}^{d\times N}$ con $d < D$ que preserve la estructura o propiedades de $\mathbf{Y}$ tanto como sea posible bajo un criterio establecido [27], [29]. El método de RD conocido como análisis de componentes principales (PCA), es una proyección lineal que intenta preservar la varianza a partir de los valores y vectores propios de la matriz de covarianza [28], [32]. Además, cuando una matriz de datos es centrada, es decir que el valor medio de las filas (características) es igual a cero, la preservación de la varianza puede ser vista como una preservación del producto interno euclidiano [28].

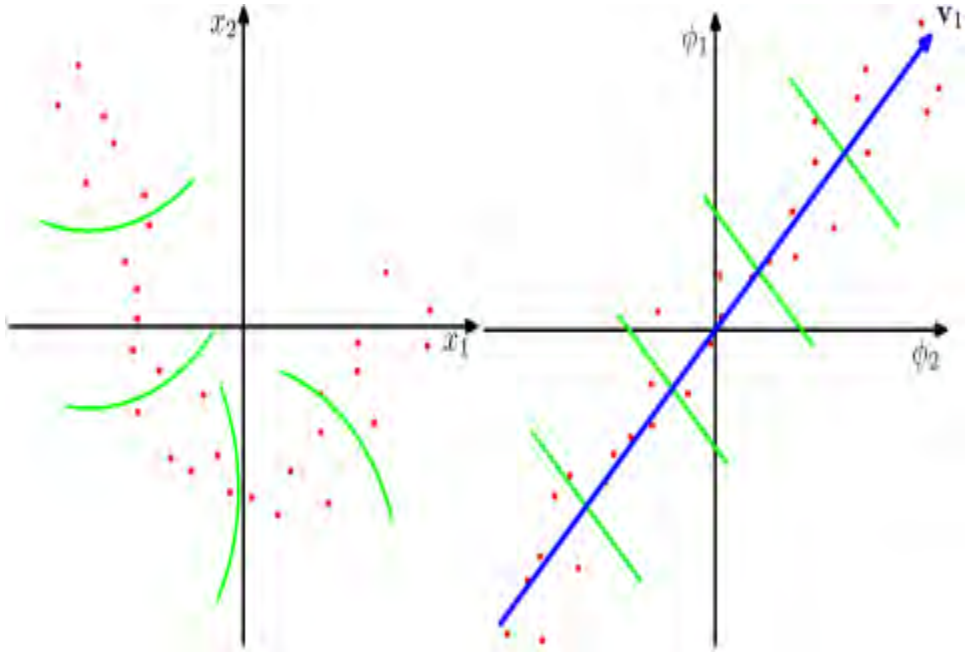


Figura 9. Ejemplo ilustrativo de *kernel* PCA. **Fuente:** [58]

En la **Figura 9** se observa a la izquierda puntos de datos que se encuentran principalmente a lo largo de una curva en 2D. PCA no puede reducir la dimensionalidad de dos a uno, porque los puntos no se encuentran a lo largo de una línea recta. Aun así, los datos están tentativamente ubicados alrededor de una curva no lineal. Por lo tanto, para este caso PCA no tendría un buen desempeño. *Kernel* PCA puede encontrar esta variedad no lineal, mapeando los datos en un espacio de dimensiones superiores. Esto puede parecer una contradicción, pero no lo es. Los datos están mapeados en un espacio de dimensiones superiores,

para luego encontrar un subespacio dimensional inferior. Entonces se aumenta la dimensionalidad para poder disminuirla. La esencia del "truco del *kernel* " es que no se necesita considerar explícitamente el espacio de mayor dimensión, por lo que este salto potencialmente confuso en la dimensionalidad se realiza completamente encubierto. A continuación, se explica el procedimiento que realiza *kernel* PCA a través de un pseudo-código:

Pseudo-código para implementación de *Kernel* PCA

1. **Calcular:** Calcula la matriz aproximada $\boldsymbol{K} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}}$**,** según el *kernel* deseado.
2. **Ajustar:** Centraliza la matriz $\boldsymbol{K}$
3. **Descomposición espectral:** Realiza la descomposición de $\boldsymbol{K}$ en valores y vectores propios.
4. **Organizar:** Organiza los valores propios en orden descendente.
5. **Seleccionar:** Selecciona los $\boldsymbol{d}$ vectores propios correspondientes a los más grandes valores propios obtenidos.
6. **Representación en baja dimensión:** Se organizan los vectores propios seleccionados en una matriz $\mathbf{X} \in \mathbb{R}^{\boldsymbol{d}\times\boldsymbol{N}}$

3.3 MODELOS DE INTERACCIÓN DE MÉTODOS RD PARA LA VISUALIZACIÓN INTERACTIVA

La visualización interactiva involucra al usuario con los métodos RD directamente y dado que entender el trasfondo del funcionamiento de estos puede ser tedioso y muchas veces no exitoso ya que se hace necesario de usuarios expertos para que ellos manipulen adecuadamente los métodos y lograr así la visualización deseada. Por lo anterior buscar una herramienta de interacción entre el usuario y los métodos de RD, sin que sea necesario un previo conocimiento, se hace indispensable para el objetivo principal de la visualización. La función principal de estos modelos es el integrar al proceso de reducción y visualización al usuario dando la libertad de que este escoja la combinación de métodos más adecuada. De manera general, cualquier conjunto de métodos se puede representar como un conjunto de métodos de RD $\{\boldsymbol{f}_1, \ldots\ldots, \boldsymbol{f}_{\boldsymbol{M}}\}$ de tal forma que *M* es el número de métodos considerados por el usuario.

En esta investigación se planteó que los parámetros a ser combinados son las matrices kernel obtenidas de las representaciones de los métodos espectrales de RD presentadas anteriormente, cada matriz corresponde a cada uno de los *M* métodos de RD considerados, esto es $\{\boldsymbol{K}^{(1)}, \ldots\ldots, \boldsymbol{K}^{(\boldsymbol{M})}\}$. Por consiguiente, se obtiene una matriz kernel $\widehat{\boldsymbol{K}}$ resultante de la mezcla de las M matrices kernel, tal

que:

$$\widehat{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(M)} \,, \tag{4}$$

definiendo a α_m como el factor de ponderación correspondiente al método *m* y a $\alpha = \{\alpha_1, \ldots\ldots, \alpha_M\}$ como el vector de ponderación estos parámetros serán definidos por los modelos de interacción, los cuales de manera diferente e independiente establecen el valor de dichos factores. A continuación, se presenta los dos modelos de interacción utilizados en la metodología de la herramienta, se explicará cómo cada uno de ellos selecciona dichos factores de ponderación para la mezcla de métodos RD.

3.3.1 MODELO CROMÁTICO

Este modelo descrito en [59] está basado en el espacio de color RGB y trabaja de la siguiente manera: Una imagen normalizada definida como una matriz descrita por la función $\boldsymbol{I} : N^3 \rightarrow [0,1]$, donde par de números $x, y : N^2$ son conocidos como pixeles y cada valor de intensidad $I(x, y, c)$ es asociado a un pixel (x, y) del canal c [61]. La descomposición es asociada con los valores de intensidad de los canales, los cuales están entre 0 y 1 (si la imagen está normalizada), de modo que un valor de 0 indica la completa ausencia de color (color negro) y el valor 1 es relacionado con el máximo valor de intensidad (color blanco) [60].

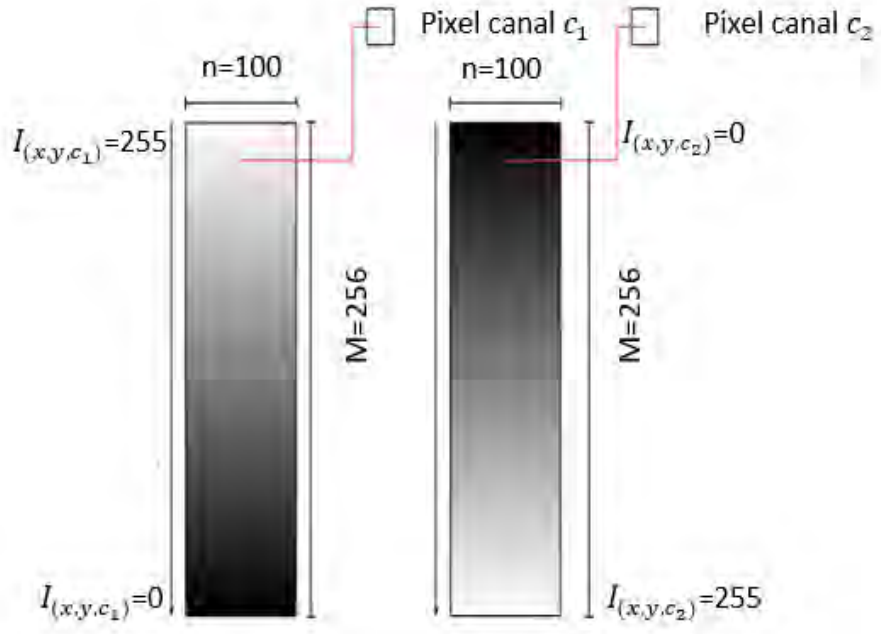
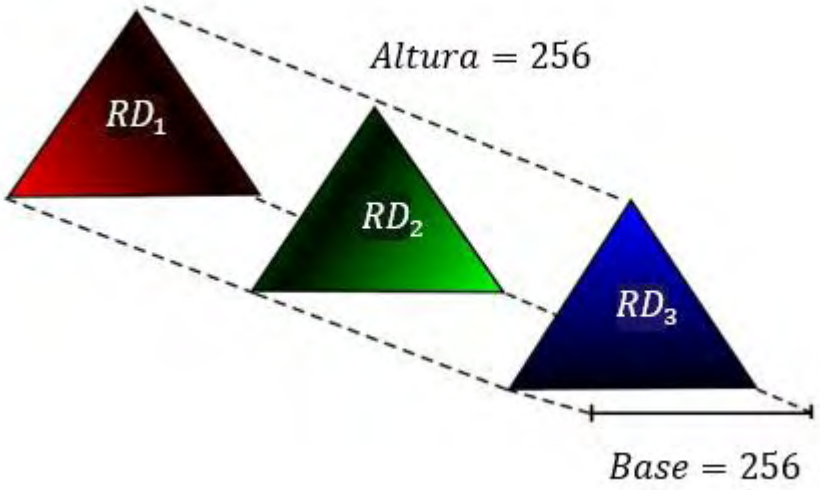
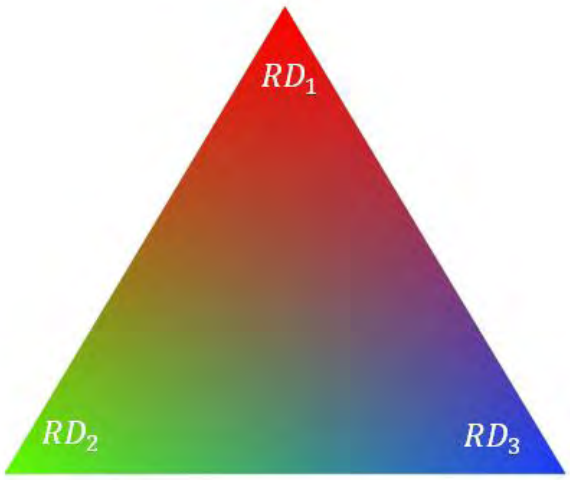


Figura 10. Imagen de dos canales, con un valor de intensidad descendente para el canal 1 (c1) y un valor ascendente para el canal 2 (c2) de tal forma que la suma

de intensidad de ambos canales sea igual a 255. **Fuente:** [59]



(a) Superposición de canales



(b) Modelo cromático

Figura 11. A partir de la superposición de los tres canales (a) se tiene como resultado el modelo cromático propuesto en [41]. (b) en donde los vértices

representarán un método de RD en particular o una combinación de métodos si se elige un punto dentro de la superficie. **Fuente:** [59].

Este modelo aprovecha dos propiedades de las imágenes RGB: La resolución espacial quien se define como el número de píxeles que una imagen tiene, y puede ser calculo por: $píxeles = m * n$ donde m representa el número de filas y n el número de columnas (el ancho y alto de la imagen).por otro lado la resolución de intensidad es el rango de valores de intensidad que cada pixel puede tener, para esta investigación la resolución es de 8 bits es decir $2^8 - 1 = 255$ (**Figura 10**) debido a que el valor de intensidad 0 es considerado [60].

En el espacio de color RGB existen tres canales $c_1 = R, c_2 = G, c_3 = B$ donde cada canal representa un método de reducción de dimensión, Sin embargo, cada píxel de la imagen tendrá tres valores de intensidad debido a sus tres canales. Además, si los valores de intensidad de los dos canales son sumados el resultado siempre será igual a 255 (1 si una normalización es realizada). De forma que, el color rojo, color verde y color azul representan respectivamente los métodos RD_1, RD_2, RD_3 (**Figura 11**). Sin embargo, cada canal tiene diferente dirección de cambio de intensidad, este modelo permite al usuario escoger múltiples combinaciones de métodos de RD como rango de colores de una combinación escogida.

3.3.2 MODELO BASADO EN ÁNGULOS

Como este modelo se basa en la geometría de un triángulo, se hace uso del teorema de ángulo externo e interno de la geometría euclídea, el cual enuncia que todo ángulo exterior de un triángulo es igual a la suma de los dos ángulos interiores no adyacentes $\sphericalangle D = \sphericalangle B + \sphericalangle A$. En la **Figura 12** se puede evidenciar intuitivamente que $\sphericalangle D + \sphericalangle C = 180$ y por consiguiente $\sphericalangle C + \sphericalangle A + \sphericalangle B = 180°$.

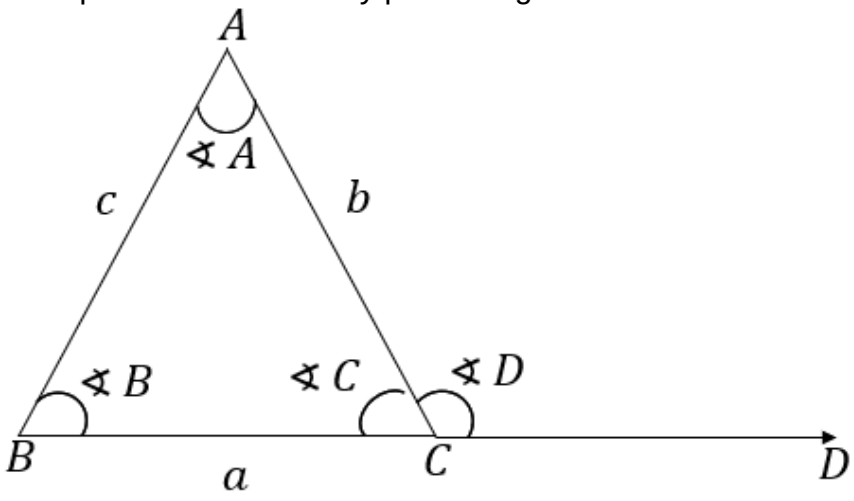


Figura 12. Ilustración de la Proposición 1.16 en los Elementos de Euclides. **Fuente:** Esta investigación

Teniendo claro cómo funciona la geometría de un triángulo se puede establecer que el modelo propuesto en esta investigación hace uso de un triángulo inscrito en un cuadrado **figura 13** donde cada uno de los vértices corresponde a una aproximación kernel de los métodos espectrales RD implementados en este trabajo, el usuario tiene la liberta de mover cada uno de los vértices que se diferencian con tres esferas de diferentes colores alrededor del cuadro cambiando así la medida de los ángulos y por consiguiente el factor de ponderación de cada uno de los métodos. 180° hace referencia a 1 en los factores por lo que se muestra el porcentaje que corresponde a cada método en la mezcla de sus aproximaciones kernel.

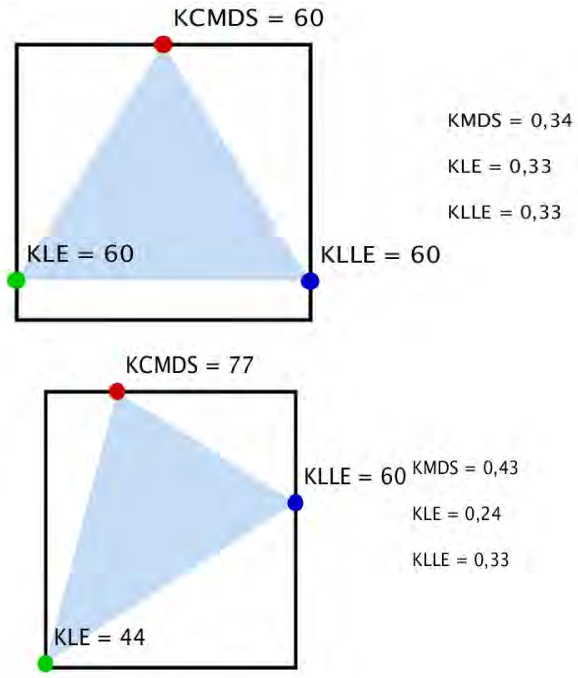


Figura 13. Modelo de interacción propuesto basado en la medida de los ángulos de un triángulo inscrito en un cuadrado. **Fuente:** Esta investigación

3.4 SUBMATRICES LOCALMENTE LINEALES

Un aporte importante en este trabajo de grado fue la implementación de un algoritmo de optimización del tiempo utilizado por la herramienta en realizar la descomposición espectral de la combinación de las matrices *kernel* con el fin de que el usuario al buscar la reducción de dimensión adecuada según su criterio, por medio de los modelos de interacción observe cambios en tiempo real en el espacio embebido resultante haciendo que la herramienta sea realmente interactiva.

Matemáticamente el problema tiene como entradas una matriz simétrica positiva semidefinida A_{NxN} (matriz kernel para este caso), que contiene la información sobre la similitud entre pares de puntos en la matriz de datos $Y \in \mathbb{R}^{DxN}$, y una matriz simétrica positiva definida B_{NxN} que por lo general ajusta la escala de la solución. Dadas estas matrices, se busca una solución $X \in \mathbb{R}^{d_{xN}}$ al siguiente problema espectral generalizado:

$$min_X \, \boldsymbol{tr}(\boldsymbol{XAX}^T) \;\; s.t. \;\; \boldsymbol{XBX}^T = \boldsymbol{I} \tag{5}$$

La solución al problema espectral (5) está dado por $\boldsymbol{X} = \boldsymbol{U}_d^T \boldsymbol{B}^{-\frac{1}{2}}$, donde $\boldsymbol{U}_d = (\boldsymbol{u_1}, \ldots, \boldsymbol{u_d})$ son los últimos vectores propios de la matriz $\boldsymbol{C} = \boldsymbol{B}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{B}^{-\frac{1}{2}}$.

En bases de datos donde *N* es la cantidad de puntos es muy grande, se presenta alto coste computacional en el proceso de obtener la matriz $\boldsymbol{U}_d$. En [14] se propone un método llamado "*Locally Linear Landmarks*" (LLL), que se basa en la idea de seleccionar un subconjunto de $L << N$ puntos de referencia, obteniendo el conjunto $\widetilde{\boldsymbol{Y}} \in \mathbb{R}^{D \times L}$, con el propósito de aproximar los datos a través de una estructura globalmente no lineal pero localmente lineal alrededor de estos puntos de referencia, y luego restringir la solución X a seguir esta estructura localmente lineal. El mapeo lineal local está dado por una matriz de proyección $\boldsymbol{Z} \in \mathbb{R}^{L \times N}$ que satisface

$$\boldsymbol{Y} = \widetilde{\boldsymbol{Y}} \boldsymbol{Z}, \tag{6}$$

en el espacio de alta dimensión, y al imponerlo en el espacio de baja dimensión, se puede expresar el problema (5) como un nuevo problema espectral con un número menor de variables L.

La matriz de proyección $\boldsymbol{Z}$ se construye estableciendo el parámetro Kz previamente. Kz establece el número de vecinos más cercanos pertenecientes al subconjunto de puntos de referencia (*landmarks*) seleccionados con respecto a cada una de las muestras que conforman la base de datos original. Según [14] Kz se recomienda establecerla dentro del intervalo [d+1, D+1]**.** A continuación, se obtiene una relación de distancia entre la muestra o individuo en cuestión y sus Kz vecinos más cercanos, con el fin de tener una proyección de los datos más aproximada a la base de datos original, construyendo así la matriz $\boldsymbol{Z}$.

La suposición fundamental en el método LLL es que la dependencia local de los puntos con respecto a los puntos de referencia (*landmarks*) que ocurre en alta dimensión (6), se preserva en el espacio embebido:

$$\boldsymbol{X} = \widetilde{\boldsymbol{X}} \boldsymbol{Z}, \tag{7}$$

sustituyendo (7) en el problema espectral (5) se obtiene el siguiente problema espectral reducido:

$$min_{\tilde{X}} \, tr(\widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{X}}^T) \;\; s.t. \;\; \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{X}}^T = \boldsymbol{I}, \tag{8}$$

donde las matrices:

$$\widetilde{\boldsymbol{A}} = \boldsymbol{ZAZ}^T, \qquad \widetilde{\boldsymbol{B}} = \boldsymbol{ZBZ}^T . \tag{9}$$

Las matrices descritas en (9) serán por lo tanto de dimensiones $\mathbb{R}^{LxL}$. La solución para el problema reducido es $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}_d^T \widetilde{\boldsymbol{B}}^{-\frac{1}{2}}$, donde $\widetilde{\boldsymbol{U}}_d$ son los últimos vectores propios de la matriz $\widetilde{\boldsymbol{C}} = \widetilde{\boldsymbol{B}}^{-\frac{1}{2}} \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{B}}^{-\frac{1}{2}}$. Después de encontrar la solución para los

puntos de referencia, los valores de $\boldsymbol{X}$ pueden ser recuperadas aplicando la ecuación (7) de nuevo.
A continuación, se explica paso a paso la implementación del método LLL a través de un pseudo-código:

Pseudo-código para implementación de LLL

1. **Definir:** Defina los parámetros $\boldsymbol{Kz}$ en el intervalo [d+1, D+1] y $\boldsymbol{L}$ como la cantidad de puntos de referencia (landmarks).
2. **Seleccionar:** Selecciona un subconjunto de muestras $\boldsymbol{L}$ que represente la base de datos original.
3. **Calcular:** Calcula la matriz de proyección $\boldsymbol{Z}$ basándose en $\boldsymbol{Kz}$.
4. **Calcular:** Calcula las matrices de similitud reducidas $\widetilde{\boldsymbol{A}}\ \boldsymbol{y}\ \widetilde{\boldsymbol{B}}$ en ecuación (9).
5. **Calcular:** Calcula la matriz $\widetilde{\boldsymbol{C}} = \widetilde{\boldsymbol{B}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{B}}^{-\frac{1}{2}}$.
6. **Descomposición espectral:** Realiza la descomposición de $\widetilde{\boldsymbol{C}}$ en valores y vectores propios.
7. **Organizar:** Organiza los valores propios en orden ascendente.
8. **Seleccionar:** Selecciona los $\boldsymbol{d}$ vectores propios correspondientes a los más pequeños valores propios obtenidos.
9. **Agrupar:** Se organizan los vectores propios seleccionados en una matriz $\widetilde{\boldsymbol{U}}_{\boldsymbol{d}} \in \mathbb{R}^{\boldsymbol{Lxd}}$.
10. **Calcular:** Calcula la solución $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}_{\boldsymbol{d}}^{\boldsymbol{T}}\widetilde{\boldsymbol{B}}^{-\frac{1}{2}}$, al problema espectral reducido de la ecuación (21).
11. **Representación en baja dimensión:** Aplica la ecuación (7) para obtener el espacio embebido $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{dxN}}$.

3.5 MEDIDA DE CALIDAD

Es importante conocer y evaluar la reducción de dimensión que se realiza por medio de los métodos RD mencionados y la combinación de estos, Para este propósito se utiliza un criterio de calidad mediante la conservación de los k-ésimos vecinos más cercanos desarrollada en [13]. El objetivo de integrar esta medida de calidad como parte de la metodología de desarrollo es cuantificar el desempeño de los métodos RD en cuanto a la preservación de la topología de los datos en el espacio de baja dimensión con respecto al de alta dimensión.

Tal medida es ampliamente aceptada como una medida adecuada no supervisada [1] [63], quien permite evaluar el espacio embebido de la siguiente manera: El rango de $\mathcal{E}_j$ respecto a $\mathcal{E}_i$ en el espacio de alta dimensión se denota como: $\boldsymbol{pij} = |\{k: \delta ik < \delta ij\ o\ (\delta ik = \delta ij\ y\ 1 \leq k < j \leq N)\}|$ donde $|\cdot|$ denota la cardinalidad del conjunto. Similarmente, en [13] define que el rango de xj respecto a xi en el espacio de baja dimensión es: $\boldsymbol{rij} = |\{k: dik < dij\ o\ (dik = dij\ y\ 1 \leq k < j \leq N)\}|$. Los k-ésimos vecinos de $\boldsymbol{\xi i}$ y de $\boldsymbol{xi}$ son los conjuntos definidos $\boldsymbol{v_i}^{\boldsymbol{k}} = \{j: 1 \leq pij < K\}$ y $\boldsymbol{n_i}^{k} = \{j: 1 \leq rij < K\}$, respectivamente. Un primer índice de rendimiento puede ser denotado como:

$$\mathrm{Q}_{NX}(\mathrm{K}) \;=\; \sum_{i=1}^{N} \frac{|\boldsymbol{v_i}^k \cap \boldsymbol{n_i}^k|}{KN} = 1 \qquad (10)$$

La ecuación (10) resulta en valores comprendidos entre 0 y 1 y mide el promedio normalizado de acuerdo con los *k*-ésimos vecinos correspondientes entre los espacios de alta dimensión y baja dimensión. Definiendo de esta manera una matriz de co-clasificación $[Q = q_{NX}]$, $j \leq N-1$ con $q_{kl} = |\{(i,j): p_{ij} = k \text{ y } r_{ij} = l|$. Por lo tanto, $\mathbf{Q_{NX}(K)}$ cuenta k-por-k bloques de Q, el rango preservado (en la diagonal principal) y las permutaciones dentro de los vecinos (en cada lado de la diagonal) [57].

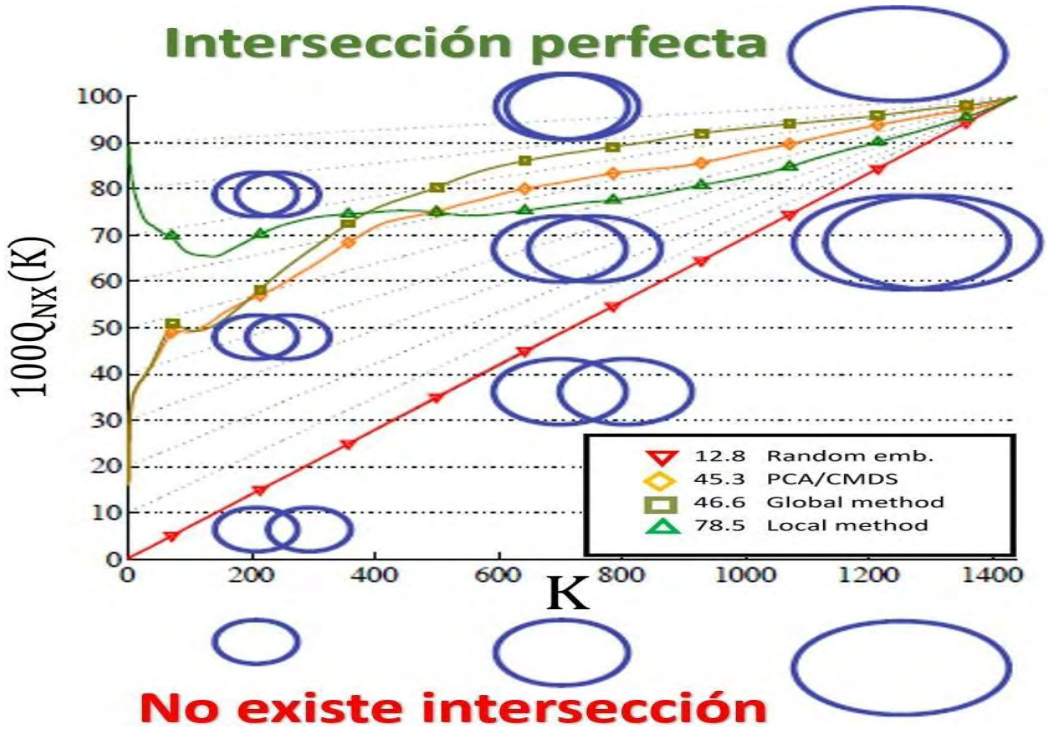


Figura 14. Ejemplo de la curva $Q_{NX}(K)$, que permite un primer acercamiento a una medida de calidad para el espacio embebido generado. **Fuente:** [6]

La **Figura 14** presenta una primera aproximación para evaluar la preservación de la topología de los datos en el espacio embebido debido a un método de RD en particular. Sin embargo, en [57] se realiza un ajuste a la curva $Q_{NX}(K)$ con el fin de que el área bajo la curva sea un buen indicador de la preservación de la topología de los datos embebidos generados, por lo tanto, la curva de calidad que se integra a la metodología de visualización está dada por:

$$\mathrm{R}_{NX}(\mathrm{K}) = \frac{(N-1)\,\mathrm{Q}_{NX}(\mathrm{K})-K}{N-1-K}\ , \qquad (11)$$

con el fin de dar una noción al usuario acerca de la calidad de la representación escogida. Un ejemplo de este tipo de curva puede ser observada en la **Figura 15.**

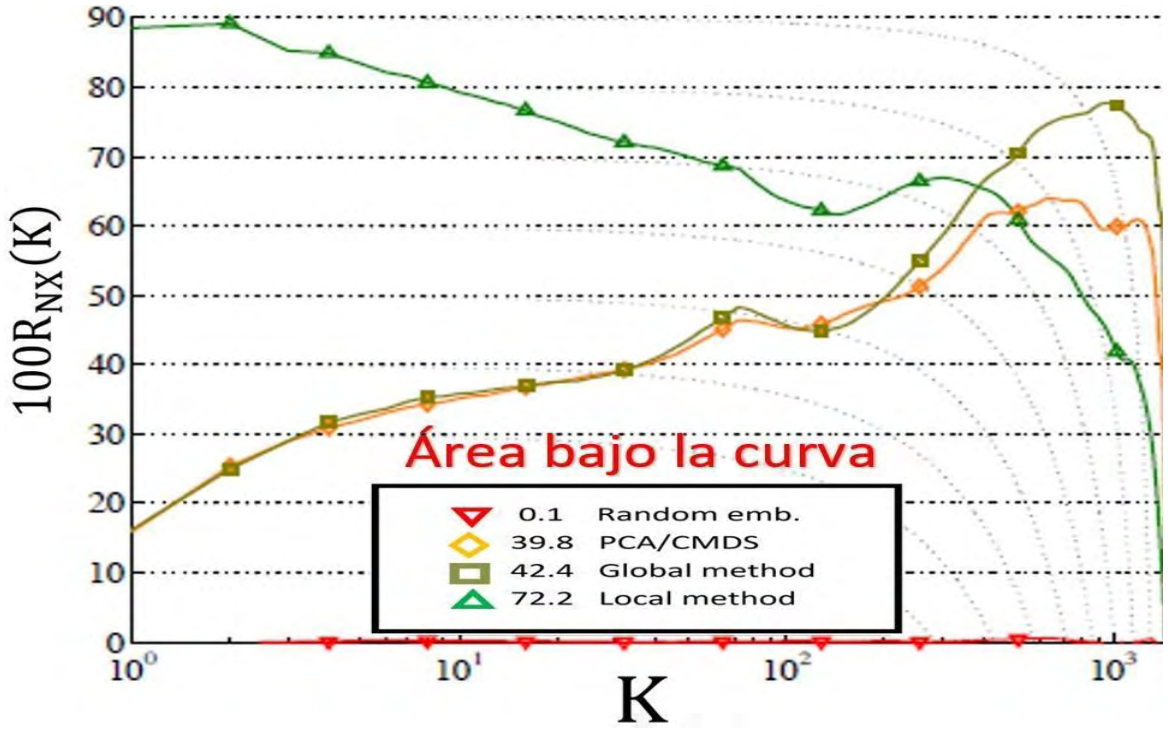
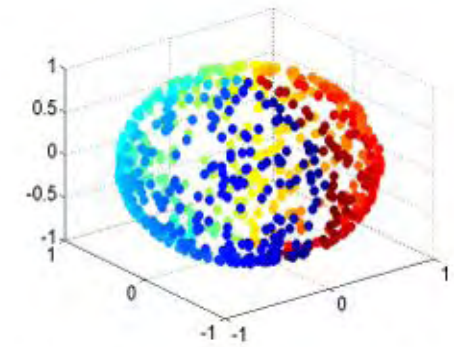


Figura 15. Medida de calidad $R_{NX}(K)$ empleada en esta investigación como un indicador del desempeño del espacio embebido. **Fuente:** [6].

3.6 BASES DE DATOS

En este trabajo los diferentes experimentos que prueban la herramienta de visualización son llevadas a cabo en tres bases de datos, una base de datos reales y dos bases de datos artificiales (**Figura 16**). La primera base de datos es una esfera conformada por 700 puntos y tres características (**Figura 16a**). La segunda base de datos es conocida como rollo suizo (**Figura 16b**) y está formada por 700 puntos con tres características. El tercer conjunto de datos es un subconjunto seleccionado al azar del banco de imágenes MNIST (**Figura 16c**), que está formado por 6000 imágenes en escala de grises de cada uno de los 10 dígitos (N = 700 puntos de datos -70 Casos para todos dígitos 10- y D = 784).



(a) Cascarón esférico 3D

(b) Rollo suizo

(c) Mnist

Figura 16. Las tres bases de datos consideradas para probar la herramienta de visualización y cada uno de sus módulos en el marco experimental propuesto. **Fuente:** https://archive.ics.uci.edu/ml/datasets.html

4. RESULTADOS

El desarrollo de la herramienta de visualización que integra modelos de interacción y técnicas de visualización se implementó en NetBeans que es uno de los entornos de desarrollo para aplicaciones Java, para que esta sea de pocas dependencias y sea ejecutado en cualquier dispositivo, también se hizo uso de algunos recursos de Processing [64] que permite desarrollar de una manera más simple las partes gráficas de la herramienta, favoreciendo la interactividad de esta. En la **figura 17** se muestra la interfaz de la herramienta que hace uso de los dos modelos de interacción unos basado en el espacio de color RGB, y el segundo basado en ángulos. Así mismo se muestran las dos técnicas de visualización diagrama de dispersión y coordenadas paralelas.

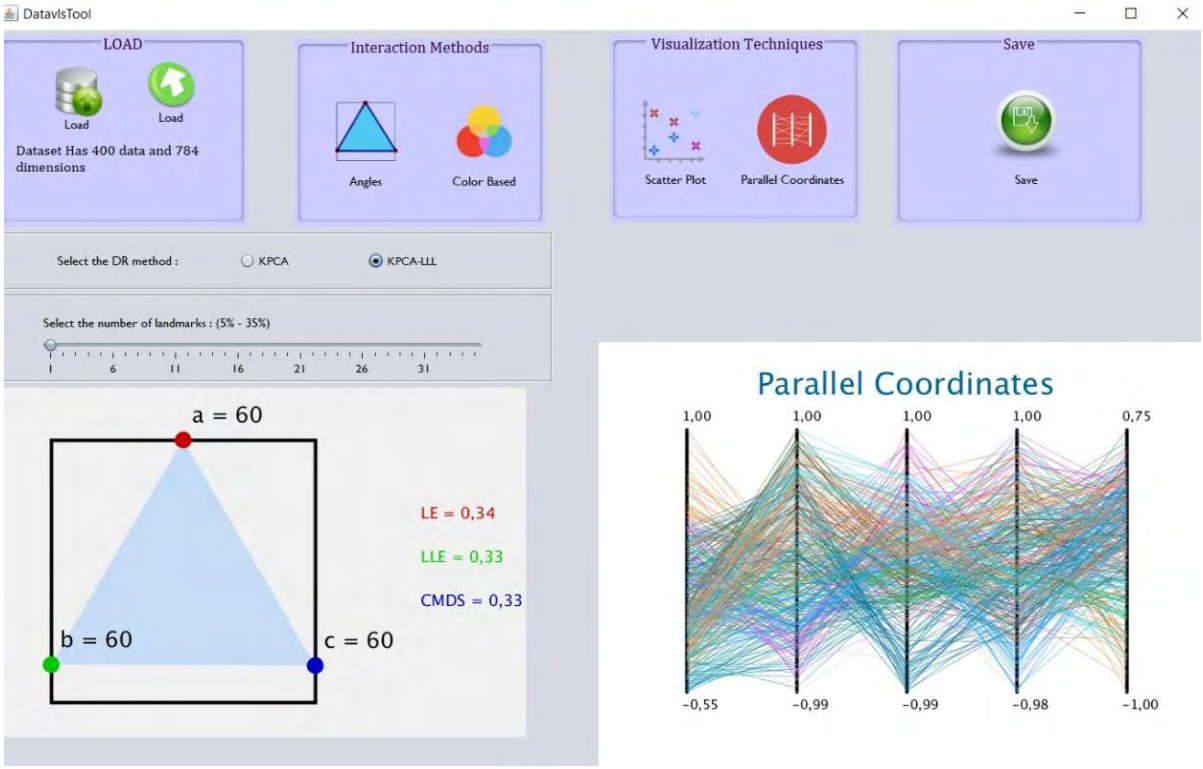


Figura 17. Herramienta de visualización implementada en el entorno de programación libre NetBeans para el desarrollo de aplicaciones Java. **Fuente:** Esta investigación.

4.1 INTERACTIVIDAD DE LA HERRAMIENTA PROPUESTA

A lo largo de la **sección 1 y 2** se argumentó por qué desarrollar herramientas y metodologías de visualización es una necesidad inminente, puesto que estas permiten que la brecha existente entre un repositorio de datos y su conocimiento oculto dentro de ellas se vea disminuido. El ser humano y la máquina trabajan conjuntamente para que dicho conocimiento se convierta en una base que permita formular hipótesis y posteriormente determinar decisiones o procesos a seguir a partir del entendimiento de los datos. La herramienta creada en Java se ambienta de manera afable, de manera que el usuario, además de hacer uso de su propio repositorio de datos, pueda aplicar un método de reducción en particular o una mezcla de estos, sin tener conocimiento previo de la fundamentación de los métodos de RD.

La herramienta cuenta con 5 módulos (**Figura 18**) los cuales se utilizan secuencialmente; en primera instancia está el módulo de cargar, (**Figura 18a**) en el que se tienen dos opciones, hacer uso de bases de datos precargadas o cargar su propio repositorio en archivos de tipo CSV con etiquetas o sin ellas. El segundo módulo (**Figura 18b**) permite escoger el modelo de interacción a usar por parte del usuario, el primero es el modelo cromático (**Sección 3.3.1**) y el segundo propuesto por esta investigación es el modelo basado en ángulos (**Sección 3.3.2**), cada uno de los modelos muestra a su lado los factores de ponderación para cada método de RD. El tercer módulo (**Figura 18c**) permite seleccionar una de las dos técnicas de visualización implementadas diagrama de dispersión o coordenadas paralelas (**Sección 2.4**) en donde se representarán el espacio embebido de hasta 10 dimensiones obtenido en el proceso de reducción.

El cuarto módulo (**Figura 18d**) permite guardar como imagen PNG la representación de los datos de baja dimensión escogida por el usuario, así como los datos embebidos en un archivo plano CSV. Finalmente, un plus de la herramienta es la optimización del cálculo del espacio embebido (**Figura 18e**) después de la mezcla de las representaciones kernel, para esto el usuario puede elegir si hacer obtener los datos embebidos por el método normal Kernel PCA (**Sección 3.2**) o por la optimización Kernel PCA con LLL (**Sección 3.4**), en este último se puede elegir el porcentaje de la base de datos originales con la que se ejecutara las matrices localmente lineales (LLL), entre más porcentaje mayor será el tiempo de ejecución pero en algunas ocasiones mejorará la representación. La conformación de estos 5 módulos hace la herramienta completa como se observa en la **figura 17** donde los espacios restantes se utilizan para ubicar los modelos de interacción y técnicas de visualización que más se ajusten a el deseo del usuario quien es el que tiene la libertad de escoger y combinar entre estas para tener una representación deseada.

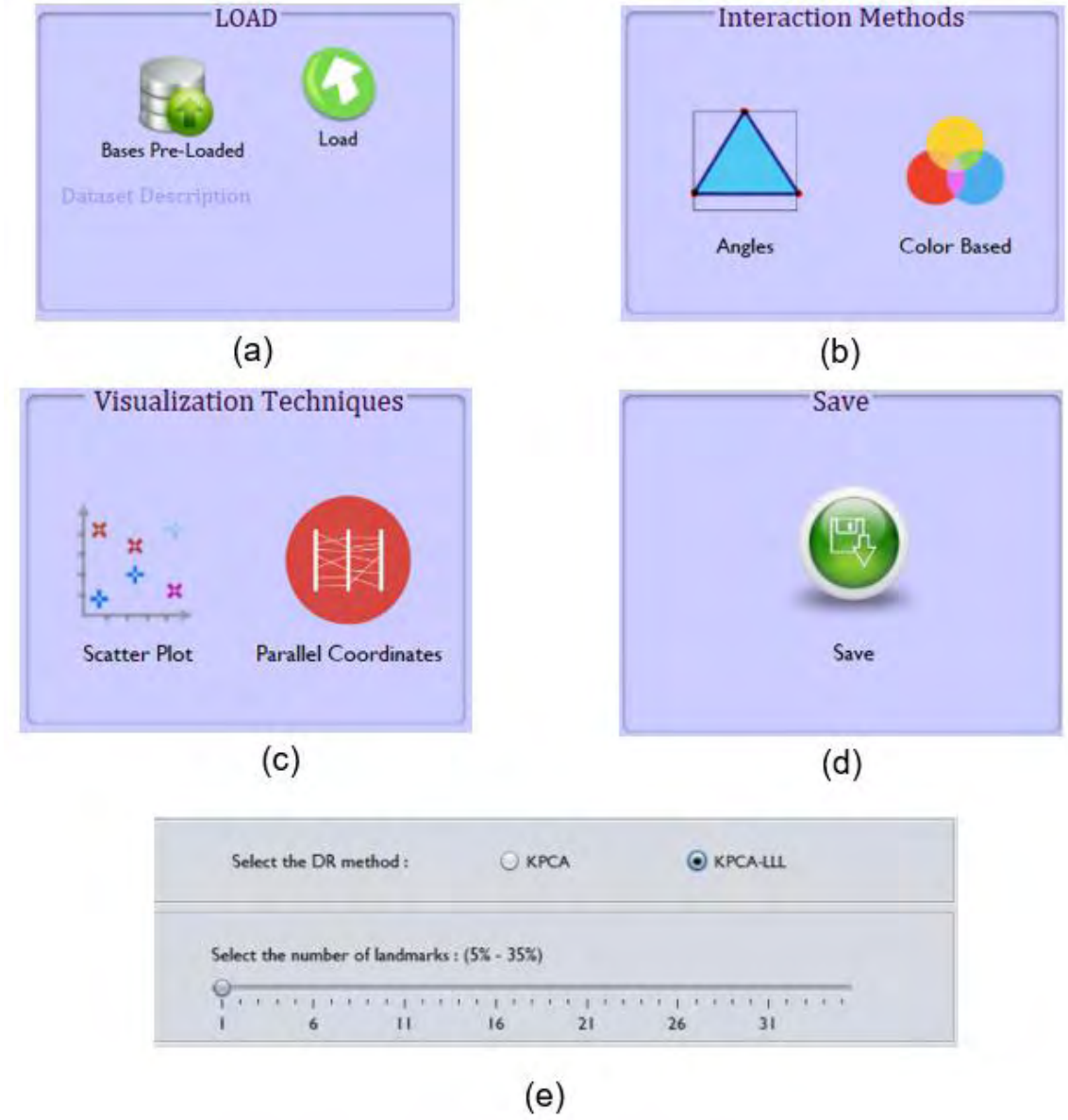


Figura 18. Se indican los diferentes módulos que constituyen la herramienta de visualización permiten el desarrollo del proceso de representación de datos. Entre las principales funciones tenemos cinco módulos para elegir bases precargadas o cargar su propio repositorio (a), elegir qué modelo de interacción usara (b), así mismo que técnica de visualización (c), un botón de guardado (d) y el modelo de optimización del tiempo de ejecución (e). **Fuente**: Esta investigación.

4.2 MARCO EXPERIMENTAL EXPERIMENTO 1: PARA PROBAR LA CONTROLABILIDAD E INTERACTIVIDAD DE LA HERRAMIENTA

Para probar la controlabilidad e interactividad de la herramienta se consideran los modelos de interacción descritos en la **sección 3.3** aleatoriamente se escogieron diferentes factores ponderación para cada uno de ellos, además observar de acuerdo con el criterio de $\mathrm{R}_{NX}(\mathrm{K})$ (**Sección 3.5**) si la mezcla escogida de métodos de RD puede conservar la topología de los datos aún mejor que las aproximaciones kernel de los métodos convencionales. La **figura 19** ilustra un diagrama del proceso de la representación visual donde el usuario puede interactuar con la mezcla de los métodos y controlar la representación de los datos resultantes, por medio de los modelos de interacción. Por lo tanto, lo que se busca evaluar es, cómo mediante la interacción con cada modelo el usuario puede controlar el tipo de representación de los datos en baja dimensión, ya sea seleccionando un método en particular o una mezcla ponderada de los mismos. De esta manera el usuario al interactuar directamente con los modelos puede elegir los factores de ponderación α de la mezcla lineal de las aproximaciones kernel, obteniendo un gran número de representación hasta obtener la que se ajuste a sus necesidades.

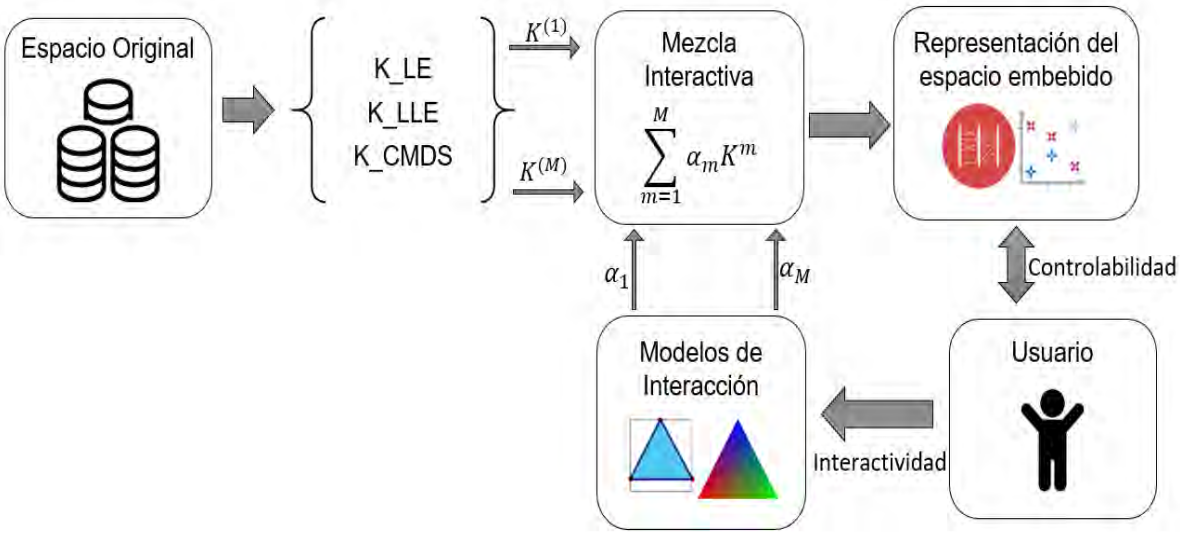


Figura 19. Diagrama de la metodología interactiva de representación de datos, donde el usuario mediante realimentación visual puede controlar e interactuar con el proceso. **Fuente:** Esta investigación.

4.3 MARCO EXPERIMENTAL EXPERIMENTO 2: PARA EVALUAR EL RENDIMIENTO DE LAS SUBMATRICES LOCALMENTE LINEALES COMO MÉTODO PARA REDUCIR EL COSTO COMPUTACIONAL

En la **sección 3.4** como parte de la metodología de la herramienta se propone implementar el método de submatrices localmente lineales (LLL) introducida en [14]. En el cual se plantea reducir el tiempo de procesamiento de KPCA (**Sección 3.2**), método utilizado en esta investigación para reducir la dimensión a partir de matrices kernel. Este método cuenta con un parámetro que permite variar el número de submatrices el cual entre mayor sea más aproximado será el subespacio que se escoja al espacio original tal y como se explicó en la **sección 3.4,** el cual se podrá controlar por una barra deslizable mostrada en la **figura 18e.** Como se observa en el diagrama de la **figura 20** el usuario entonces será capaz de controlar e interactuar con la representación de datos no solo por medio de los modelos de interacción si no por el número de submatrices del método LLL. Este experimento consistirá entonces en probar aleatoriamente número de submatrices, cuantificar el tiempo de ejecución y comparar los espacios embebidos obtenidos entre sí y con KPCA sin LLL.

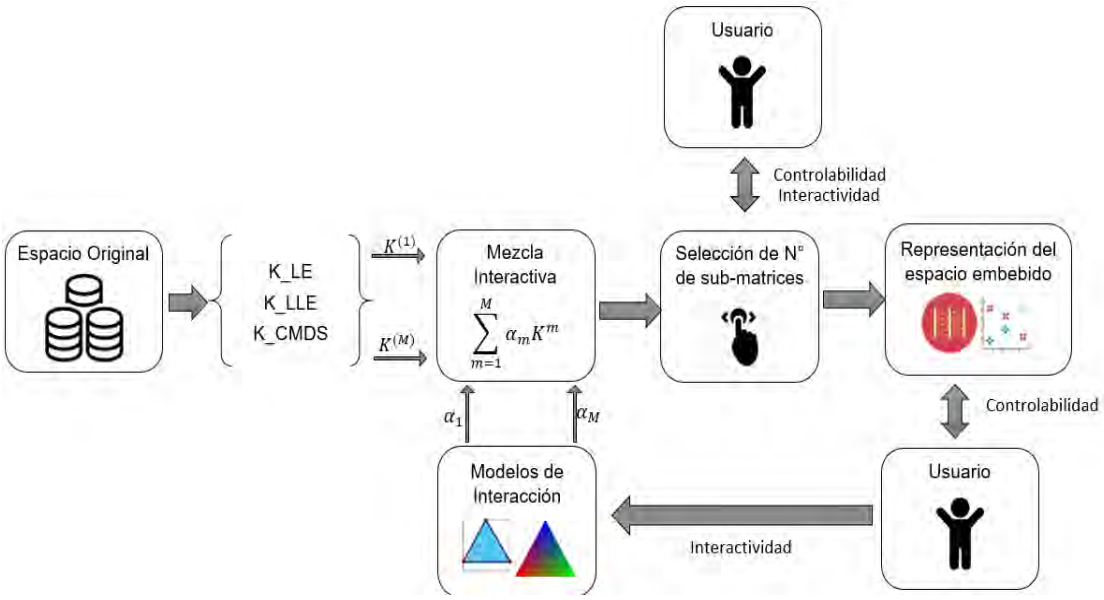


Figura 20. Diagrama de la metodología interactiva de representación de datos, donde el usuario mediante realimentación visual puede controlar e interactuar no solo con el modelo de interacción, también con el método de reducción de coste computacional LLL. **Fuente:** Esta investigación.

4.4 RESULTADOS EXPERIMENTO 1

4.4.1 Resultados obtenidos para el modelo cromático

Este modelo fue diseñado en Processing y permite al usuario tener una mejor noción de la mezcla de métodos de RD a través de un color seleccionado en la superficie de un triángulo Cromático (**Figura 21**), además este cuenta con unas barras de porcentaje que especifican cuánto le corresponde de rojo, verde y azul.

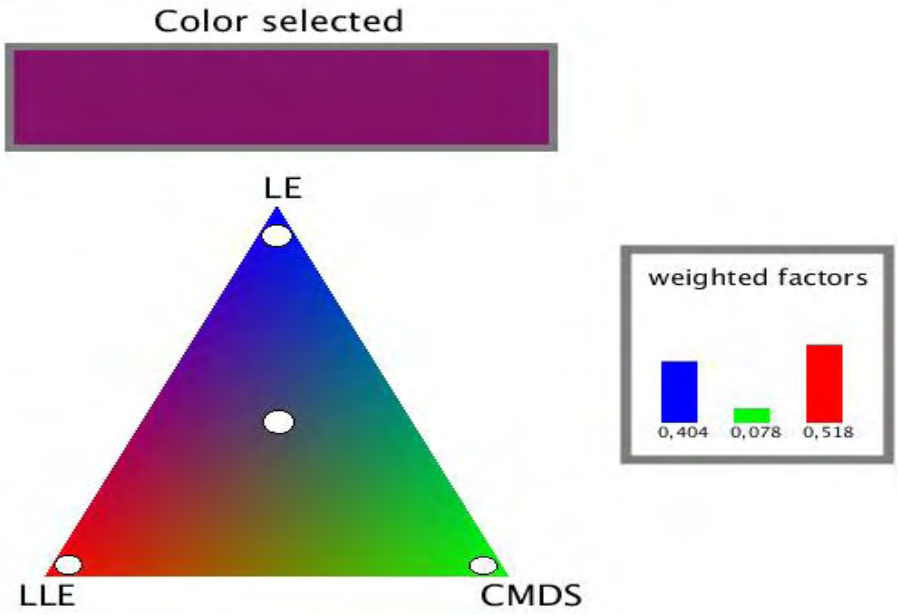
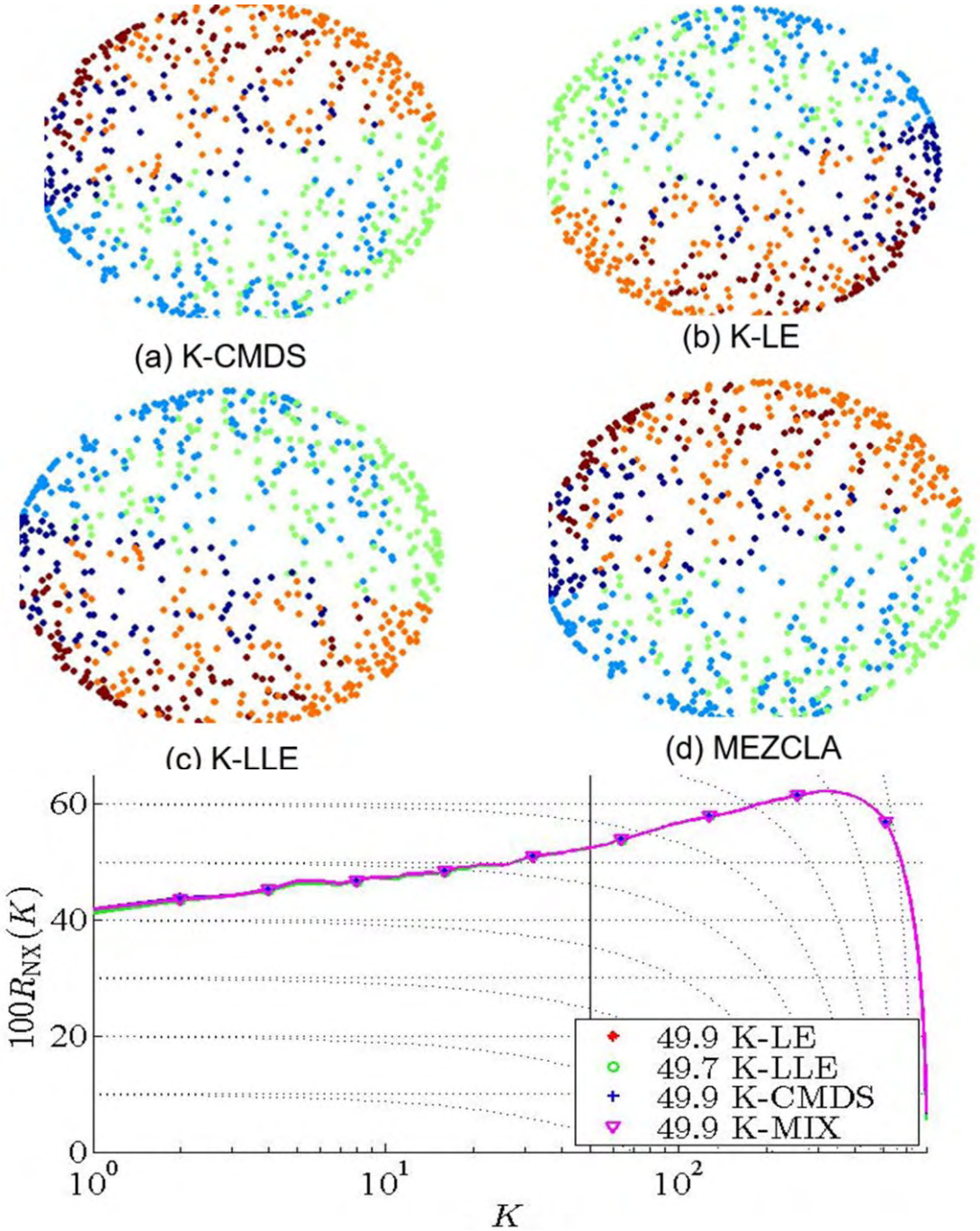


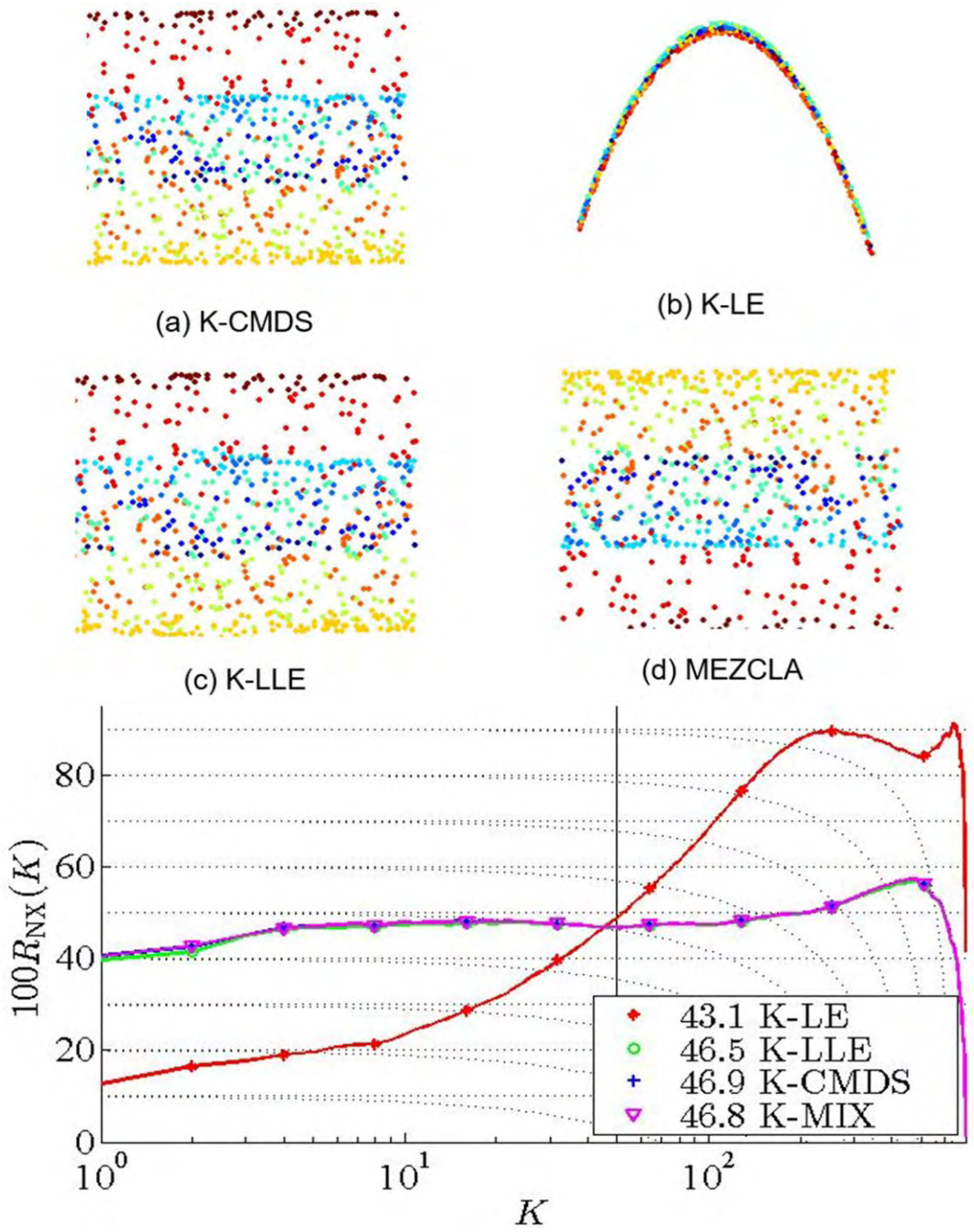
Figura 21. Modelo cromático implementado en Processing, además se muestran los puntos aleatorios en la superficie del triángulo cromático escogidos para probar la interactividad y controlabilidad del modelo. **Fuente:** Esta investigación.

En la **figura 21** se observa que se escogieron aleatoriamente 4 puntos en la superficie del triángulo cromático, de los cuales 3 están cercanos a los vértices que representan a cada uno de los métodos de RD utilizados en esta investigación, y el cuarto corresponde al centro del triángulo que representa la mezcla por igual porcentaje de cada método. Los resultados de este experimento aplicados a las bases de datos de la **sección 3.6** se muestran en las **figuras 22 a 24.** Las (a), (b), (c) representan los espacios embebidos de los vértices del triángulo donde se selecciona un punto exacto donde es respectivamente $\boldsymbol{K}_{CMDS}, \boldsymbol{K}_{LE}\ y\ \boldsymbol{K}_{LLE}$, el resultado (d) se obtiene mediante la mezcla ponderada de las matrices kernel donde $\alpha_{CMDS} = \alpha_{LE} = \alpha_{LLE} = 0{,}33$.



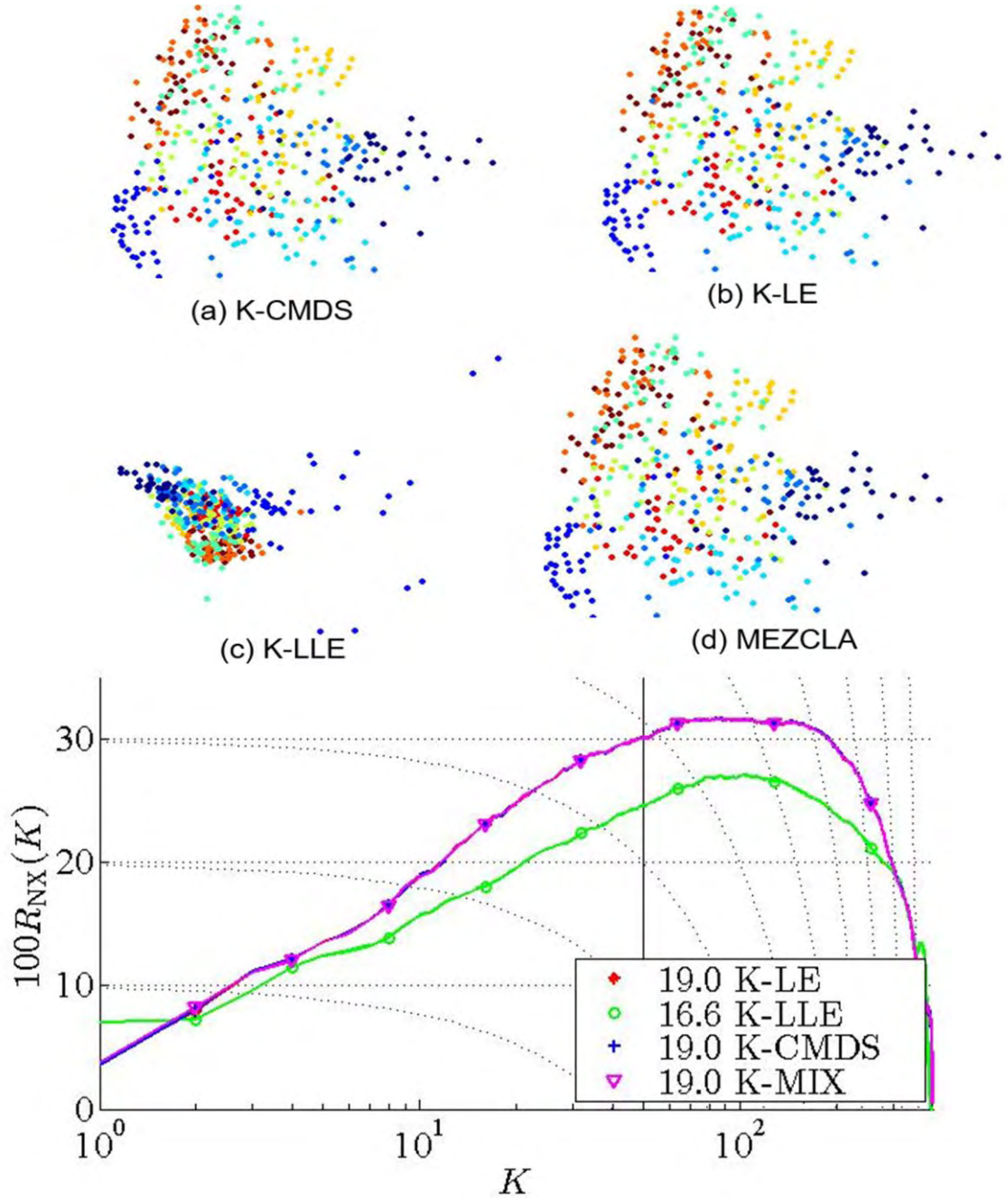
(e) $\mathrm{R}_{NX}(\mathrm{K})$

Figura 22. Resultados obtenidos a partir de la herramienta de visualización usando el modelo de interacción cromático para la base de datos cascaron esférico en 3D. Las figuras (a)-(d) indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo. La figura (e) indica la medida de calidad $R_{NX}(K)$. **Fuente**: Esta investigación.



(e) $\mathrm{R}_{NX}(\mathrm{K})$

Figura 23. Resultados obtenidos a partir de la herramienta de visualización usando el modelo de interacción cromático para la base de datos conocida como rollo suizo. Las figuras (a)-(d) indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo. La figura (e) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$. **Fuente**: Esta investigación.



(e) $\mathrm{R}_{NX}(\mathrm{K})$

Figura 24. Resultados obtenidos a partir de la herramienta de visualización usando el modelo de interacción cromático para la base de datos conocida como Mnist. Las figuras (a)-(d) indican los espacios embebidos resultantes a partir de la selección de cuatro puntos aleatorios dentro del modelo. La figura (e) indica la medida de calidad $R_{NX}(K)$. **Fuente**: Esta investigación.

4.4.2 Resultados obtenidos para el modelo basado en ángulos

Este modelo también fue diseñado en Processing, no obstante, el usuario interactúa de manera diferente con este modelo, pero se basa en el mismo funcionamiento del modelo cromático. Para probar la controlabilidad e interactividad del modelo propuesto en esta investigación, se tomaron 3 diferentes mezclas (**Figura 25**), la primera corresponde a $\alpha_{CMDS} = 0{,}25\ \alpha_{LLE} = 0{,}3\ \alpha_{LE} = 0{,}45$ (Mix 1) (**Figura 25a**), la segunda corresponde a $\alpha_{CMDS} = 0{,}4\ \alpha_{LLE} = 0{,}3\ \alpha_{LE} = 0{,}3$ ((Mix2) **Figura 25b**), y última corresponde a $\alpha_{CMDS} = 0{,}2\ \alpha_{LLE} = 0{,}5\ \alpha_{LE} = 0{,}3$ (Mix 3) (**Figura 25c**).

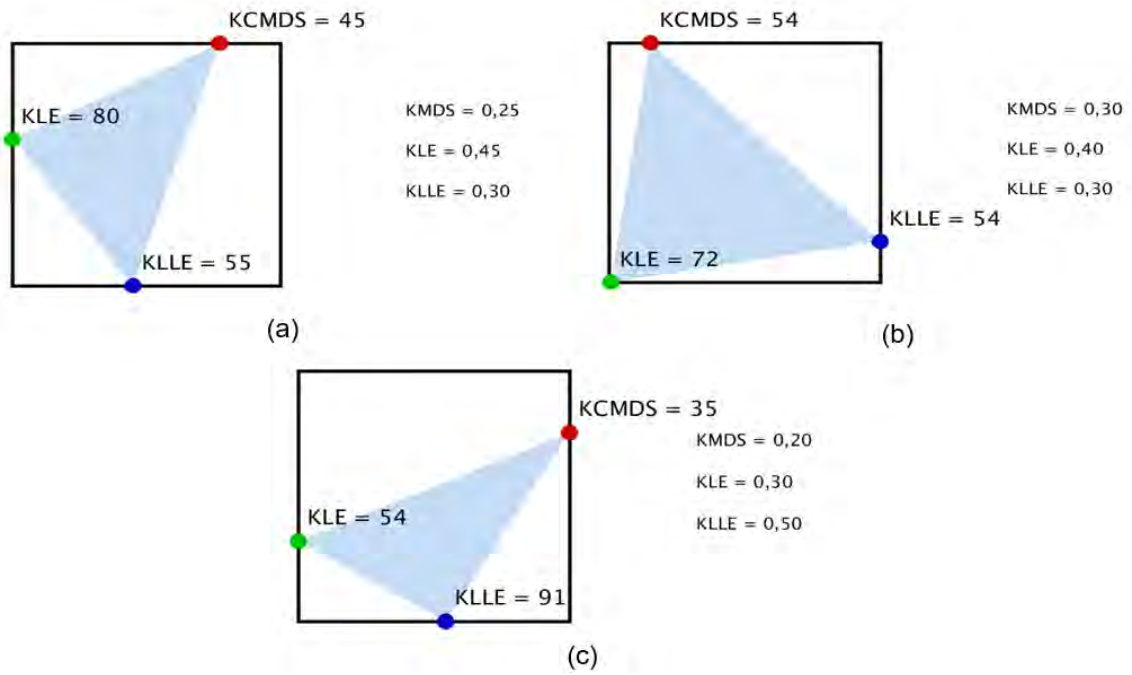
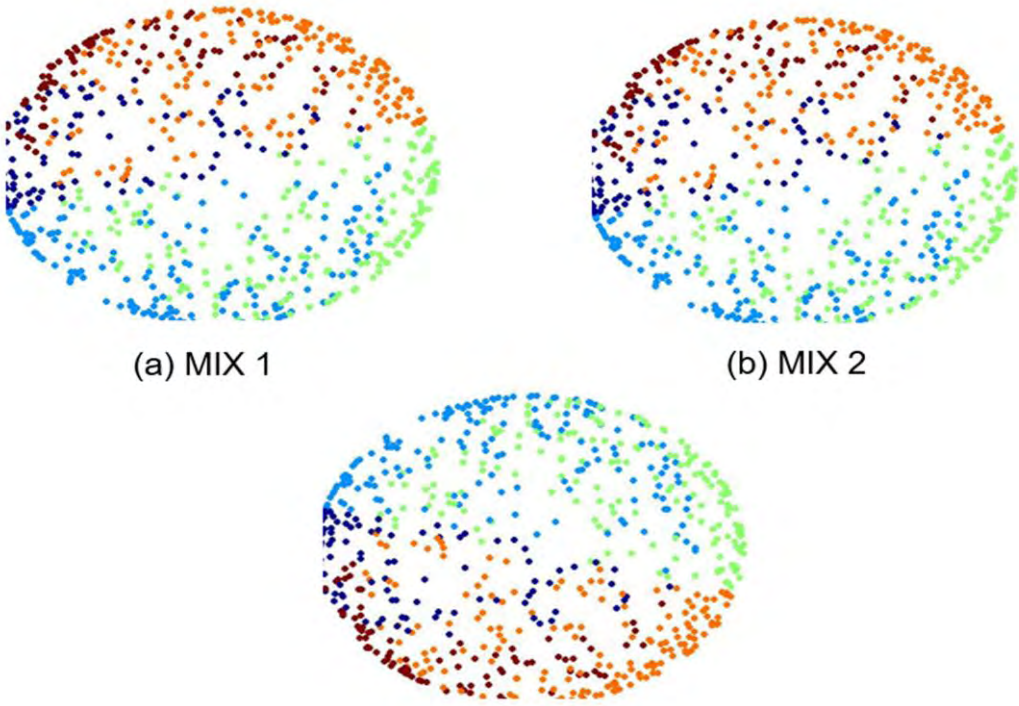


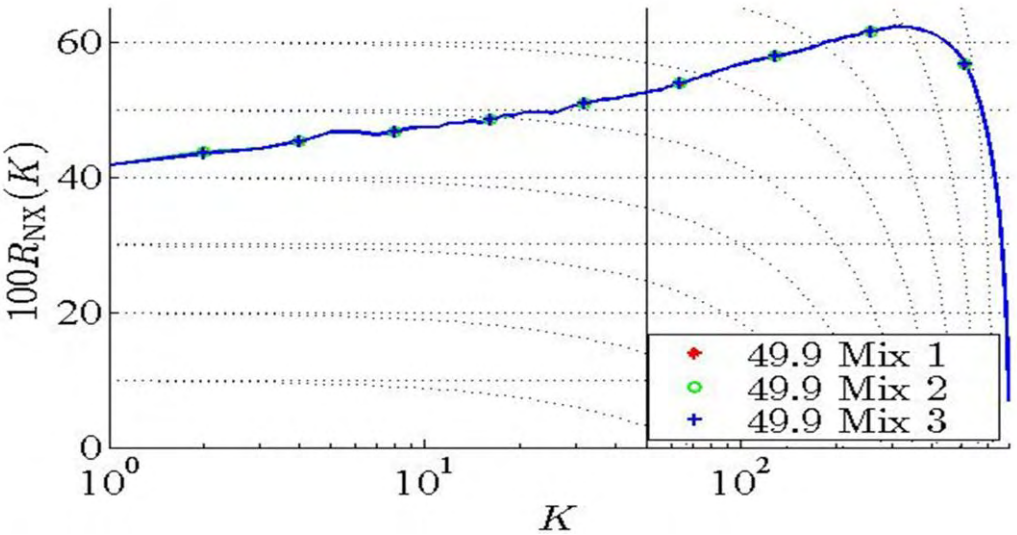
Figura 25. Modelo basado en ángulos implementado en Processing, además se muestran las representaciones aleatorias (a) a (c) para realizar las pruebas de interactividad y controlabilidad del modelo en la herramienta implementada. **Fuente:** Esta investigación.

Los resultados para este experimento se encuentran en las **figuras 26 a 28**, en las que se evidencian los espacios embebidos (**Figuras 26 a 28 (a), (b), (c)**) y la cuantificación de la conservación de la topología de los datos originales en dichos espacios obtenidos (**Figuras 26 a 28 (d)**). A pesar de que los espacios difieran de representación bien sea de forma o traslación en los colores que se determinan por el vector de etiquetas, el área bajo la curva $R_{NX}(K)$ no varía lo que indica que a pesar de que se realicen diferentes mezclas se sigue conservando la topología de los datos.



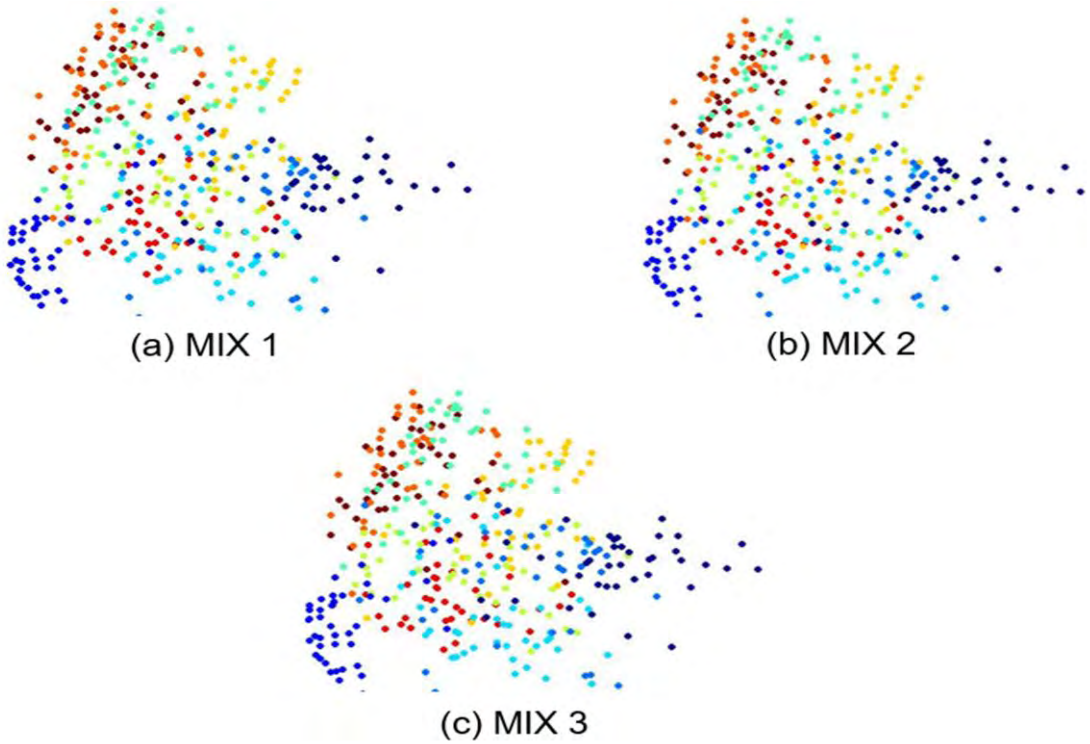
(a) MIX 1 (b) MIX 2

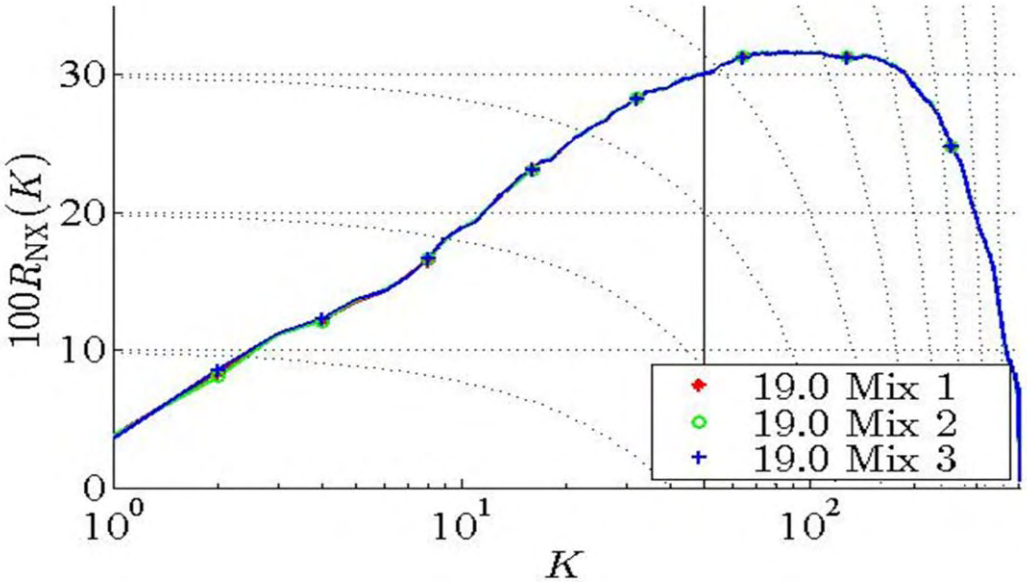
(c) MIX 3



(d) $\mathrm{R}_{NX}(\mathrm{K})$

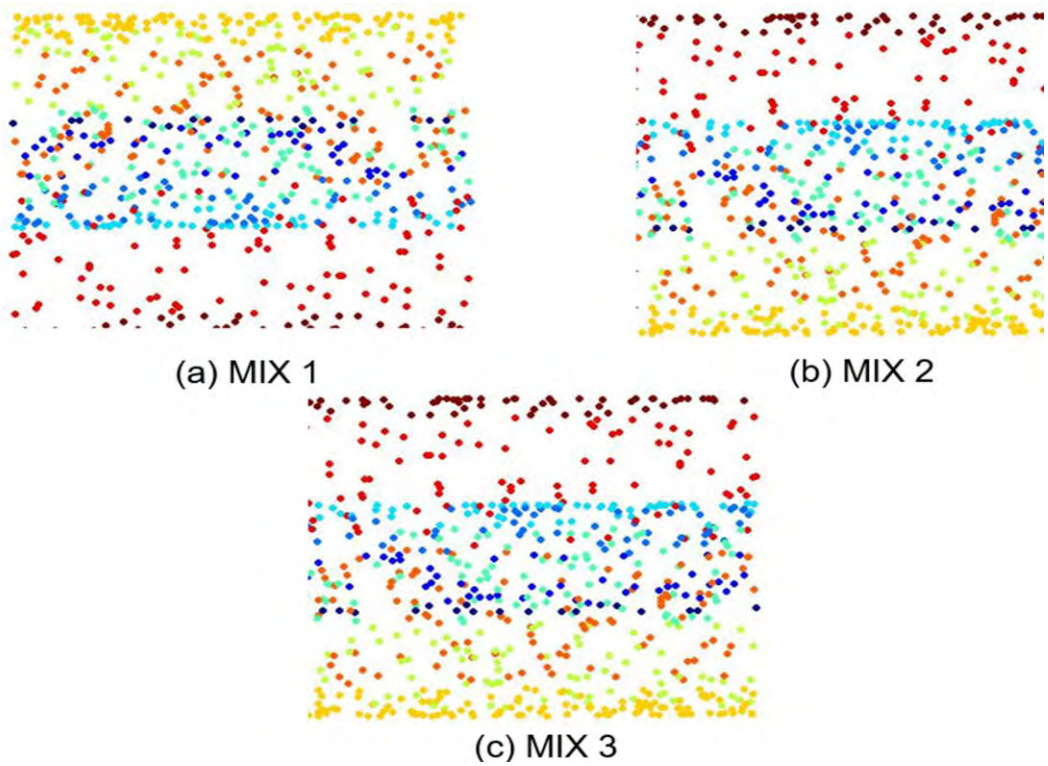
Figura 26. Resultados obtenidos a partir de la herramienta de visualización usando el modelo de interacción basado en ángulos para la base de datos conocida como cascaron esférico 3D. Las figuras (a)-(c) indican los espacios embebidos resultantes a partir de la selección de las tres representaciones (**Figura 25**). La figura (d) indica la medida de calidad $R_{NX}(K)$. **Fuente**: Esta investigación.





(d) $\mathrm{R}_{NX}(\mathrm{K})$

Figura 27. Resultados obtenidos a partir de la herramienta de visualización usando el modelo de interacción basado en ángulos para la base de datos conocida como Mnist. Las figuras (a)-(c) indican los espacios embebidos resultantes a partir de la selección de las tres representaciones (**Figura 25**). La figura (d) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$. **Fuente**: Esta investigación.



(d) $\mathrm{R}_{NX}(\mathrm{K})$

Figura 28. Resultados obtenidos a partir de la herramienta de visualización usando el modelo de interacción basado en ángulos para la base de datos conocida como Rollo suizo. Las figuras (a)-(c) indican los espacios embebidos resultantes a partir de la selección de las tres representaciones (**Figura 25**). La figura (d) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$. **Fuente**: Esta investigación.

4.4.3 Discusión

En los resultados se pueden apreciar los diagramas de dispersión en donde se representan los datos embebidos, además se dan a conocer las curvas de calidad que permiten de alguna manera establecer un criterio sobre el desempeño de la representación en un espacio de baja dimensión, teniendo en cuenta la preservación de la topología de los datos a través del área bajo la curva. De esta manera, si el valor del área bajo la curva es mayor, el rendimiento de los datos embebidos será mejor debido a la conservación de los k-vecinos más cercanos. Además, los modelos de interacción permiten dotar al usuario de propiedades como la interactividad y la controlabilidad, donde el usuario puede apreciar representaciones dinámicas y que la percepción del funcionamiento de la reducción de dimensión se vuelva más dinámica y de mayor entendimiento para cualquier usuario, puesto que como se ha dicho anteriormente somos seres humanos visuales capaces de analizar información visualmente de manera más sencilla. Por lo tanto, la controlabilidad de la herramienta se reduce a seleccionar los factores de ponderación bien sea en el modelo cromático por medio de colores o en el modelo basado en ángulos donde el triángulo cambia la medida de estos, así como la técnica de visualización donde se representarán los datos de baja dimensión bien sea el diagrama de dispersión o coordenadas paralelas que permiten apreciar más de 10 dimensiones. Si se comparan las dos secciones anteriores el realizar una mezcla ponderada de los métodos RD (**Sección 4.4.2**) conserva mejor la topología de los datos que si se realiza una reducción de dimensión solo haciendo uso de uno de los métodos (**Sección 4.4.1**), esto resulta evidente ya que se estaría aprovechando las ventajas de cada uno de los métodos espectrales.

4.5 RESULTADOS EXPERIMENTO 2

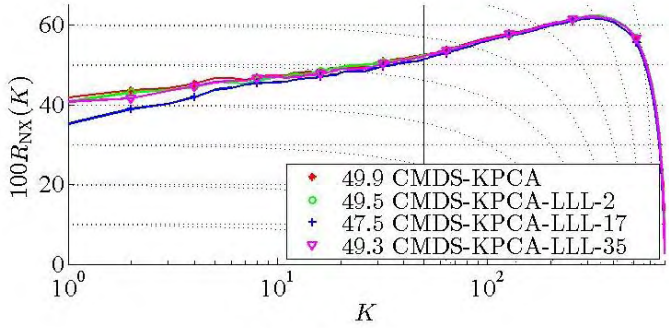
Como se implementaron dos modelos de interacción las pruebas de este experimento se realizaron de manera similar que en la **sección 4.4** haciendo las mismas variaciones para cada modelo de interacción, sin embargo, para cada prueba se variara el número de submatrices lineales en porcentaje de 2%, 17%, y 35%, además se ejecutó el proceso de reducción con KPCA sin uso del método LLL, esto con el fin de cuantificar y comparar el tiempo de ejecución en KPCA y KPCA-LLL, para obtener el tiempo de ejecución se ejecutó cada escenario 20 veces para calcular la desviación estándar de los valores de tiempo obtenidos. También se mostrarán las curvas de calidad $\mathrm{R}_{NX}(\mathrm{K})$ que nos permitirán determinar si la disminución del tiempo de ejecución debido a la aplicación del método de submatrices localmente lineales se ve reflejado en la mejora o desmejora de la representación final de los datos embebidos con respecto a la preservación de la topología de los datos originales.

4.5.1 Resultados obtenidos para el modelo cromático

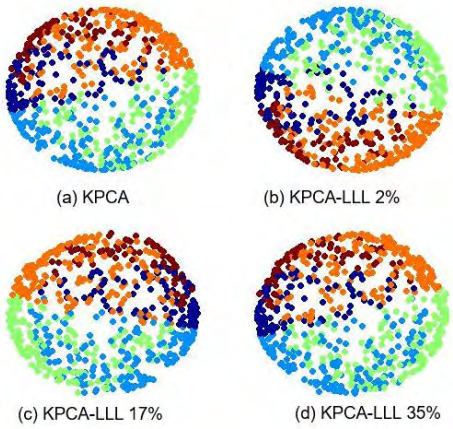
Para esta prueba se escogieron los mismos factores de ponderación aleatorios presentados en la **sección 4.4.1,** en las **figuras 29 a 31** se muestran los resultados para el experimento dos, en cada una de estas figuras se indican los espacios obtenidos de las variaciones de los factores de ponderación que son 4 diferentes escenarios (1) $\alpha_{CMDS} = 1\ y\ \alpha_{LLE} = \alpha_{LE} = 0$, (2) $\alpha_{LE} = 1\ y\ \alpha_{CMDS} = \alpha_{LLE} = 0$, (3) $\alpha_{LLE} = 1\ y\ \alpha_{CMDS} = \alpha_{LE} = 0$, (4) $\alpha_{LE} = \alpha_{CMDS} = \alpha_{LLE} = 0{,}33$. Además, para la obtención de los espacios embebidos se aplicaron 4 diferentes maneras de obtener los espacios embebidos (a) aplicando KPCA, (b) KPCA-LLL con 2% de submatrices, (c) KPCA-LLL con 17% y (d) KPCA-LLL con 35% de submatrices. Con esto entonces visualizan por cada base de datos 16 diferentes representaciones de espacios embebidos. Adicionalmente cada uno de los escenarios de (1) a (4) contiene la curva de calidad en las que se comparan los espacios obtenidos a partir de las pruebas (a) a (d).
Como resultados para este experimento también se obtienen los tiempos de ejecución para cada prueba realizada descrita anteriormente los cuales se muestran en la **tabla 1** donde además se relaciona la desviación estándar determinada tras probar la herramienta completa 20 veces por cada prueba aplicada a todas las bases de datos. Se puede apreciar que el costo computacional disminuye en algunas ocasiones hasta el 80% con relación a KPCA, de manera que el aplicar KPCA-LLL permite una representación más dinámica de los datos garantizando una mayor interactividad con la herramienta.

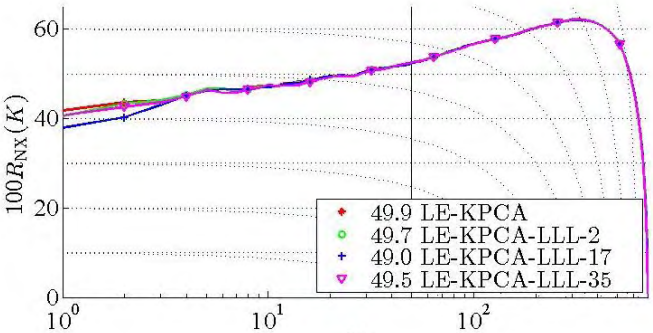
Tabla 1. Tiempo de ejecución para los diferentes escenarios del experimento 2

Factores de ponderación	Base de datos	Tiempo con KPCA	Tiempo con KPCA-LLL 2%	Tiempo con KPCA-LLL 17%	Tiempo con KPCA-LLL 35%
$\alpha_{LLE} = 1$ $\alpha_{CMDS} = 0$ $\alpha_{LE} = 0$	Esfera 3D	3804,7 ± 555,7	78,4 ± 8,8	573,1 ± 36,1	1796,8 ± 54,7
	Rollo suizo	2758,8 ± 149,8	67,8 ± 4,3	548,2 ± 17	1771,2 ± 77,1
	Mnist	544,8 ± 9,4	83,1 ± 13,2	247,6 ± 8,7	582,6 ± 18,7
$\alpha_{LLE} = 0$ $\alpha_{CMDS} = 1$ $\alpha_{LE} = 0$	Esfera 3D	4191,3 ± 476,8	68,9 ± 4,1	569,7 ± 23,1	1733,7 ± 41,7
	Rollo suizo	3021,6 ± 204,6	72,2 ± 3,4	565,2 ± 18,2	1725,4 ± 51,5
	Mnist	492,8 ± 18,7	59,3 ± 8,1	226,1 ± 5,2	579,5 ± 10,7
$\alpha_{LLE} = 0$ $\alpha_{CMDS} = 0$ $\alpha_{LE} = 1$	Esfera 3D	3244,8 ± 393,7	76,1 ± 4,9	568,2 ± 12,5	1802,3 ± 54,9
	Rollo suizo	3355,2 ± 367,6	79,3 ± 6,9	567,2 ± 14,9	1798,2 ± 35,6
	Mnist	477,8 ± 10,8	61,7 ± 2,9	228,6 ± 3,3	564,2 ± 21,8
$\alpha_{LLE} = 0{,}33$ $\alpha_{CMDS} = 0{,}33$ $\alpha_{LE} = 0{,}33$	Esfera 3D	3178,8 ± 584,8	96,4 ± 16,8	577,9 ± 45,1	1793,8 ± 136,1
	Rollo suizo	2701,1 ± 477,1	75 ± 7,2	556,3 ± 25,3	1793 ± 84,1
	Mnist	488,5 ± 22,8	63 ± 15,1	228,7 ± 7,1	582,1 ± 9,8

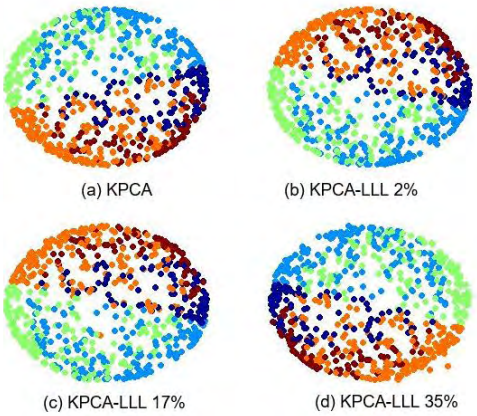


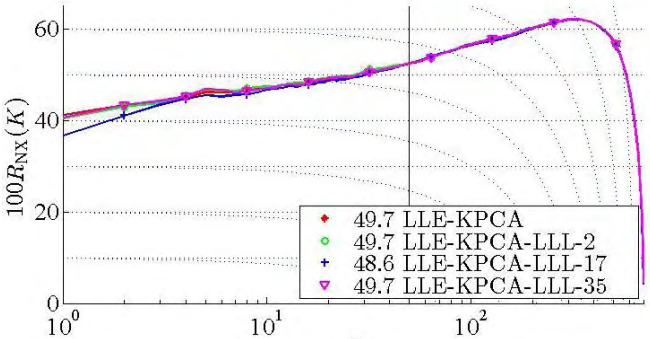
(e) $R_{NX}(K)$



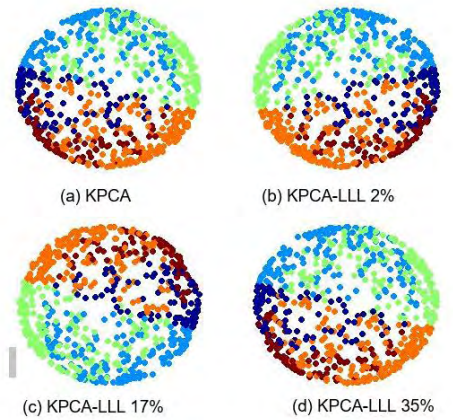


(e) $R_{NX}(K)$





(e) $R_{NX}(K)$



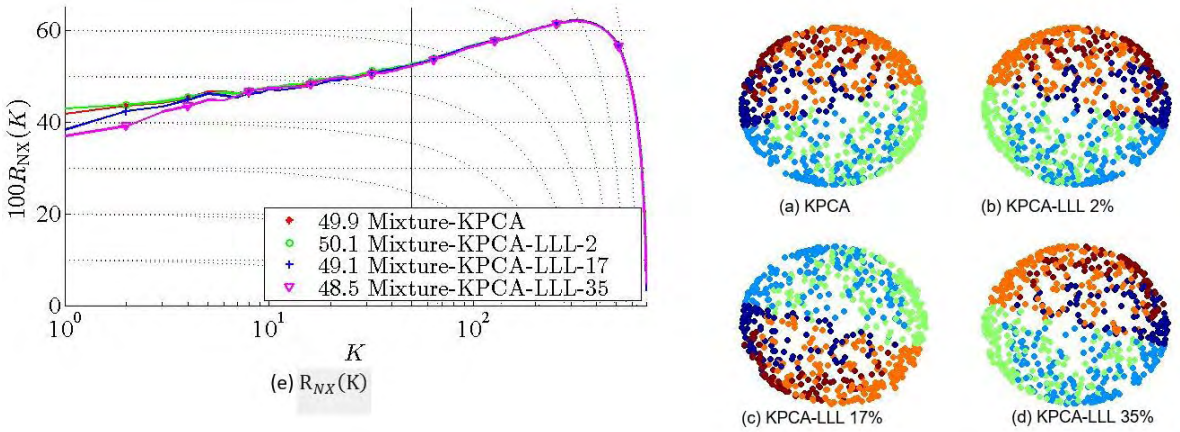
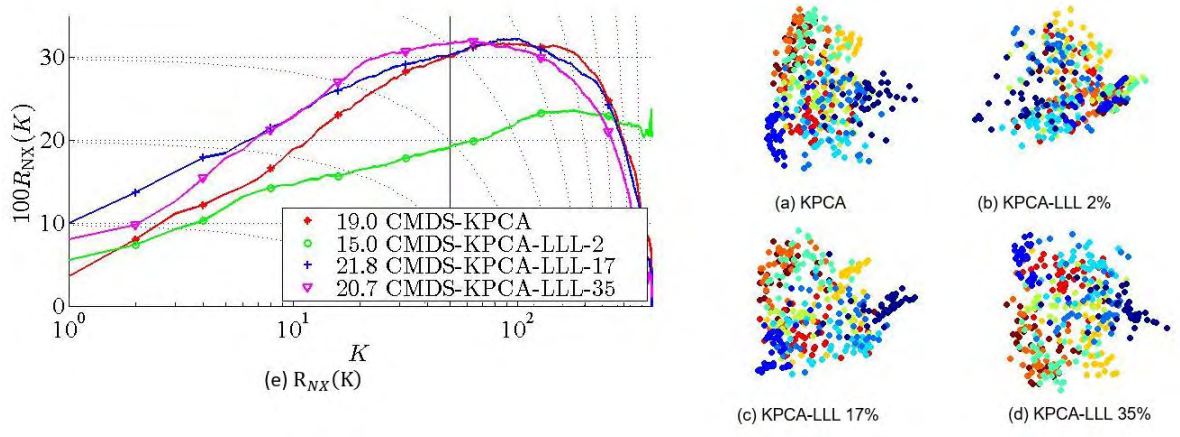


Figura 29. Resultados obtenidos para la base de datos cascarón esférico 3D con diferentes factores de ponderación (1) $\alpha_{CMDS} = 1\ y\ \alpha_{LLE} = \alpha_{LE} = 0$, (2) $\alpha_{LE} = 1\ y\ \alpha_{CMDS} = \alpha_{LLE} = 0$, (3) $\alpha_{LLE} = 1\ y\ \alpha_{CMDS} = \alpha_{LE} = 0$, (4) $\alpha_{LE} = \alpha_{CMDS} = \alpha_{LLE} = 0{,}33$, escogidos con el modelo de interacción cromático. En las figuras principales (1) a (4) cada ilustración (a) es el espacio embebido obtenido de la aplicación de KPCA y (b) a (d) son los espacios obtenidos por medio de KPCA-LLL con el número de submatrices al 2%, 17% y 35%. La figura (e) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$ de los datos embebidos de la (a) a (d). **Fuente**: Esta investigación.

Para la base de datos del cascarón esférico (**Figura 29**) se obtienen resultados similares y con cambios no muy significativos en cuanto la representación bidimensional obtenida y las curvas de calidad. Pero si evidencia disminución de hasta la 4 parte del tiempo de ejecución en el escenario de KPCA-LLL en el 2% tiempo que va aumentando progresivamente cuando se aumenta al 17% y 35% de la totalidad de los datos.



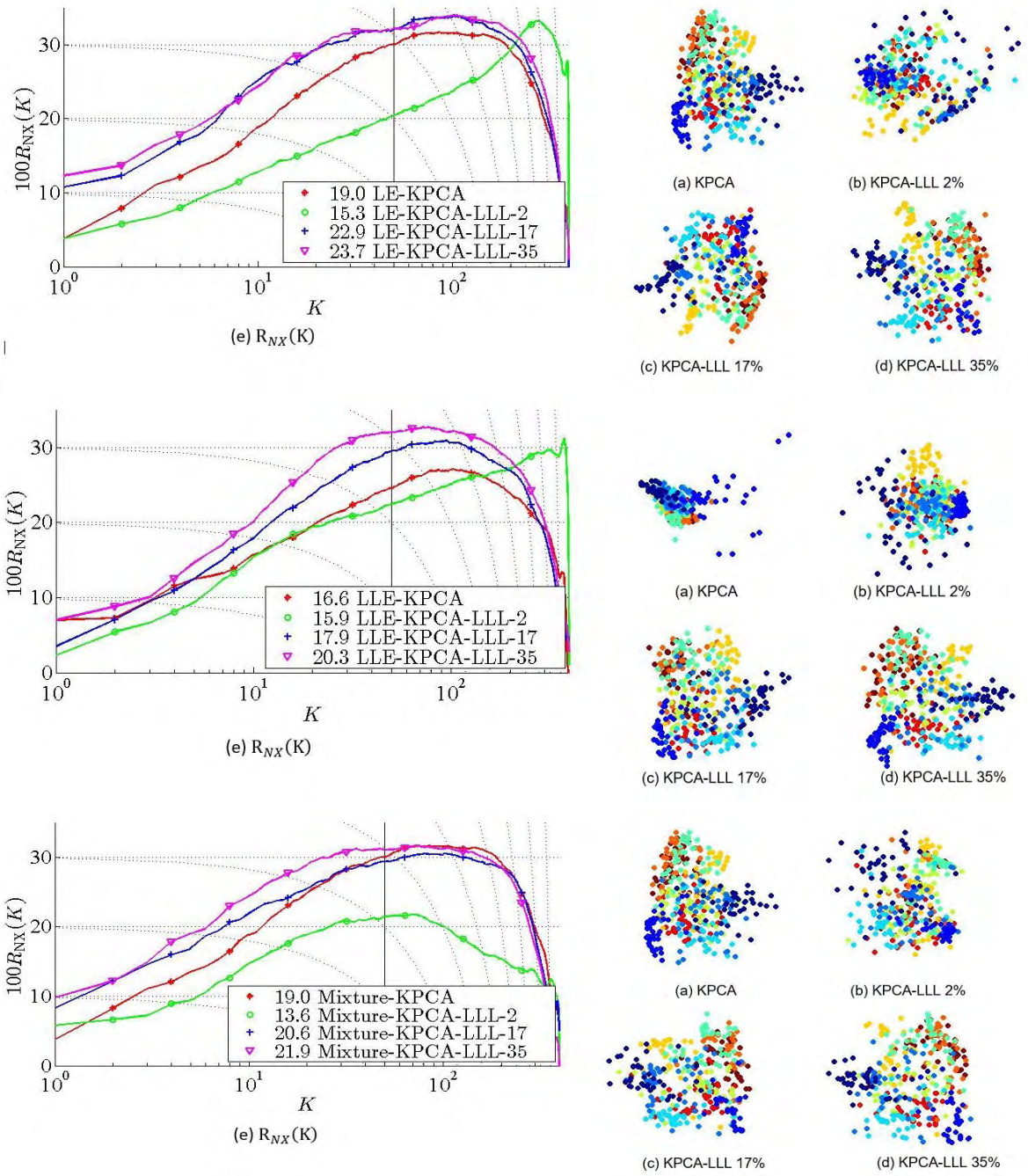
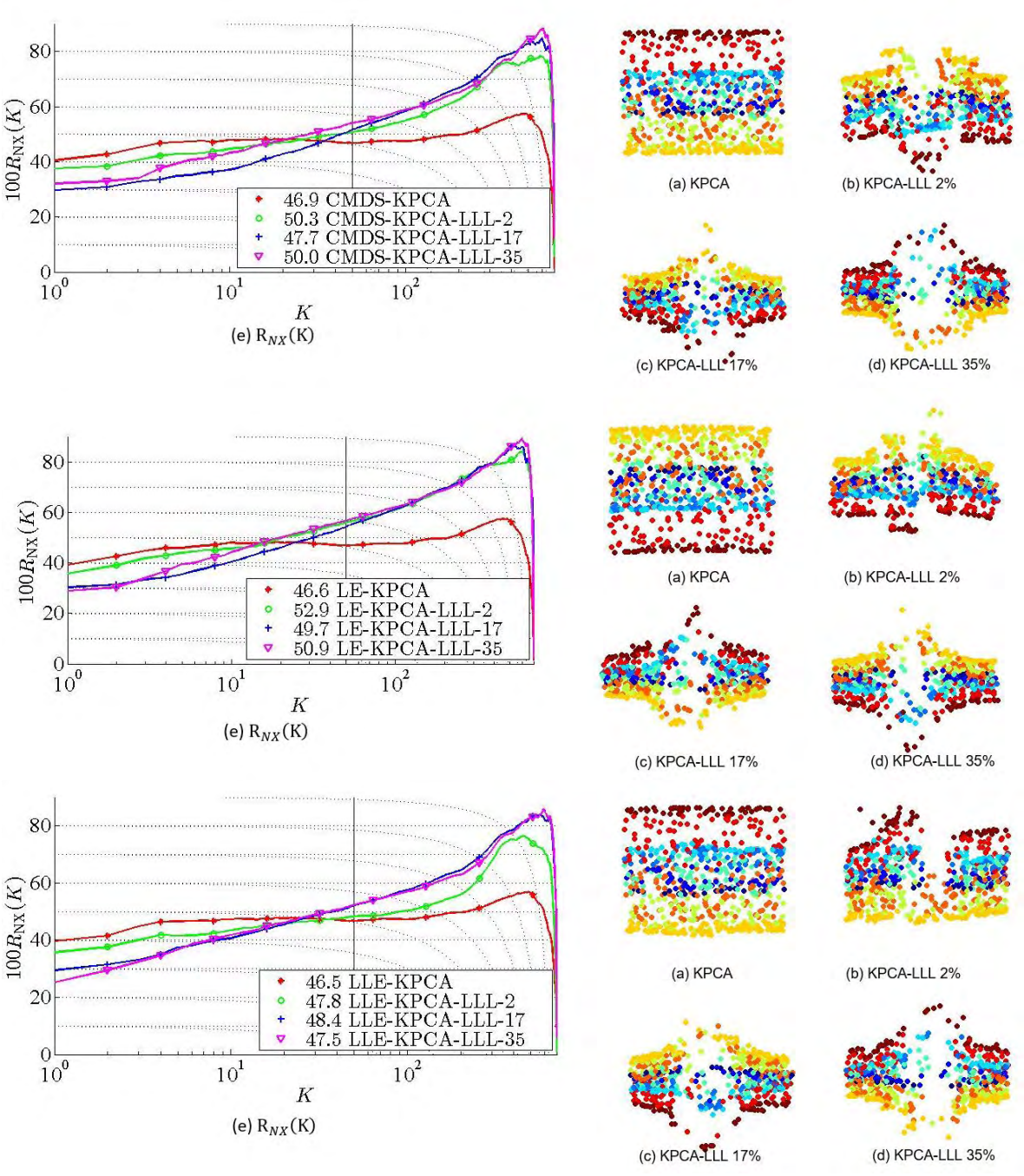


Figura 30. Resultados obtenidos para la base de datos Mnist con diferentes factores de ponderación (1) $\alpha_{CMDS} = 1\ y\ \alpha_{LLE} = \alpha_{LE} = 0$, (2) $\alpha_{LE} = 1\ y\ \alpha_{CMDS} = \alpha_{LLE} = 0$, (3) $\alpha_{LLE} = 1\ y\ \alpha_{CMDS} = \alpha_{LE} = 0$, (4) $\alpha_{LE} = \alpha_{CMDS} = \alpha_{LLE} = 0{,}33$, escogidos con el modelo de interacción cromático. En las figuras principales (1) a (4) cada ilustración (a) es el espacio embebido obtenido de la aplicación de *KPCA* y (b) a (d) son los espacios obtenidos por medio de KPCA-LLL con el número de submatrices al 2%, 17% y 35%. La figura (e) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$ de los datos embebidos de la (a) a (d). **Fuente**: Esta investigación.

En el caso de MNIST (**Figura 30.**) se obtiene un cambio en la curva de calidad en cualquiera de los escenarios en el que se use KPCA-LLL en el 2%, en este caso el área bajo la curva se ve afectada puesto que no se preserva la topología de los datos de igual manera que en los demás escenarios, en los cuales no se obtiene mayor diferencia pero esto dependerá de los fines del usuario.



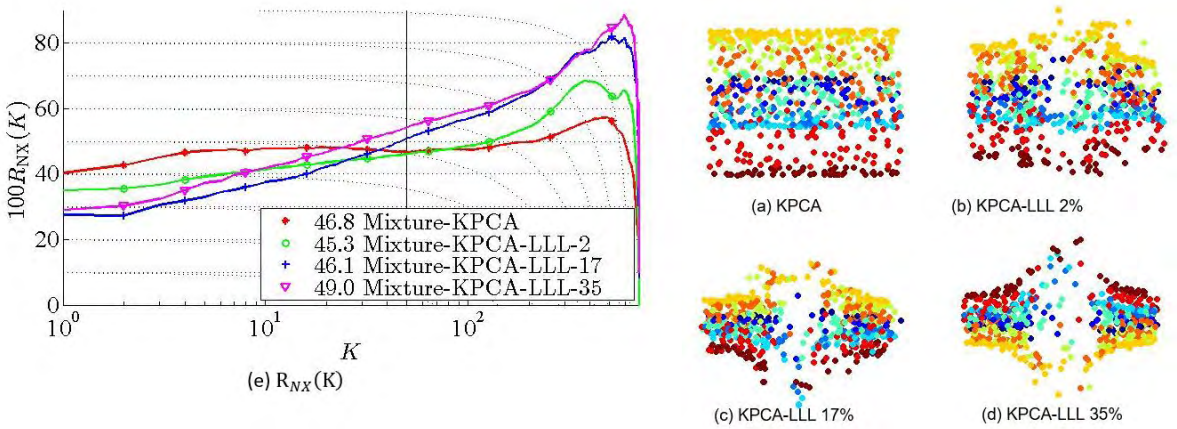


Figura 31. Resultados obtenidos para la base de datos rollo suizo con diferentes factores de ponderación (1) $\alpha_{CMDS} = 1\ y\ \alpha_{LLE} = \alpha_{LE} = 0$, (2) $\alpha_{LE} = 1\ y\ \alpha_{CMDS} = \alpha_{LLE} = 0$, (3) $\alpha_{LLE} = 1\ y\ \alpha_{CMDS} = \alpha_{LE} = 0$, (4) $\alpha_{LE} = \alpha_{CMDS} = \alpha_{LLE} = 0{,}33$, escogidos con el modelo de interacción cromático. En las figuras principales (1) a (4) cada ilustración (a) es el espacio embebido obtenido de la aplicación de KPCA y (b) a (d) son los espacios obtenidos por medio de KPCA-LLL con el número de submatrices al 2%, 17% y 35%. La figura (e) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$ de los datos embebidos de la (a) a (d). **Fuente**: Esta investigación.

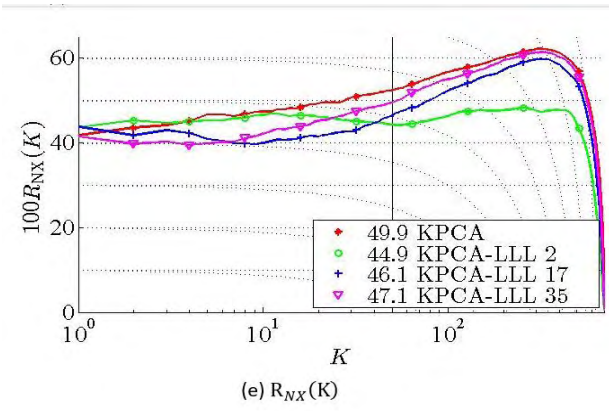
Para la base de datos del rollo suizo (**Figura 31**) se obtiene unos resultados resaltables puesto que en ninguno de los escenarios de este experimento se obtiene la misma representación si solo se usa KPCA. Sin embargo, en cuanto a el área bajo la curva de calidad en algunos casos se obtienen mayor área usando KPCA-LLL en 35% o 17%.

4.5.2 Resultados obtenidos para el modelo basado en ángulos

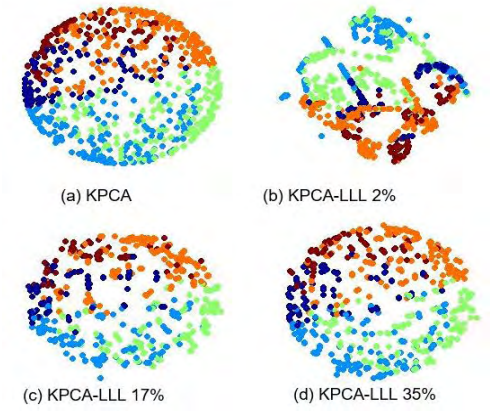
Para esta sección del experimento 2 se usan los mismos escenarios que para la **sección 4.4.2,** en este orden de ideas para cada base de datos se usó con el modelo basado en ángulos en las tres circunstancias presentadas en la **figura 24**. Los resultados obtenidos se muestran en las **figuras 32 a 34** En cierto modo los resultados obtenidos en esta sección comparados con los de la sección anterior (**sección 4.5.1**) muestran que el número de submatrices está ligado a la representación visual final, puesto que con menor número de estas las representaciones generadas no son muy similares a las generadas con KPCA, esto se evidencia en las tres bases de datos donde se presentan deformidades para KPCA-LLL con 2% de la cantidad de datos originales, problema que se soluciona si se incrementa hasta el 35%. Adicionalmente se midieron los tiempos de ejecución de cada uno de los escenarios de esta parte del experimento dos, evidentemente al aplicar KPCA-LLL el tiempo se ve disminuido aun así la representación que se obtiene con esta técnica puede no ser la misma y perder completamente la esencia de los datos.

Tabla 2. Tiempo de ejecución para los diferentes escenarios del experimento 2

Factores de ponderación	Base de datos	Tiempo con KPCA	Tiempo con KPCA-LLL 2%	Tiempo con KPCA-LLL 17%	Tiempo con KPCA-LLL 35%
$\alpha_{LLE} = 0{,}3$ $\alpha_{CMDS} = 0{,}25$ $\alpha_{LE} = 0{,}45$	Esfera 3D	5254 ± 1701	67,7 ± 6,3	548,4 ± 24,9	1712,4 ± 29,3
	Rollo suizo	3437,4 ± 377,6	70,1 ± 5,1	559,4 ± 27,6	1715,8 ± 52,1
	Mnist	509,6 ± 18,7	56,5 ± 5,6	224,7 ± 9,9	543,1 ± 14,8
$\alpha_{LLE} = 0{,}3$ $\alpha_{CMDS} = 0{,}4$ $\alpha_{LE} = 0{,}3$	Esfera 3D	3339,8 ± 665,1	67,1 ± 3,8	579,6 ± 73,8	1754,3 ± 96,8
	Rollo suizo	3118,3 ± 360,5	72,2 ± 4,1	541,8 ± 22,6	1818,9 ± 47,3
	Mnist	499,9 ± 43,3	55,1 ± 4,9	220,8 ± 11,6	562,7 ± 22,1
$\alpha_{LLE} = 0{,}3$ $\alpha_{CMDS} = 0{,}2$ $\alpha_{LE} = 0{,}5$	Esfera 3D	3469,4 ± 805,7	77,4 ± 11,5	580,4 ± 45,4	1801,3 ± 33,2
	Rollo suizo	2870,4 ± 198,5	78,7 ± 9,3	578,9 ± 28,4	1802 ± 54,7
	Mnist	463,7 ± 13,6	55,1 ± 8,8	222,3 ± 6,9	554,7 ± 20,4



(e) $R_{NX}(K)$



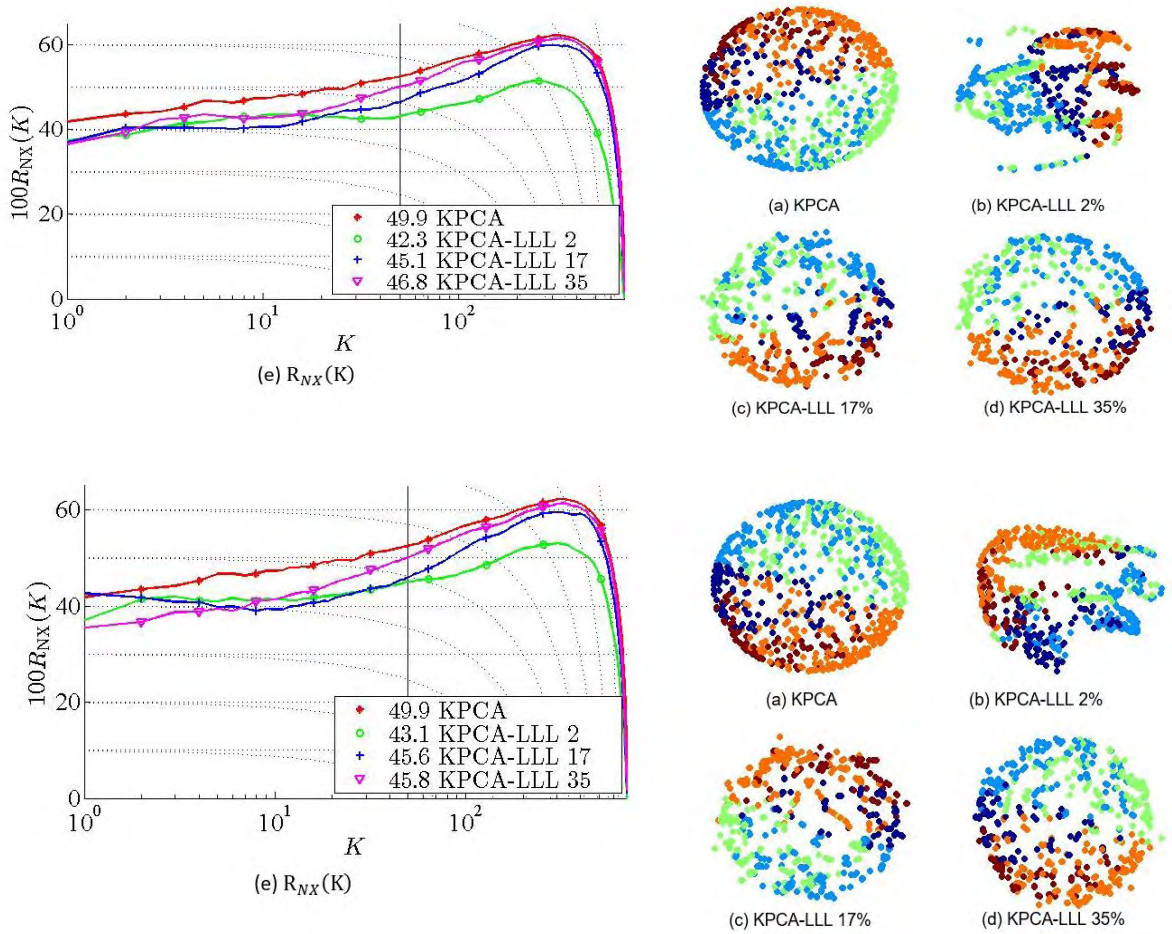


Figura 32. Resultados obtenidos para la base de datos cascaron esférico 3D con diferentes factores de ponderación (1) $\alpha_{CMDS} = 0{,}25\ \alpha_{LLE} = 0{,}3\ \alpha_{LE} = 0{,}45$, (2) $\alpha_{CMDS} = 0{,}4\ \alpha_{LLE} = 0{,}3\ \alpha_{LE} = 0{,}3$, (3) $\alpha_{CMDS} = 0{,}2\ \alpha_{LLE} = 0{,}5\ \alpha_{LE} = 0{,}3$, escogidos con el modelo de interacción basado en ángulos. En las figuras principales (1) a (3) cada ilustración (a) es el espacio embebido obtenido de la aplicación de KPCA y (b) a (d) son los espacios obtenidos por medio de KPCA-LLL con el número de submatrices al 2%, 17% y 35% respectivamente. La figura (e) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$ de los datos embebidos de la (a) a (d). **Fuente**: Esta investigación.

Se puede evidenciar de acuerdo con la **Tabla 2** y la **Figura 32**, que el tiempo de ejecución para la base de datos del cascaron esférico se reduce en hasta un 90% si se usa solo el 2% de la totalidad de muestras de la base de datos, sin embargo esto puede afectar la representación final, puesto que esta se ve afectada y la topología de los datos no se preserva de igual manera que para 17% o 35% de los datos totales. Sin embargo, no se obtiene mucha diferencia en cuanto al área bajo la curva entre el algoritmo generalizado de KPCA y KPCA-LLL con el 35% siendo este un buen resultado en menos de la 3 parte del tiempo de ejecución.

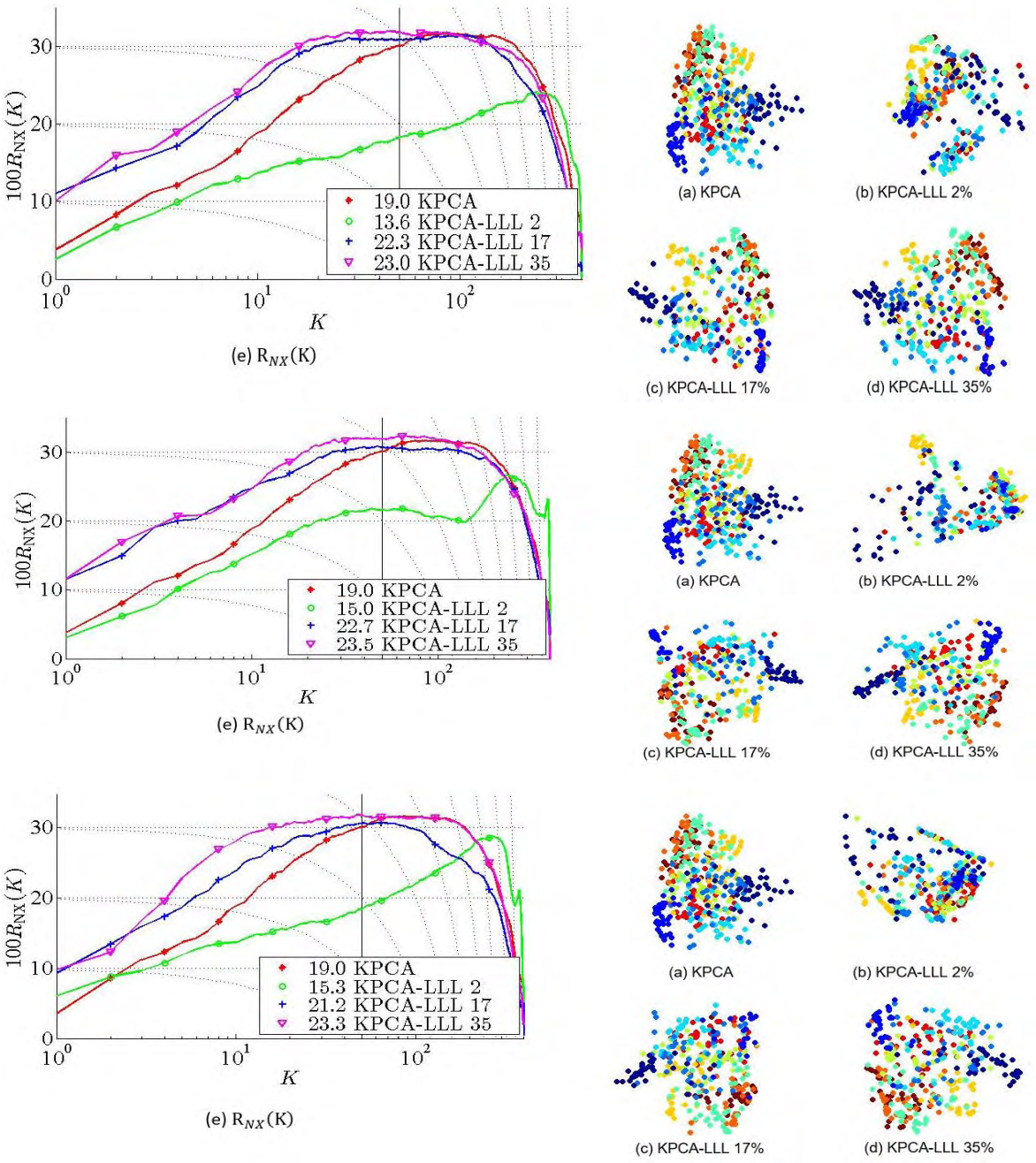
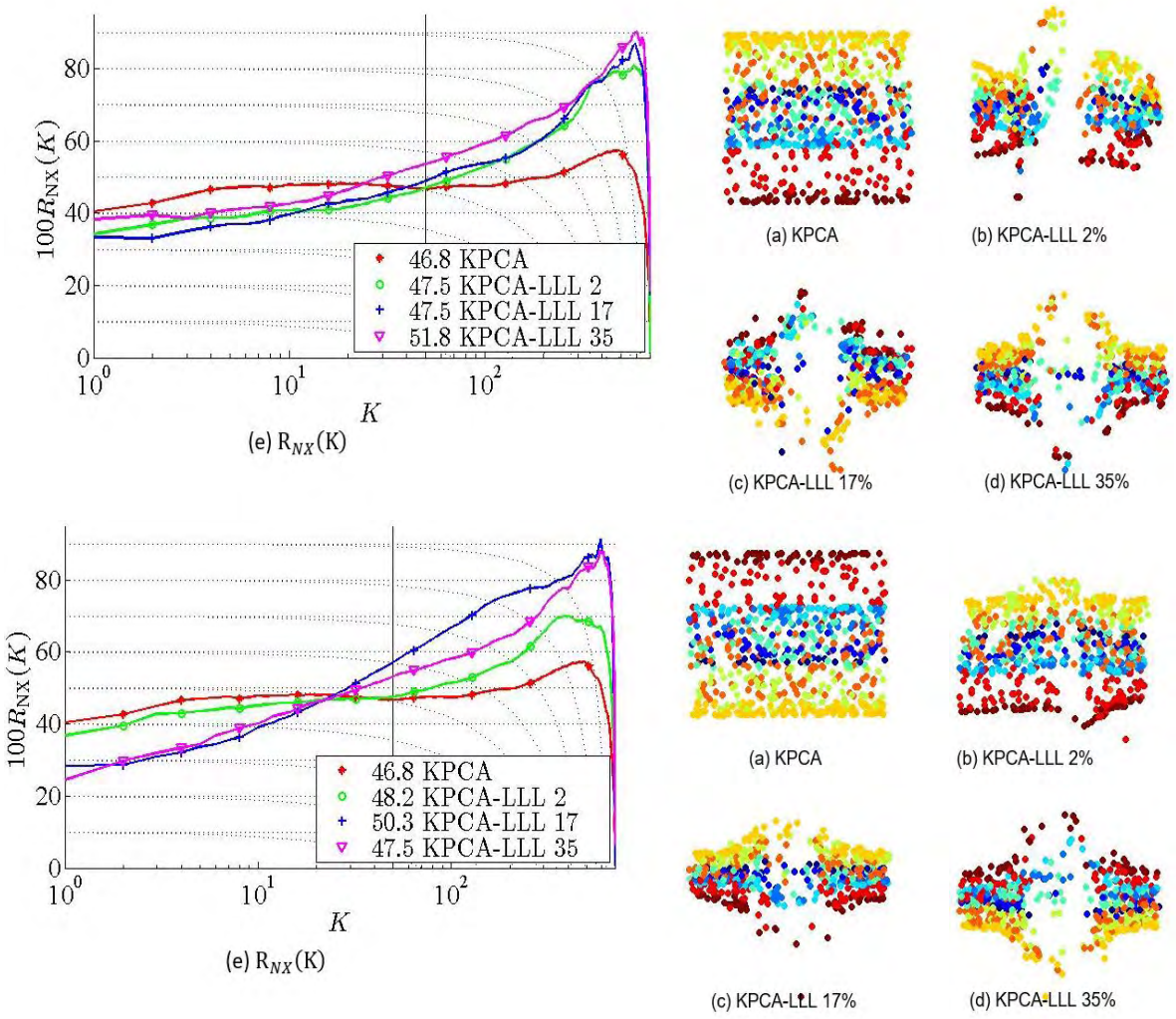


Figura 33. Resultados obtenidos para la base de datos Mnist con diferentes factores de ponderación (1) $\alpha_{CMDS} = 0{,}25$ $\alpha_{LLE} = 0{,}3$ $\alpha_{LE} = 0{,}45$, (2) $\alpha_{CMDS} = 0{,}4$ $\alpha_{LLE} = 0{,}3$ $\alpha_{LE} = 0{,}3$, (3) $\alpha_{CMDS} = 0{,}2$ $\alpha_{LLE} = 0{,}5$ $\alpha_{LE} = 0{,}3$, escogidos con el modelo de interacción basado en ángulos. En las figuras principales (1) a (3) cada ilustración (a) es el espacio embebido obtenido de la aplicación de KPCA y (b) a (d) son los espacios obtenidos por medio de KPCA-LLL con el número de submatrices al 2%, 17% y 35% respectivamente. La figura (e) indica la medida de

calidad $\mathrm{R}_{NX}(\mathrm{K})$ de los datos embebidos de la (a) a (d). **Fuente**: Esta investigación.

En la base de datos MNIST se encuentra una particularidad, puesto que los resultados obtenidos con KPA-LLL en el 17% y 35% mejoran la preservación de la topología de los datos obteniendo mayor área bajo la curva que el algoritmo sin optimización KPCA. Sin embargo, para KPCA-LLL en el 2% se ve una disminución significativa en el área bajo la curva y en la representación final obtenida que según las finalidades del usuario pueda afectar la visualización requerida.



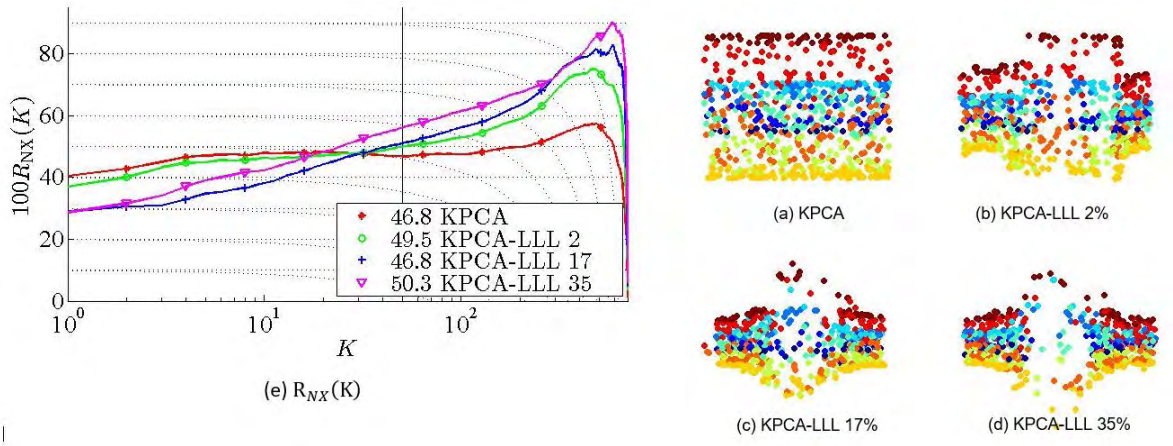


Figura 34. Resultados obtenidos para la base de datos rollo suizo con diferentes factores de ponderación (1) $\alpha_{CMDS} = 0{,}25\ \alpha_{LLE} = 0{,}3\ \alpha_{LE} = 0{,}45$, (2) $\alpha_{CMDS} = 0{,}4\ \alpha_{LLE} = 0{,}3\ \alpha_{LE} = 0{,}3$, (3) $\alpha_{CMDS} = 0{,}2\ \alpha_{LLE} = 0{,}5\ \alpha_{LE} = 0{,}3$, escogidos con el modelo de interacción basado en ángulos. En las figuras principales (1) a (3) cada ilustración (a) es el espacio embebido obtenido de la aplicación de KPCA y (b) a (d) son los espacios obtenidos por medio de KPCA-LLL con el número de submatrices al 2%, 17% y 35% respectivamente. La figura (e) indica la medida de calidad $\mathrm{R}_{NX}(\mathrm{K})$ de los datos embebidos de la (a) a (d). **Fuente**: Esta investigación.

4.6.3 Discusión

Como se pudo apreciar en los resultados obtenidos en el experimento 2 es posible reducir el costo computacional si no se considera todo el conjunto de datos de entrada para el proceso de la RD espectral, si no una porción de esta que se representa con el número de submatrices L, sin embargo este parámetro representa un factor muy importante a la hora de aplicar KPCA-LLL ya que se debe escoger el correcto número de submatrices para aproximar el conjunto original sin pérdida significativa de información. En la mayoría de las relaciones presentadas en las curvas de calidad en las **figuras 29 a 31** de la primera parte de este experimento, los espacios embebidos obtenidos con KPCA-LLL no varían mucho con respecto a los obtenidos con KPCA, sin embargo la segunda parte (**Figuras 32 a 34**) al aplicar KPCA-LLL con bajo número de submatrices (entre el 2% a 17% de la base de datos original) afecta a la representación obtenida ya que esta cambia de forma ocasionando que el usuario no pueda apreciar de manera correcta la topología del espacio original, este es el caso de rollo suizo cuyo espacio embebido se deforma cuando se aplica dicho método. Por lo anterior se recomienda que se use más del 20% de submatrices para que la representación que se genere sea interactiva y correcta, sin embargo, esto depende de la base de datos en la que se esté trabajando.

4.7 EXPLORACIÓN SEGUNDA TÉCNICA DE VISUALIZACIÓN

Para la prueba de la segunda técnica de visualización implementada (Coordenadas paralelas) se uso la base de datos MNIST (**Figura 16c**) puesto que esta cuenta con 784 dimensiones y la técnica de coordenadas paralelas fue implementada para que se hagan visualizaciones en mas de 3 dimensiones.

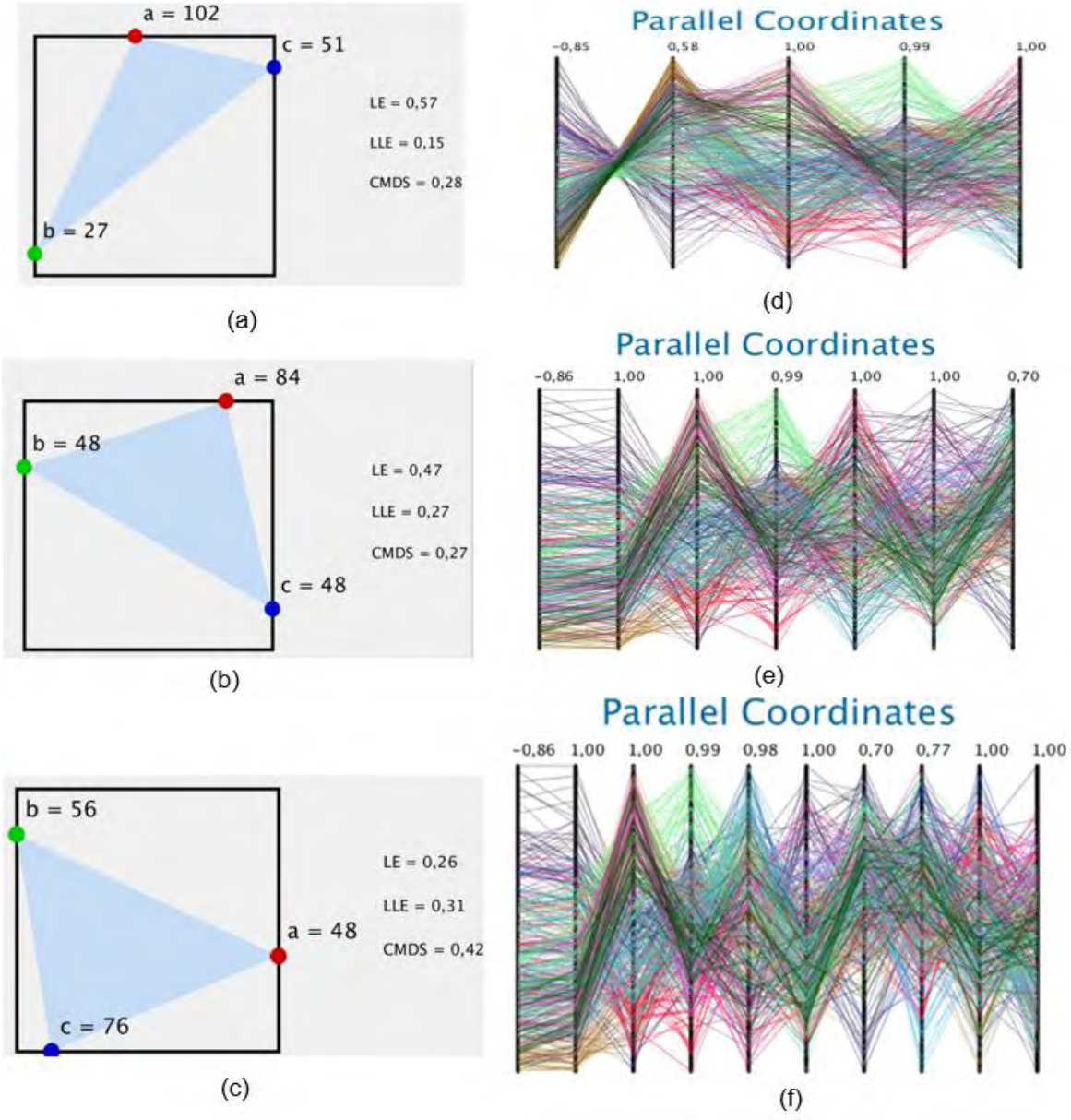


Figura 35. Resultados obtenidos al aplicar 3 combinaciones diferentes (a, b, c) en

la base de datos MNIST, las representaciones obtenidas son con la técnica de visualización coordenadas paralelas para 4 (d), 7 (e) y 10 (f) dimensiones reducidas de 784 datos originales de la base de datos.

Se escogieron 3 diferentes combinaciones haciendo uso del modelo de interacción basado en ángulos (**Figura 35 (a), (b), (c)**), para cada combinación se obtuvo una representación diferente en 4 dimensiones (**Figura 35 (d)**), 7 (**Figura 35 (e)**) y 10 dimensiones (**Figura 35 (f)**) siendo esta ultima la máxima permitida. Se evidencia que esta técnica permite al usuario tener un rango mayor de visualización que solo en un plano bidimensional el cual muchas veces suele ser un poco más sencillo, pero limitante según la aplicación que requiera el usuario.

5. CONCLUSIONES

Existe una necesidad latente del proponer herramientas que ayuden a los seres humanos a entender la gran cantidad de datos que hoy en día en muchas disciplinas de la vida diaria se recolectan. Por lo anterior la herramienta de visualización propuesta les permite a los usuarios inexpertos interactuar de manera dinámica y didáctica con los datos, teniendo la oportunidad no solo de apreciar la representación en baja dimensión sino además interactuar con el proceso de la mezcla de métodos espectrales de RD quienes al ser más versátiles son representados mediante aproximaciones kernel. Si bien la aplicación de métodos de RD puede ser ambigua para un usuario inexperto, la herramienta propuesta representa una forma alternativa de visualización de datos donde con la ayuda de los modelos de interacción y las técnicas de visualización se reduce la brecha entre los usuarios, los repositorios de datos, y la reducción de dimensión.

Se puede concluir, basados en los resultados presentados en la **sección 4** y en el **anexo 3** como parte de esta investigación, que la mezcla de métodos de RD en sus aproximaciones en matrices kernel, permite al usuario tener diferentes representaciones en baja dimensión aumentando las posibilidades de obtener la adecuada bajo su criterio. Dichas representaciones preservan en gran medida la topología de los datos, ofreciendo al usuario espacios embebidos más aproximados a sus datos originales.

El rendimiento y los espacios de baja dimensión obtenidos a partir del uso de la herramienta dependerá siempre de la topología de los datos en cuestión. Por ejemplo: en bases de datos como el cascarón esférico 3D no se obtuvieron muchos cambios en cuanto a la forma de representación de sus espacios embebidos. En contraste, para bases de datos como el rollo suizo y MNIST dependiendo de los factores de ponderación escogidos, se obtiene una variación tanto para las representaciones en baja dimensión como en la preservación de la topología de los datos originales.

Dadas las características fundamentales de la herramienta -genérica, versátil y de software libre-, está resulta ser única en su clase, puesto que a la fecha no se encuentran desarrolladas herramientas de dedicación exclusiva a la visualización interactiva de datos, que se fundamente en la reducción de dimensión incorporando modelos de interacción y técnicas de visualización. Además, está resulta ser un software de apoyo para el análisis de datos que brinda un ambiente intuitivo y didáctico al usuario, quien interactúa directamente con el proceso de visualización sin tener mucho conocimiento de las metodologías utilizadas.

A pesar de que las aproximaciones kernel son una adecuada y versátil forma de implementar métodos RD espectrales, uno de los principales problemas que

presentan estas es que después del cálculo de la matriz de afinidad (de similitud , disimilitud o kernel) que es de dimensión $N\ x\ N$ donde N es el número de registros de los datos, se debe realizar una descomposición espectral de dicha matriz, lo que implica un costo computacional elevado, reduciendo la interactividad que no permite una representación dinámica de los datos. En efecto al trabajar con bases de datos con elevado número de registros se es recomendable el uso de un algoritmo que disminuya este tiempo de ejecución que es el caso de las sub-matrices localmente lineales implementadas en la herramienta. Este resultó ser de gran utilidad puesto que reduce el tiempo de ejecución de la descomposición espectral en hasta un 80%, con espacios embebidos similares a los obtenidos con el algoritmo generalizado de PCA. El usuario entonces podrá evidenciar cambios en tiempo real en los espacios embebidos aun si su repositorio cuenta con grande número de registros.

6. RECOMENDACIONES

Debido a que la generación de bases de datos y la facilidad de acceder a estos es una tarea interdisciplinar, es importante el explorar otras modelos de interacción, métodos de reducción de dimensión y algoritmos o metodologías que estén orientados a la reducción del costo computacional de descomposición espectral para que la representación de datos sea de tipo dinámica con la menor perdida de información y la mayor conservación de la topología de los datos originales.

Se recomienda el aplicar algoritmos o estrategias que permitan determinar el número de submatrices que pueden representar el espacio original con la menor cantidad de datos posibles como LLL. Este valor depende no solo del tipo de datos si no de la cantidad de $\boldsymbol{Kz}$, por lo cual se recomienda escoger un d+1< $\boldsymbol{Kz}$ < D+1 donde d es la dimensión a la que se quiere reducir y el parámetro $\boldsymbol{L}$ mayor al 20% de la base de datos original, según los resultados obtenidos en esta investigación.

.

BIBLIOGRAFÍA

[1] Han, Jiawei, Jian Pei, y Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[2] Leskovec, Jure, Anand Rajaraman, y Jeffrey David Ullman. Mining of massive datasets. Cambridge university press, 2014

[3] Turner, Vernon, et al. The digital universe of opportunities: Rich data and the increasing value of the internet of things. IDC Analyze the Future. Pp. 5. 2014.

[4] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[5] Daza Santacoloma, Genaro. Metodología de reducción de dimensión para sistemas de reconocimiento automático de patrones sobre bioseñales. Diss. Universidad Nacional de Colombia-Sede Manizales, 2006.

[6] Peña-Unigarro, D. F. Metodología de visualización interactiva de datos de alta dimensión a partir de un modelo intuitivo de reducción de dimension. Pasto, 2016. Trabajo de grado (Ingeniería electrónica). Universidad de Nariño-Colombia. Facultad de ingeniería

[7] Fayyad, Usama M., Andreas Wierse, y Georges G. Grinstein, eds. Information visualization in data mining and knowledge discovery. Morgan Kaufmann, 2002.

[8] Gisbrecht, Andrej, y Barbara Hammer. "Data visualization by nonlinear dimensionality reduction." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5.2 (2015): 51-73.

[9] Fayyad, Usama M. "Data mining and knowledge discovery: Making sense out of data." IEEE Expert: Intelligent Systems and Their Applications 11.5 (1996): 20-25.

[10] Cios, Krzysztof J., et al. Data mining: a knowledge discovery approach. Springer Science & Business Media, 2007.

[11] Murray, Scott. Interactive Data Visualization for the Web: An Introduction to Designing with. " O'Reilly Media, Inc.", 2017.

[12] M. Bostock y J. Heer, «Protovis: A graphical toolkit for visualization,» IEEE transactions on visualization and computer graphics , pp. 1121-1128, 2009.

[13] J. A. Lee y M. Verleysen, «Quality assessment of dimensionality reduction: Rankbased criteria,» Neurocomputing, vol. 72, nº 7, 2009

[14] Vladymyrov, Max, and Miguel Á. Carreira-Perpinán, «Locally linear landmarks for large-scale manifold learning,» Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2013.

[15] M. Sedlmair, T. Munzner y M. Tory, «Empirical Guidance on Scatterplot and Dimension Reduction,» IEEE Transactions on Visualization and Computer Graphics, vol. 19, no 12, pp. 2634-2643, 2013.

[16] S. H. Bae, J. Qiu y G. Fox, «High performance multidimensional scaling for large high-dimensional data visualization,» IEEE Transaction of Parallel and Distributed System., 2012.

[17] D. Z. L. Sacha, M. L. J. A. Sedlmair, J. Peltonen, D. Weiskopf y D. A. Keim, «Visual interaction with dimensionality reduction: a structured literature analysis,» IEEE Transactions on Visualization and Computer Graphics, 2016.

[18] M. Sedlmair, M. Brehmer, S. Ingram y T. Munzner, «Dimensionality reduction in the wild: Gaps and guidance,» Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., 2012.

[19] M. Sedlmair y M. Aupetit, «Data-driven Evaluation of Visual Quality Measures,» Computer Graphics Forum, vol. 34, nº 3, pp. 201-210, 2015

[20] G. Shmueli, N. R. Patel y P. C. Bruce, Data Mining for Business Analytics: Concepts, Techniques, and Applications in XLMiner, 2016.

[21] M. Scholz, «Approaches to analyse and interpret biological profile data,» Universitat Potsdam, 2006.

[22] W. Dai y P. Hu, «Research on Personalized Behaviors Recommendation System Based on Cloud Computing,» Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, nº 2, pp. 1480-1486, 2013.

[23] M. L. Kersten, S. Idreos, S. Manegold y E. Liarou, «The researcher's guide to the data deluge: Querying a scientific database in just a few seconds.,» PVLDB Challenges and Visions, 2011.

[24] C. Ware, Information visualization: perception for design, Elsevier, 2012.

[25] González-Torres, A., F. J. García-Peñalvo y R. Therón, «Human–computer interaction in evolutionary visual software analytics,» Computers in Human Behavior, vol. 29, nº 2, pp. 486-495, 2013.

[26] J. R. Harger y P. J. Crossno, «Comparison of open-source visual analytics toolkits,» In IS&T/SPIE Electronic Imaging International , 2012.

[27] D. H. Peluffo Ordóñez, A. E. Castro Ospina, J. C. Alvarado Pérez y E. J. Revelo Fuelagán, «Multiple Kernel Learning for Spectral Dimensionality Reduction,» IberoAmerican Congress on Pattern Recognition (CIARP), pp. 626-634, 2015.

[28] D. H. Peluffo-Ordóñez, J. A. Lee y M. Verleysen, «Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In Computational Intelligence and Data Mining,» Computational Intelligence and Data Mining (CIDM), pp. 171-177, 2014.

[29] D. H. Peluffo Ordoñez, J. A. Lee y M. Verleysen, « Recent methods for dimensionality reduction: a brief comparative analysis,» European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2014.

[30] Salazar castro, J. A. Rosas Narvaez, Y. C. Implementación de una interfaz de visualización de datos eficiente e interactiva a partir de una perspectiva geométrica. Pasto, 2015. Trabajo de grado (Ingeniería electrónica). Universidad de Nariño-Colombia. Facultad de ingeniería

[31] J. Ham, D. D. Lee, S. Mika y B. Schölkopf, «A kernel view of the dimensionality reduction of manifolds.,» Proceedings of the twenty-first international conference on Machine learning (ICML), p. 47, 2004.

[32] J. Gijón Gómez, «Visualización bidimensional de problemas de clasificación en alta dimensión,» PROYECTO FIN DE CARRERA. UNIVERSIDAD CARLOS III DE MADRID, 2013.

[33] SAS. 2012. Visualización de datos: hacer Big Data accesible y valioso. Papel blanco. Una encuesta sobre visualización de información: avances y desafíos recientes. (Noviembre 2015)

[34] D. H. Peluffo-Ordónez, J. A. Lee y M. Verleysen, «Dimensionality reduction methods based on stochastic neighbour embedding. In Advances in Self-Organizing Maps and Learning Vector Quantization,» Springer International Publishing, pp. 65-74, 2014.

[35] V. Vasudevan, "Supporting High Bandwidth Navigation in Object-Bases," Proc. 10th Int'l Conf. Data Engineering, Houston, Tex., pp. 294-301,1994.

[36] I.T. Jolliffe, Principal Component Analysis (Springer-Verlag, New York, 1989).

[37] T. Cox and M. Cox. Multidimensional Scaling (Chapman & Hall, London, 1994)

[38] Saul, Lawrence K., and Sam T. Roweis. "An introduction to locally linear embedding." unpublished. Available at: http://www. cs. toronto. edu/~ roweis/lle/publications. html (2000).

[39] M. Belkin y P. Niyogi, «Laplacian eigenmaps for dimensionality reduction and data representation.,» Neural computation, vol. 15, nº 6, pp. 1373-1396, 2003.

[40] Ward, M. O., Grinstein, G., & Keim, D. (2010). Interactive data visualization: foundations, techniques, and applications. CRC Press

[41] Keim, D. A. (2002). Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics, 8(1), 1-8.

[42] Keim, D. A. (1996). Pixel-oriented visualization techniques for exploring very large data bases. Journal of Computational and Graphical Statistics, 5(1), 58-77

[43] Keim, D. A., & Kriegel, H. P. (1996). Visualization techniques for mining large databases: A comparison. IEEE Transactions on knowledge and data engineering, 8(6), 923-938.

[44] R.A. Becker, S.G. Eick, and G.J. Wills, "Visualizing Network Data," IEEE Trans. Visualizations and Graphics, vol. 1, no. 1, pp. 1,628,1995.

[45] Kaski, S., & Peltonen, J. (2011). Dimensionality reduction for data visualization [applications corner]. IEEE Signal Processing Magazine, 28(2), 100-104.

[46] Aravindh Krishnamoorthy, «Symmetric QR Algorithm with Permutations,» arXiv preprint arXiv:1402.5086 (2014).

[47] John GF Francis, «The QR transformation a unitary analogue to the LR transformation (part 1),» The Computer Journal 4.3 (1961), pp. 265–271.

[48] John GF Francis, «The QR transformation (part 2),» The Computer Journal 4.4 (1962), pp.

332–345.

[49] Vera N Kublanovskaya, «On some algorithms for the solution of the complete eigenvalue problem,» USSR Computational Mathematics and Mathematical Physics 1.3 (1962), pp. 637–657

[50] A. Sameh, «Jacobi and Jacobi-like algorithms for a parallel computer,» Mathematics of Computation, 25:579-590, 1971

[51] Si, Si, et al, « Multi-scale spectral decomposition of massive graph,» Advances in Neural Information Processing Systems. 2014.

[52] B. N. Parlett, «The Symmetric Eigenvalue Problem, » Prentice-Hall, 1980 Williams, Seeger, «Using the Nystrom method to speed up kernel machines,» Advances in Neural Information Processing Systems (NIPS). Volume 13., MIT Press, Cambridge, MA (2001) 682–688

[53] Drineas, P., Mahoney, «On the Nystrom method for approximating a Gram matrix for improved kernel-based learning,» J. Machine Learning Research 6 (December 2005) 2153–2175

[54] J. A. Lee and M. Verleysen, Nonlinear dimensionality reduction. Springer Science & Business Media, 2007

[55] H. Strange and R. Zwiggelaar, Open Problems in Spectral Dimensionality Reduction. Springer, 2014.

[56] M L. A. Belanche, «Developments in kernel design,» EESANN, 2013.

[57] J. Cook, I. Sutskever, A. Mnih y G. E. Hinton, «Visualizing Similarity Data with a Mixture of Maps,» AISTATS, vol. 7, pp. 67-74, 2007

[58] C. M. Bishop, Bishop Pattern Recognition and Machine Learning., New York: Springer, 2006

[59] Peña-ünigarro, Diego F., et al. "Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction." Signal Processing, Images and Artificial Vision (STSIVA), 2016 XXI Symposium on. IEEE, 2016.

[60] R. C. Gonzalez y R. E. Woods, Digital image processing, (2002).

[61] M. Meyer, A. Barr, H. Lee y M. Desbrun, «Generalized barycentric coordinates on irregular polygons. Journal of graphics tools,» Journal of graphics tools, vol. 7, nº 1, pp. 13-22, 2002

[62] N. Cristianini y J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, UK: Cambridge university press, 2000.

[63] Y. Aflalo y R. Kimmel, «Spectral multidimensional scaling.,» Proceedings of the National Academy of Sciences, vol. 110, nº 45, pp. 18052-18057, 2013.

[64] D. Shiffman, Learning Processing, Elsevier, 2008

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 5.

ANEXO 1. LISTA DE ACRÓNIMOS

DCBD Descubrimiento de conocimiento en base de datos

KDD *Knowledge Discovery in Databases*

RD Reducción de dimensión

DR *Dimensionality Reduction*

RGB Rojo (*red*), verde (*green*) y azul (*blue*)

3-D Tres dimensiones

2-D Dos dimensiones

PCA Análisis de componentes principales (*principal component analysis*)

LLE *Locally Linear Embedding*

CMDS *Classical Multidimensional Scaling*

LE *Laplacian Eigenmaps*

KPCA *Kernel principal component analysis*

LLL *Locally lineal landmarks*

ANEXO 2. EJECUTABLE PARA EL USO DE LA HERRAMIENTA

Para la implementación de la herramienta se usa la plataforma de NetBeans bajo el lenguaje de programación Java. Sin embargo, para la implementación de los modelos de interacción y la técnica de visualización coordenadas paralelas se utilizó el software processing que al estar basado en Java resulta muy fácil el emigrar aplicaciones de una plataforma a otra. La principal librería que se utilizo para el desarrollo de la

herramienta es la librería JAMA[1] que permite realizar operaciones entre matrices y calcular la descomposición espectral de la mezcla de matrices kernel.

El código fuente de la aplicación en NetBeans y las librerías necesarias para el correcto funcionamiento, además, del archivo ejecutable para probar la aplicación sin necesidad de tener un programa en específico instalado en el equipo, están disponibles en la página web descrita en el **ANEXO 5.**

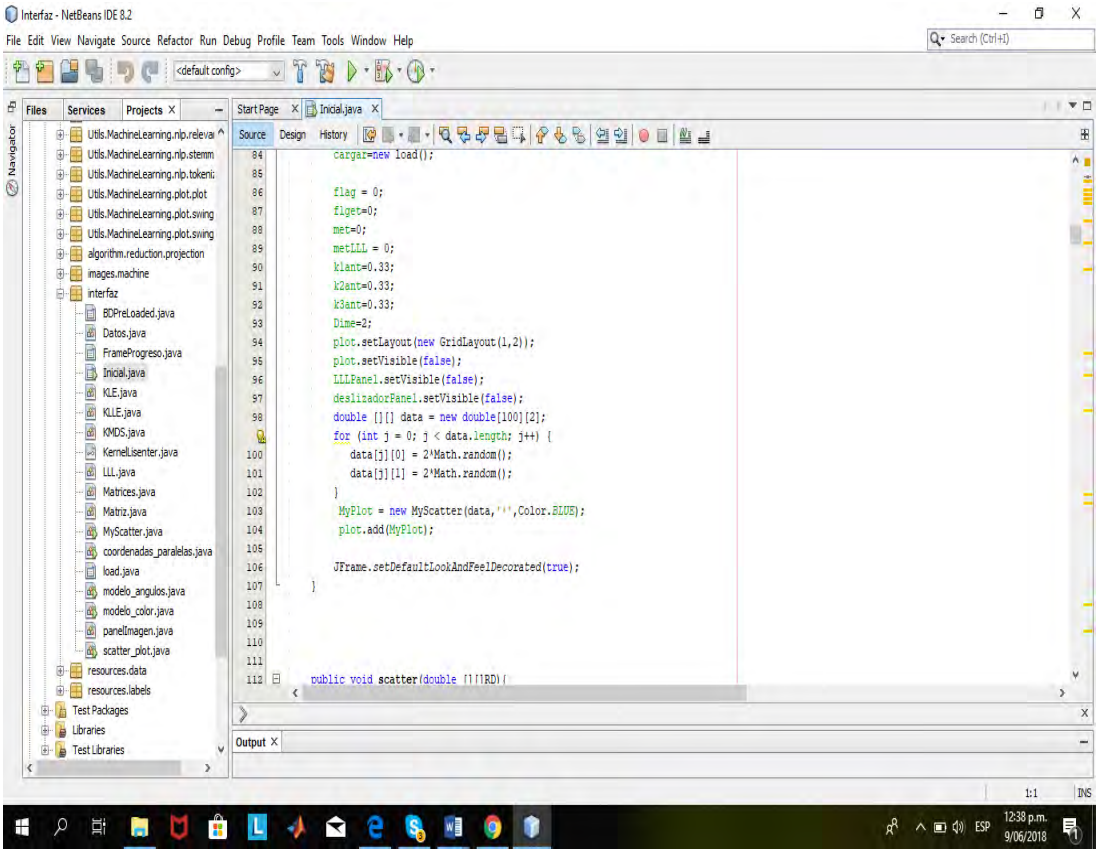


Figura 36. Frame principal de la herramienta que contiene todo lo implementado en processing y lo implementado en Java. Este archivo y las librerías necesarias están disponibles en la página web. **Fuente:** Esta investigación

[1] https://math.nist.gov/javanumerics/jama/

ANEXO 3. ARTICULO: CCC'18

Comparative analysis between embedded-spaces-based and kernel-based approaches for interactive data representation

C. K. Basante-Villota, C. M. Ortega-Castillo, J. E. Revelo-Fuelagán[1] and D. H. Peluffo-Ordóñez[2,3]

[1] Universidad de Nariño, Colombia.
[2] Corporación Universitaria Autónoma de Nariño, Colombia.
[3] Yachay Tech , Ecuador

Abstract. This work presents a comparative analysis between the linear combination of em-bedded spaces resulting from two approaches: 1) The application of dimensional reduction methods (DR) in their standard implementations, and 2) Their corresponding kernel-based approximations. Namely, considered DR methods are: CMDS (Classical Multi- Dimensional Scaling), LE (Laplacian Eigenmaps) and LLE (Locally Linear Embedding). This study aims at determining -through objective criteria- what approach obtains the best performance of DR task for data visualization. The experimental validation was performed using four databases from the UC Irvine Machine Learning Repository. The quality of the obtained embedded spaces is evaluated regarding the $R_{NX}(K)$ criterion. The $R_{NX}(K)$ allows for evaluating the area under the curve, which indicates the performance of the technique in a global or local topology. Additionally, we measure the computational cost for every comparing experiment. A main contribution of this work is the provided discussion on the selection of an interactivity model when maturing DR methods, which is a crucial aspect for information visualization purposes.

Keywords: Artificial intelligence, dimensionality reduction methods, kernel, Kernel PCA, CMDS, LLE, LE.

1 Introduction

Nowadays, the large volumes of data are accompanied by the need of powerful tools for analysis and representation, as, you could have a dense repository of data, but without the appropriate tools the information obtained may not be very useful [1]. The need arises to find different techniques and tools that help researchers or analysts in tasks such as obtaining useful patterns for large volumes of data, these tools are the subject of an emerging field of research known as Knowledge Discovery in Bases of Data (KDD). Dimension reduction (DR) is considered within the KDD process as a pre-processing stage because it projects the data to a space where the original data is represented with fewer attributes or characteristics, preserving the greater intrinsic information of the original data to enhance tasks such as data mining and machine learning. For example,

in classification tasks knowing the representation of the data as well as knowing whether these have separability characteristics, make easier to engage and interpret by the user [2], [3].

We have two method PCA (Principal Component Analysis) and the CMDS (Classical Multi-Dimensional Scaling) which are part of those classic RD methods whose objective is to preserve variance or distance [4]. Recently, the focus of DR methods is based on criteria aimed at preserving the data topology. A topology of this type could be represented in an undirected and weighted graph based on data constructed whose points represent the nodes, and their edge's weights are contained in an affinity and non-negative similarity matrix. This representation is leveraged by methods based on spectral and divergence approaches, for the spectral approach we can represent the weights of the distances in a similarity matrix, such as with the LE (Laplacian Eigenmaps) method [5] and using a matrix of unsymmetrical similarity and focusing on the local structure of the data, the method called LLE (Locally Linear Embedding) arises [6]. There is also the possibility of working on the high-dimensional space with the advantage of greatly enhancing the representation and the embedded data visualization of the original space mapped to the high-dimensional space, from the calculation of the eigen decomposition. An estimate of the inner product (kernel) can be designed based on the function and application which one wants to develop [7], in this work the kernel matrices will represent distance or similarity functions associated with a dimension reduction method.

In this research three spectral dimension reduction methods are considered, trying to encompass different criteria which CMDS, LLE and LE are based on, these are used under two approaches, one of them is the representation of their embedded spaces obtained from their standard algorithms widely explained in [5], [8], [6], and the second is based on the kernel approaches of the same methods. After obtaining each of the embedded spaces, a linear weighting is performed for combine the different approaches leveraging each of the RD methods, the same is done for the kernel ma-trices obtained from the approximations of the spectral methods. subsequently the Kernel PCA technique is applied to reduce the dimension to obtain the embedded space from the combination of the kernel-based approach. The combination of embedded spaces already obtained from the RD methods is not clear and intuitive mathematically, on the other hand, the linear combination of kernel or similarity matrices which are represented in the same infinite space is more intuitive and concise mathematically. Nevertheless, in tasks such as visualization of information, choosing any of the two interaction methods for dimension reduction is a crucial task on which the representation of the data and also the interpretation by the user will depend, therefore this research proposes the quantitative and qualitative comparison in addition to the demonstration of the previous assumption in order to contribute to machine learning tasks, visualization data, data mining where dimension reduction execute an imperative role. For example, perform tasks of classification of high dimension data, it is necessary to visualize them in such a way that they are understandable for non-expert users who want to know he topology of the data and characteristics such as separability which aid to determine which classifier could be adequate for determinate data record.

2 Methodology

Mathematically, the objective of dimension reduction is to map or project (linear transformation) data from a high-dimensional space $\boldsymbol{Y} \in \mathbb{R}^{D \times N}$ a low-dimensional space $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, where $d < D$, therefore, The original data and the embedded data will consist of N points or registers, denoted respectively by $\boldsymbol{y_i} \in \mathbb{R}^D$ and $\boldsymbol{X_i} \in \mathbb{R}^d$ with $\{\boldsymbol{K}^{(1)}, \cdots, \boldsymbol{K}^{(M)}\}$ [5], [6]. It means that the number of samples in the high-dimensional data matrix would not be affected when the number of attributes or characteristics is reduced. In order to represent the resulting embedded space in a two-dimensional Cartesian plane, this research takes into account only the two main characteristics in the kernel matrix, which represent most of the information in the original space.

2.1 Kernel based approaches

The RD method known as principal component analysis (PCA) is a linear projection that tries to preserve the variance from the values and eigenvectors of the covariance matrix [9], [10]. Moreover, when a data matrix is centered, which means that the average value of the rows (characteristics) is equal to zero, the preservation of variance could be named as a preservation of the Euclidean internal product [9].

Kernel PCA method is as similar as PCA method which maximizes the variance criterion, but in this case of a kernel matrix, which is basically an internal product of an unknown space of high dimension. We define $\phi \in \mathbb{R}^{D \times N}$ a high-dimensional space with $D_h \gg D$, which is completely unknown except for its internal product that can be estimated [9]. To use the properties of this new high-dimensional space and its internal product, it is necessary to define a function $\phi(\cdot)$ that can map the data from the original space to the high-dimension (ϕ) as follows:

$$\phi(\cdot): \mathbb{R}^D \mathbb{R}^{D_h}$$

$$\boldsymbol{y_i} \Rightarrow \phi(y_i), \tag{1}$$

where the i-th vector column of the matrix $\phi = \phi(\boldsymbol{y_i})$.

Considering the conditions of Mercer [11], and the matrix Ïł is centered, the internal product of the kernel function $\boldsymbol{K}(\cdot, \cdot)$ can be calculated as follows: $\phi(y_i)^T \phi(y_i) = \boldsymbol{K}(y_i, y_j)$. In short, the kernel function can be understood as a composition of the mapping generated by $\phi(\cdot)$ and its scalar product as follows: $\phi(y_i)^T \phi(y_i)$, so for each pair of elements of the set Y its scalar product is directly assigned without going through the mapping (ϕ). Organizing all possible internal products in a $\boldsymbol{K}_{N \times N}$ array will result in a kernel matrix:

$$\boldsymbol{K}_{N \times N} = \varphi^T{}_{D_h \times N} \varphi_{D_h \times N}. \tag{2}$$

The advantage of working with the high-dimensional space (ϕ) is that it can greatly improve the representation and visualization of the embedded data from the original space mapped to the high-dimensional space, from the calculation of the eigenvalues and eigenvectors of its product internal. An estimation of the internal product (kernel) can be designed based on the function and application that the user wants to develop [12], in this case the kernel matrices will represent distance functions associated

with a dimension reduction method, approximations kernels presented below are widely explained in [13]. The kernel representation for the CMDS reduction method is defined as the distance matrix $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ doubly centered, that is, making the mean of the rows and columns zero, as follows:

$$\boldsymbol{K}_{CMDS} = -\frac{1}{2}(\boldsymbol{I}_N - \mathbf{1}_N \mathbf{1}_N^\top)\boldsymbol{D}(\boldsymbol{I}_N - \mathbf{1}_N \mathbf{1}_N^\top), \tag{3}$$

where the ij entry of $\boldsymbol{D}$ is given by the Euclidean distance:

$$d_{ij} = ||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2. \tag{4}$$

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 $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{W})(\boldsymbol{I}_N - \boldsymbol{W}^\top)$ and λ_{max} as the largest eigenvalue of $\boldsymbol{M}$. Kernel matrix for LLE is in the form:

$$\boldsymbol{K}_{LLK} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \tag{5}$$

Considering that kernel PCA is a maximization problem in the high-dimensional covariance represented by a kernel, LE can be represented as the pseudo-inverse matrix of the graph $\boldsymbol{L}$, as shown in the following expression:

$$\boldsymbol{K}_{LK} = \boldsymbol{L}^\dagger, \tag{6}$$

where $\boldsymbol{L} = \boldsymbol{\mathcal{D}} - \boldsymbol{S}$, $\boldsymbol{S}$, such that $\boldsymbol{S}$ is a dissimilarity matrix and $\boldsymbol{\mathcal{D}} = \text{Diag}(\boldsymbol{S}\mathbf{1}_N)$ is the degree matrix is the matrix of the degree of $\boldsymbol{S}$. The similarity matrix $\boldsymbol{S}$ is organized in such a way that the relative width parameter is estimated by maintaining the entropy of the distribution with the nearest neighbor with approximately $\log K$, where K is the given number of neighbors as explained in [14]. For this investigation the number of neighbors was established as the integer closest to 10% of the amount of data.

Finally, to project the data matrix $\boldsymbol{Y} \in \mathbb{R}^{D \times N}$ into an embedded space $\boldsymbol{X} \in \mathbb{R}^{d \times N}$ we use the PCA dimension reduction method. In PCA, the embedded space is obtained by selecting the most representative eigenvectors of the covariance matrix [6], [10]. Therefore, we obtain the d most representative eigenvectors of the kernel matrix $\boldsymbol{K}_{N \times N}$ obtained previously, constructing the embedded space $\boldsymbol{X}$. As it was said for this research, the embedded space with two dimensions that represents most of the characteristics of the data is established.

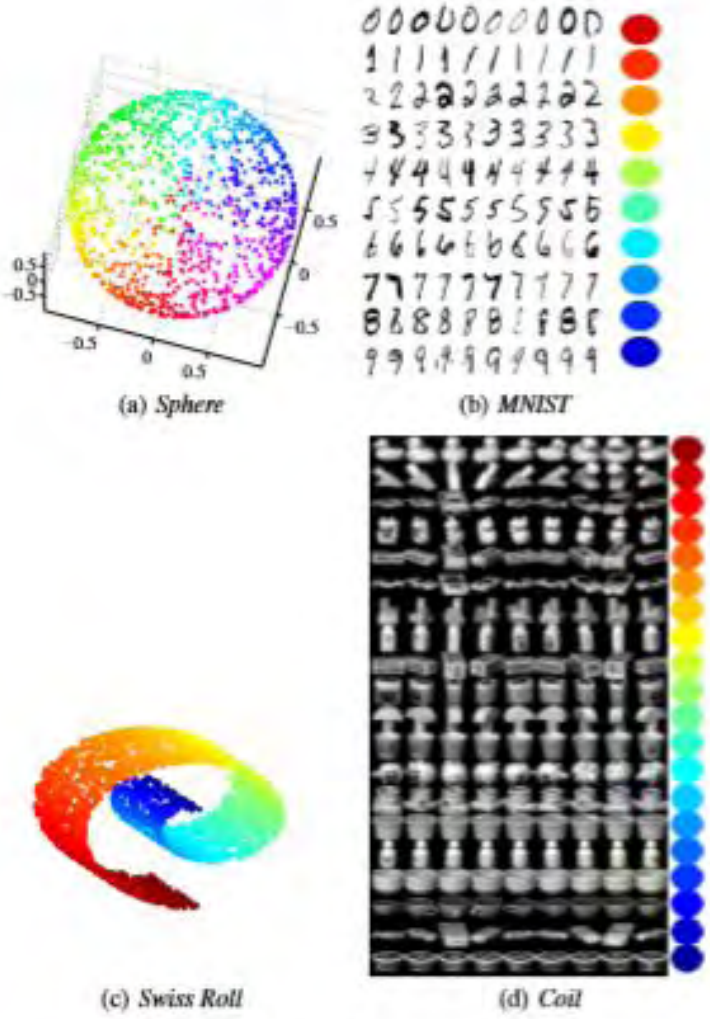
2.2 DR-Methods Mixturing

In terms of data visualization through RD methods, the parameters to be combined are the kernel matrices and the embedded spaces obtained in each method, each matrix corresponds to each of the M RD methods considered, that is $\{\boldsymbol{K}^{(1)}, \cdots, \boldsymbol{K}^{(M)}\}$. Consequently, a matrix is obtained depending on the kernel approach or final embedded space $\bar{\boldsymbol{K}}$ resulting from the mixing of the M matrices, such that:

$$\bar{\boldsymbol{K}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{K}^{(m)}, \tag{7}$$

Defining α_m as the weighting factor corresponding to the method M and $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_m\}$ as the weighting vector. In this research these parameters will be defined as 0.333 for each of the three methods used, so the sum of the three will be 1 in order to provide to each method equal priority, since the aim of this research is to present a comparison of each proposed approach in a equal conditions scenario, Each $\boldsymbol{K}(M)$ will represent the kernel matrices obtained after applying the approximations presented in equations (3), (5) and (6) or the embedded spaces obtained by applying the RD methods in their classical algorithm.

3 Results



(a) *Sphere* (b) *MNIST*

(c) *Swiss Roll* (d) *Coil*

Fig. 1. The fourth considered datasets, source: https://archive.ics.uci.edu/ml/datasets.html..

Data-sets: Experiments are carried out over four conventional data sets. The first data set (Fig. 1(a)) is an artificial spherical shell($N = 1500$ data points and $D = 3$). The second data set (Fig. ??) is a toy set here called Swiss roll ($N = 3000$ data points and $D = 3$). The third data set (Fig. 1(d)) is `Coil_20` is a database of gray-scale images of 20 objects. Images of the objects were taken at pose intervals of 5 degrees. This corresponds to 72 images per object ($N = 1440$ data points 20 and $D = 1282$ -number of pixels) [15]. The fourth data set (Fig. ??) is a randomly selected subset of the MNIST image bank [11], 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 $D = 242$). Figure 1 depicts examples of the considered data sets.

Performance Measure: In dimensionality reduction, the most significant aspect, which defines why a RD method is more efficiency, is the capability of preserve the data topology in low-dimensional space regarding the high-dimension. Therefore, we apply a quality criterion used by conserving the k-th closest neighbors developed in [16], as efficacy measure for each approach proposed for the interactive RD methods mixture. This criterion is widely accepted as an adequate unsupervised measure [17], [14], which allows the embedded space to assess in the following way: The rank of ε_j with respect to ε_i in high-dimensional space is denoted as:

$$p_{ij} = |\{k : \delta_{ik} < \delta_{ij} \ \ or \ \ (\delta_{ik} = \delta_{ij} \ \ and \ \ 1 \leq k < j \leq N)\}|. \tag{8}$$

In equation (8) $|\cdot|$ denotes the set cardinality. Similarly, in [13] is defined that the range of xj with respect to xi in the low-dimensional space is:

$$r_{ij} = |\{k : d_{ik} < d_{ij} \ \ or \ \ (d_{ik} = d_{ij} \ \ and \ \ 1 \leq k < j \leq N)\}|. \tag{9}$$

The k-th neighbors of ζi and xi are the sets defined by (10) and (11), respectively.

$$v_i^k = \{j : 1 \geq pij < K\}, \tag{10}$$

$$n_i^k = \{j : 1 \geq rij < K\}. \tag{11}$$

A first performance index can be defined as:

$$Q_{NX}(K) = \sum_{i=1}^{N} \frac{|v_i^k \cap n_i^k|}{KN} - 1. \tag{12}$$

Equation (12) results in values between 0 and 1 and measures the normalized average according to the corresponding k-th neighbors between the high-dimensional and low-dimensional spaces. Defining in this way a coclasification matrix:

$$[Q = q_{NX}] \ for \ j \geq N - 1, \tag{13}$$

whit $q_{kl} = |\{(i, j) : p_{ij} = k \ \ and \ \ p_{ij} = l\}|$.

Therefore $Q_{NX}(K)$ counts k-by-k blocks of Q, the range preserved (in the main diagonal) and the permutations within the neighbors (on each side of the diagonal) [12]. This research employs an adjustment of the curve $Q_{NX}(K)$ introduced in [12] in order

that the area under the curve is an adequate indicator of the embedded data topology preservation, hence, the quality curve that is applied into the visualization methodology is given by:

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - N}{N-1-K}. \tag{14}$$

When the equation in (14) is expressed logarithmically, errors in large neighborhoods are not proportionally as significant as small ones [14]. This logarithmic expression allows obtaining the area under the curve of $R_{NX}(K)$ given by:

$$AUC \log_K(R_{NX}(K)) = \frac{\sum_{K=1}^{N-2} \frac{R_{NX}(K)}{K}}{\sum_{K=1}^{N-2} \frac{1}{K}}. \tag{15}$$

The results obtained by applying the methodology proposed over four data bases described, are shown in Fig.2, where the curve $R_{NX}(K)$ of each approach is presented as well as the AUC in (13) which assess the dimension reduction quality corresponding to each proposed combination. As a result, for RD procedure in terms of visualization we show the embedded space for each test performed. It is necessary to clarify that each combination was carried out same scenario with equal conditions which allows us to measure a computational cost in terms of execution time, which are shown in Table 1. This is an important issue if users are seeking for an interactive RD methods mixture which has a satisfactory performance, as well as an efficient computational development.

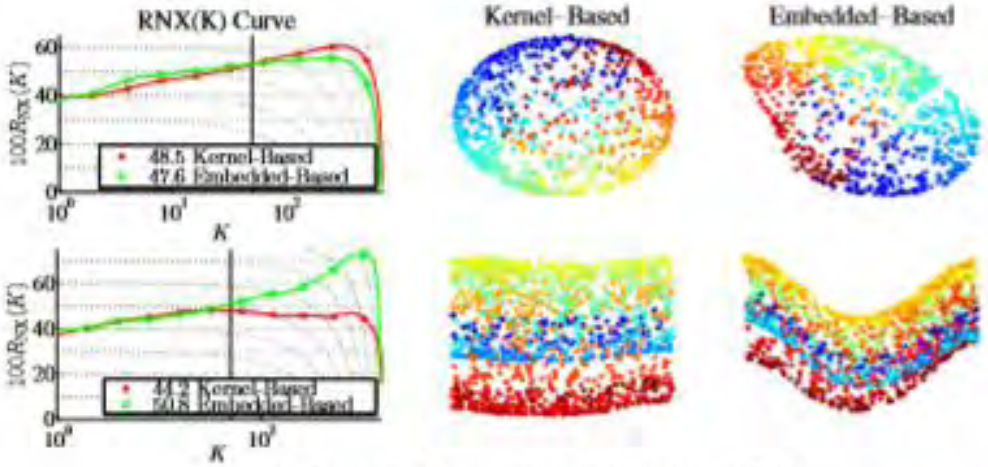
Based Approach	Dataset	Computacional time (sec)
Kernel	3D sphere	6,27
	Swiss Roll	6,43
	Coil-20	28,94
	MINST	37,87
Embedded-spaces	3D sphere	2,88
	Swiss Roll	3,09
	Coil-20	15,24
	MINST	16,24

Table 1. Consumed time for performing each approach over the fourth dataset.

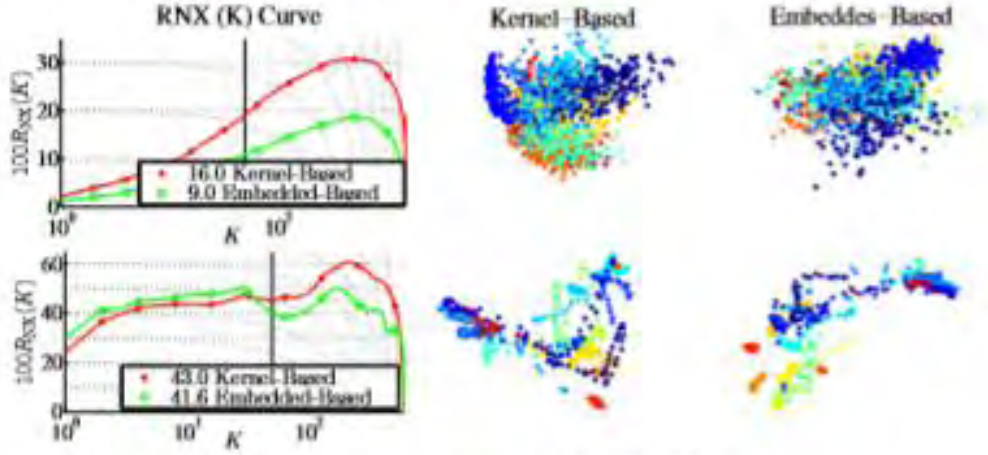
Nevertheless, results achieved in this research allows us to conclude that in data visualization terms performing an interactive mixture RD method based on kernel is more favorable than based on standard methods, mathematically combining a kernel approximations, which means that each kernel approximation is in the same high-dimensional space where all classes are separable before developing the mixture, is more appropriate than combining obtained embedded space from an unknown space which are the standard methods.

The computational cost (Table 1) allows us to infer that the cost in executing kernel approaches and PCA kernel application for dimension reduction is a slightly more elevated in all cases. This is since the databases have a high number of registers, which

means that acquiring the kernel matrices involves a lot of processing, as if the data base consists of n samples, the kernel matrix size will be $N \times N$.



(a) *Results for datasets: Sphere 3D and Swiss Roll.*



(b) *Results for datasets: MNIST and Coil-20.*

Fig. 2. Results obtained for the four experimental databases

Making a comparison of the $R_{NX}(K)$curves for each database, there is a low performance in the dimension reduction process for the case of the Coil-20 database whose AUC is the lowest among all, which means that the data topology in the embedded space obtained is not as conserved as in the other studied cases. Evidently the best performance was accomplished for 3D spherical shell and Swiss roll which obtained the best AUC and preserve the data local structure, generally preserved local structure generates superior embedded spaces [13]. On the other hand, MNIST and spherical shell

database preserved the global data structure in a preferable way as regards the other cases.

4 Conclusion

This work presented a comparative analysis of two different approaches for DRmethods mixturing which are applied in an interactive. Results obtained in this research allows us to conclude that performing an interactive DR-methods mixture could be a tough task for a dataset with a great number of points and dimensions as it was proved that the computational cost is higher but also this approach gives to users a high-quality performance since, a greater area is obtained under the quality curve which indicates that the topology of the data can be preserved more. On the other hand, embedded-spaces-based approach has a slightly difference in the $R_{NX}(K)$ AUC curve, but it is not wide so if the user wants to carry out a quicker mixture, the embedded-spaces-based approach will be more appropriate for data visualization where interactivity is the most important achievement seeking a better perception for the inexpert users of their datasets.

Acknowledgments

This work is supported by the Smart Data Analysis Systems (SDAS) Research Group (http://sdas-group.com), as well as the *"Grupo de Investigación en Ingeniería Eléctrica y Electrónica - GIIEE"* from Universidad de Nariño. Also, the authors acknowledge to the research project: *"Desarrollo de una metodología de visualización interactiva y eficaz de información en Big Data"* supported by Agreement No. 18 November 1st, 2016 by VIPRI from Universidad de Nariño.

References

1. Sacha, D., Zhang, L., Sedlmair, M., Lee, J.A., Peltonen, J., Weiskopf, D., North, S.C., Keim, D.A.: Visual interaction with dimensionality reduction: A structured literature analysis. IEEE transactions on visualization and computer graphics **23**(1) (2017) 241–250
2. Peluffo Ordoñez, D.H., Lee, J.A., Verleysen, M.: Recent methods for dimensionality reduction: A brief comparative analysis. In: 2014 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014). (2014)
3. Peluffo-Ordóñez, D.H., Castro-Ospina, A.E., Alvarado-Pérez, J.C., Revelo-Fuelagán, E.J.: Multiple kernel learning for spectral dimensionality reduction. In: Iberoamerican Congress on Pattern Recognition, Springer (2015) 626–634
4. Belanche Muñoz, L.A.: Developments in kernel design. In: ESANN 2013 proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: Bruges (Belgium), 24-26 April 2013. (2013) 369–378
5. Borg, I., Groenen, P.J.: Modern multidimensional scaling: Theory and applications. Springer Science & Business Media (2005)
6. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing **72**(7-9) (2009) 1431–1443

7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15**(6) (2003) 1373–1396
8. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. science **290**(5500) (2000) 2323–2326
9. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, IEEE (2014) 171–177
10. Gijón Gómez, J.: Visualización bidimensional de problemas de clasificación en alta dimensión. B.S. thesis (2013)
11. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the twenty-first international conference on Machine learning, ACM (2004) 47
12. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge university press (2000)
13. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing **112** (2013) 92–108
14. Cook, J., Sutskever, I., Mnih, A., Hinton, G.: Visualizing similarity data with a mixture of maps. In: Artificial Intelligence and Statistics. (2007) 67–74
15. Nene, S.A., Nayar, S.K., Murase, H., et al.: Columbia object image library (coil-20). (1996)
16. Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. Journal of the American Statistical Association **104**(485) (2009) 209–219
17. France, S., Carroll, D.: Development of an agreement metric based upon the rand index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data. In: International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer (2007) 499–517

ANEXO 4. ARTICULO: IWAIPR'18

The Angle-based model for interactive dimensionality reduction and data visualization

C. K. Basante-Villota, C. M. Ortega-Castillo, D. F. Peña-Unigarro, J. E. Revelo-Fuelagán[1], J. A. Salazar-Castro[2,3], and D. H. Peluffo-Ordóñez[3,4]

[1] Universidad de Nariño, Colombia.
[2] Universidad Nacional, sede Manizales, Colombia.
[3] Corporación Universitaria Autónoma de Nariño, Colombia.
[4] Yachay Tech - Ecuador

Abstract. In recent times, an undeniable fact is that the amount of data available has increased dramatically due mainly to the advance of new technologies allowing for storage and communication of enormous volumes of information. In consequence, there is an important need for finding the relevant information within the raw data through the application of novel data visualization techniques that permit the correct manipulation of data. This issue has motivated the development of graphic forms for visually representing and analyzing high-dimensional data. Particularly, in this work, we propose a graphical approach, which, allows the combination of dimensionality reduction (DR) methods using an angle-based model, making the data visualization more intelligible. Such approach is designed for a readily use, so that the input parameters are interactively given by the user within a user-friendly environment. The proposed approach enables users (even those being non-experts) to intuitively select a particular DR method or perform a mixture of methods. The experimental results prove that the interactive manipulation enabled by the here-proposed model -due to its ability of displaying a variety of embedded spaces-makes the taks of selecting a embedded space simpler and more adequately fitted for a specific need.

Keywords: Dimensionality reduction, data visualization, kernel PCA, pairwise similarity.

1 Introduction

Given the existence of new sources of information (sensors, mobile phone, emails, social networks, the internet in general, and among others), the emerging term so-named Big Data has taken place, which is a relatively new concept attained to encompass big volumes of data coming from several sources as well as technologies able to deal with such data. The main issues that Big Data concerns are: volume, visualization, variability and speed. Therefore, this research field has increasingly become an area of great interest in computer science and data analytics. This research is focused on the concept of data visualization through dimensionality reduction techniques. The mapping of high-dimensional data into a smaller version that depicts the most relevant information from the original data is a widely studied research area [9, 14], given its ability to reduce the

computational cost or improve the performance of pattern recognition and information visualization systems [10, 11]. Despite the existence of tools that achieve the efficient indicators in terms of computational performance, exploration and representation of high-dimensional data, they do not take into account properties such as interactivity and controllability. Therefore, an improvement in these aspects is required [10, 13]. In consequence, there is a gap between the knowledge of the users and the database to be analyzed [14]. In this connection, an interaction model is proposed that improves the manipulation of the interface and the knowledge discovery inside of the data-set. Users can obtain an overview of the data to draw conclusions and make decisions [14]. This interactive model is based on the geometry of a triangle, using mainly the theorem of internal and external angles of the Euclidean geometry. Each vertex of the triangle represents a DR method. Therefore, the vertex with the greatest angle will represent the highest model value as well as the maximum application of one particular DR method. Spectral DR methods are implementing through kernel approximations [8,9,13], which are combined to reach a final kernel matrix. Finally, such a kernel matrix feeds a generalized algorithm of kernel principal component analysis (KPCA) [8]. The benefit of this approach is that user may utilize DR methods over the data, even with no knowledge about the theoretical foundations behind them. The user controls the results simply by exploring an intuitive interface based on the angles of a triangle. The angle-based model proposed in this paper is evaluated using three DR methods, namely: locally linear embedding (LLE) [12], multidimensional classical scaling (CMD) [3] and laplacian eigenmaps (LE) [2]. The experiments are performed over a real databases (images of objects - MNIST) and two artificial data-set (Swiss roll and letter S in 3D) [1]. The DR performance is quantified by a scaled version of the average agreement rate between K-ary neighborhoods explained in [7]. This paper is organized as follows: Section 2 describes the Proposed angle-based model for the combination of DR methods. Experimental setup and results are presented in Sections 3 and 4, respectively. Finally, some final remarks are drawn in section 5.

2 Proposed model for interactive dimensionality reduction using a angle-based approach

This section describes the proposed model, here named, angle-based model that allows an interactive combination of three different unsupervised DR spectral methods, in order to obtain an improvement of the data visualization process. A suitable and versatile approximation for DR spectral methods are kernel matrices because they make possible a linear combination [8,9,13].

Our interactive model allows the mixing of kernel matrices in an intuitive way. First, the kernel matrices obtained by applying the DR methods in high-dimensional data are linearly combined, after the angle-based model is applied a new kernel matrix is created which contains a weighted combination of three different types of DR methods with the purpose of reach several low-dimensional spaces. Therefore, user can easily and intuitively select a both a unique DR method or a combination of three different DR methods in order to get a suitable data representation of one specific task by simply exploring the different positions that can be obtained. Figure 1 shows a diagram of the

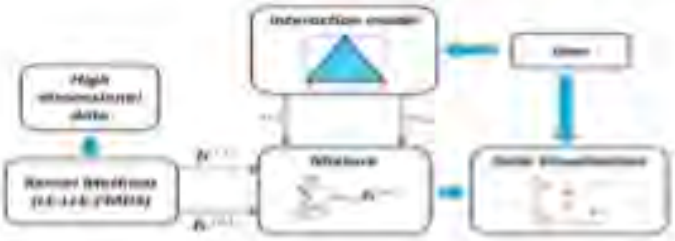


Fig. 1. Block diagram of proposed interactive data visualization using dimensionality reduction which illustrates each stage, step by step, about the interface.

interface. Since we are aiming at combining three DR methods, we propose to use a triangle, where each vertex represents a DR method. This triangle is within in a square since this geometric figure allows us to obtain angles close to 180°. The model proposed in this paper is based on the geometry of a triangle, using the external and internal angle theorem of Euclidean geometry, which say that every exterior angle of a triangle is equal to the sum of the two Non-adjacent interior angles, as shown in Figure 2(a) $\measuredangle D = \measuredangle C + \measuredangle A$, it can be intuitively demonstrated that $\measuredangle D + \measuredangle B = 180°$ and therefore $\measuredangle C + \measuredangle A + \measuredangle B = 180°$.

(a) *Illustration of the triangle geometry*

(b) *Illustration of angle-based model.*

Fig. 2. Illustration of the construction of the interaction model

The user has the freedom to move each one of the vertices that are represented by three spheres of different colors around the square, thus changing the measure of the angles and consequently the factors of each of the methods. 180° illustrates a normalized value of 1, so the percentage corresponding to each method depicts the mixture configuration of the kernel matrices. Figure 2(b) shows a graph of the model proposed.

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 $\widehat{K}$ as the mixture of M kernel matrices $\{K^{(1)}, \cdots, K^{(M)}\}$ so: $\widehat{K} = \sum_{m=1}^{M} \alpha_m K^{(m)}$, where α_m is the coefficient or weighting factor corresponding to method m and $\alpha = \{\alpha_1, \cdots, \alpha_m\}$ is weighting vector. In this work $m = 3$ and the relationship between the points within the surface and the coefficients of linear combination α_m are given by the angular measure of each vertex that makes up the triangle. Nevertheless, this interactive model differs from other approaches as [16,17] due

to the fact that angle base model has the ability to expand the number of dimensionality reduction methods which can be combined. Thus, the proposed model can be adapted into a polygonal approach, if new kernel representations are proposed, in order to increase the number of embedded spaces which can improve tasks as data visualization, pattern recognition, data mining, among others.

3 Experimental setup

Data-sets: Experiments are carried out over three conventional data sets. The first data set is a letter **S** in 3D ($N = 1000$ data points and $D = 3$). The second data set is a toy set here called Swiss roll (N = 3000 data points and D = 3). The third data set is a randomly selected subset of the MNIST image bank [6], 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 $D = 24^2$). Figure 3 depicts examples of the considered data sets.

(a) *Letter S in 3D* (b) *Swiss Roll* (c) *MNIST*

Fig. 3. The three considered datasets.

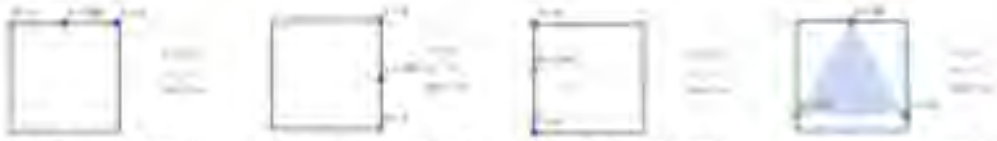
Methods: Dimensionality reduction allows the extraction of relevant information from high- dimensional data sets aimed at improving the performance of a pattern recognition system or that facilitates the visualization and analysis of data. In mathematical terms, the goal of dimensionality reduction is 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 $d < D$ [12,13]. Three spectral DR approaches are considered, namely: classical multidimensional scalling (CMDS) [3], locally linear embedding (LLE) [12], and graph Laplacian eigenmaps (LE) [2]. They are all performed in their standard algorithms. Also, in order to evaluate our framework, kernel approximations are also considered. CMDS kernel is the double centered distance matrix $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ so $\boldsymbol{K}^{(1)} = \boldsymbol{K}_{CMDS} = -\frac{1}{2}(\boldsymbol{I} - \boldsymbol{1}_N\boldsymbol{1}_N^\top)\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{1}_N\boldsymbol{1}_N^\top)$, where the ij entry of $\boldsymbol{D}$ is given by $d_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2$.

A kernel for LLE can be approximated from a quadratic form in terms of the matrix $\boldsymbol{\mathcal{W}}$ holding linear coefficients that sum to 1 and optimally reconstruct observed data. Define a matrix $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{\mathcal{W}})(\boldsymbol{I}_N - \boldsymbol{\mathcal{W}}^\top)$ and λ_{max} as the largest eigenvalue of $\boldsymbol{M}$. Kernel matrix for LLE is in the form $\boldsymbol{K}^{(2)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}$. Since kernel PCA is a maximization of the high-dimensional covariance

represented by a kernel, a feasible kernel for LE can be represented as the pseudo-inverse of the graph Laplacian L: $K^{(3)} = K_{LE} = L^{\dagger}$, where $L = D - S$, S is a similarity matrix and $D = \mathrm{Diag}(S\mathbf{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [5]. The similarity matrices are 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 [4]. For all methods, input data is embedded into a 2-dimensional space, then $d = 2$. The number of neighbors is established as $K = 30$ for all considered data sets.

Quality measures: To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [7] 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.

Experiment description: To assess the performance of the interactive visualization interface, a testing were done by moving the vertices of the triangle through the square. Doing so a collection of weighting factors are established to consequently carry out the mixture. In Figure 4 the weighting factors configuration for the experiment are defined



(a) *LE method is independent* (b) *CMDS method is independent* (c) *LLE method is independent* (d) *The three methods are mixed*

Fig. 4. The chosen positions for the experiment.

4 Results and discussion

In the experiment we considered three DR methods that are mixed through the angle-based model. We tested the interactive model in four positions as shown in Figure 4. Three of the positions represent one single DR method and in the last configuration the three methods are combined with the same percentage of participation. The results are shown in Figures 5 to 7. In the results we can see the embedded data and several curves that give a notion to the user about the performance of the low-dimensional space and the preservation of the neighbors in the high dimensional-space. If the value of the area under the curve is greater, the performance of the integrated data will be better. Therefore, we affirm that the sphere Figure 5 obtains embedded data with greater preservation of relevant information. In the Figures 6 and 7 we observe that some DR methods have better performance in one or another data set depending on their nature.

To facilitate the management of our interactive model, we implemented an interface in the NetBeans software, which allows the calculation of the DR methods and the

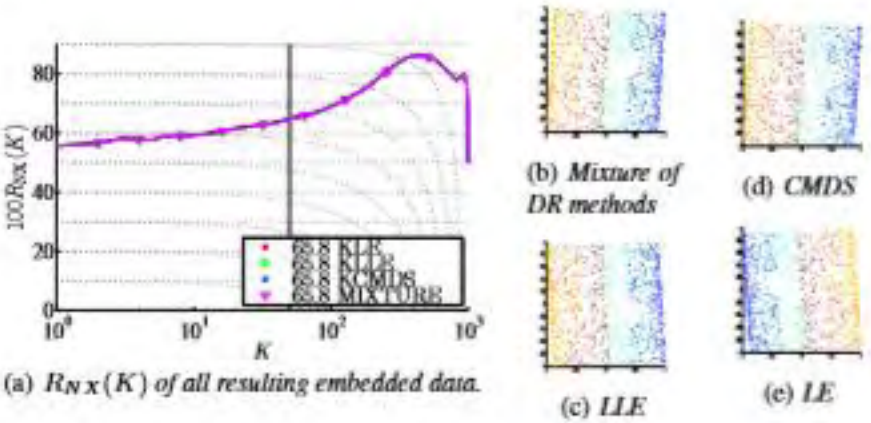


Fig. 5. Results for the letter-S data-set.

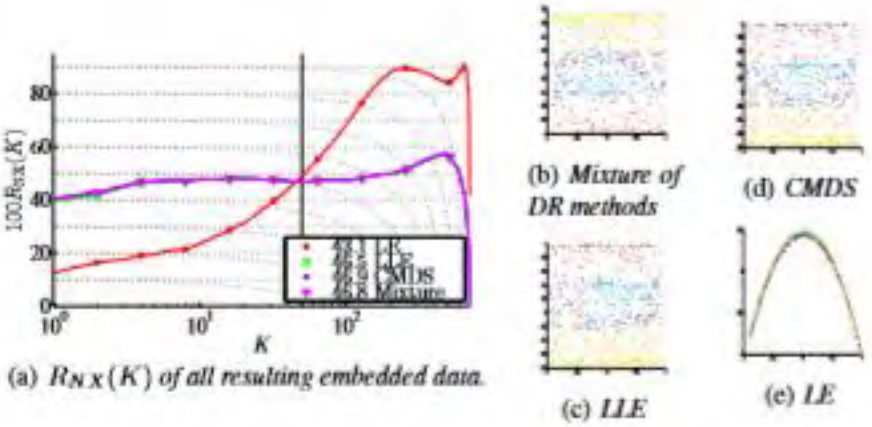


Fig. 6. Results for the Swiss roll data-set.

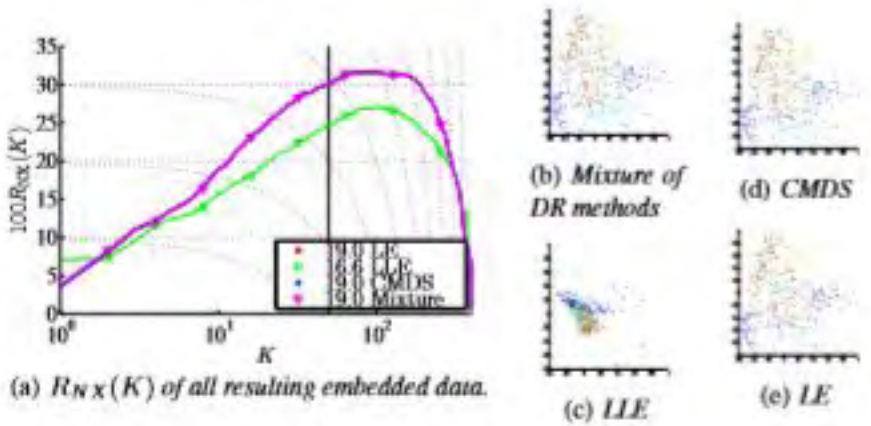


Fig. 7. Results for MNIST dataset.

visualization of their results through scatter plots. in order to create an attractive visual analysis environment. Figure 8 shows a view of the implemented interface. Therefore, a powerful tool is provided to make decisions about the most appropriate representation of the original data, as well as, the combination of the most appropriate DR methods, with the purpose of to users (even non-expert) can intuitively interact with the DR methods and its feasible combinations.

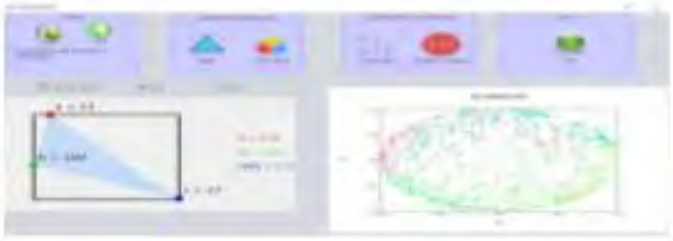


Fig. 8. View of the interface implemented on NetBeans software.

5 Conclusions and future work

The angle-based interaction model successfully generates the mixing coefficients, since it allows the three DR methods to be combined satisfactorily in any position that is required. In addition, the model allows the use of independently implemented DR methods. It is important to mention that the interaction model can be used in other types of mixtures that are based on linear combinations. The development of the interactive interface facilitates the use of the interaction model for users without experience in DR methods, due to all the results are shown graphically, allowing a more intuitive understanding of the data. The angles of the triangle turn out to be easy to understand and therefore attractive for interaction with the interface. In this sense, the user might fulfill their specific needs and parameter criteria by moving the vertices of the triangle. As future work, other DR methods could be included for mixing, since our approach has a suitable adaptation capability for the mixture of more than three kernel matrices in order to improve the results of the dimensionality reduction. Furthermore, several developed and interactive models can be explored to optimize and accelerate the interface and its performance.

Acknowledgments

This work is supported by the Smart Data Analysis Systems (SDAS) Research Group (http://sdas-group.com), as well as the "Grupo de Investigación en Ingeniería Eléctrica y Electrónica - GIIEE" from Universidad de Nariño. Also, the authors acknowledge to the research project: "Desarrollo de una metodología de visualización interactiva y eficaz de información en Big Data" supported by Agreement No. 18 November 1st, 2016 by VIPRI from Universidad de Nariño.

References

1. A. Asuncion and D. Newman. Uci machine learning repository. irvine, ca: University of california, school of information and computer science. *Available online a t http://www. ics. uci. edu/ mlearn/MLRepository. html*, 2007.
2. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.
3. I. Borg. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.
4. 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*, pages 67–74, 2007.
5. J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the twenty-first international conference on Machine learning*, page 47. ACM, 2004.
6. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
7. 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.
8. D. H. Peluffo-Ordonez, J. Aldo 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*, pages 171–177. IEEE, 2014.
9. D. H. Peluffo-Ordóñez, A. E. Castro-Ospina, J. C. Alvarado-Pérez, and E. J. Revelo-Fuelagán. Multiple kernel learning for spectral dimensionality reduction. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 626–634. Springer, 2015.
10. D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen. Recent methods for dimensionality reduction: A brief comparative analysis. In *European Symposium on Artificial Neural Networks (ESANN)*. Citeseer, 2014.
11. D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen. Short review of dimensionality reduction methods based on stochastic neighbour embedding. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 65–74. Springer, 2014.
12. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
13. J. Salazar-Castro, Y. Rosas-Narvaez, A. Pantoja, J. C. Alvarado-Perez, and D. H. Peluffo-Ordonez. Interactive interface for efficient data visualization via a geometric approach. In *Signal Processing, Images and Computer Vision (STSIVA), 2015 20th Symposium on*, pages 1–6. IEEE, 2015.
14. M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. In *Computer Graphics Forum*, volume 34, pages 201–210. Wiley Online Library, 2015.
15. M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner. Dimensionality reduction in the wild: Gaps and guidance. *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03*, 2012.
16. D. F. Peña-Unigarro, P. Rosero-Montalvo, E. J. Revelo-Fuelagán, J. A. Castro-Silva, J. C. Alvarado-Pérez, R. Therón, C. M. Ortega-Bustamante and D. H. Peluffo-Ordóñez. Interactive Data Visualization Using Dimensionality Reduction and Dissimilarity-Based Representations. *Intelligent Data Engineering and Automated Learning - IDEAL 2017*, Lecture Notes in Computer Science, vol 10585. Springer, Cham
17. P. D. Rosero-Montalvo, D. F. Peña-Unigarro, D. H. Peluffo, J. A. Castro-Silva, A. Umaquinga and E. A. Rosero-Rosero. Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction *Biomedical Applications Based on Natural and Artificial Computing*, vol. 10338, pp. 289, 2017.

ANEXO 5. ARTICULO: IDEAL'18

Generalized low-computational cost Laplacian eigenmaps

J. A. Salazar-Castro [1,2], D. F. Peña, C. Basante[3], C. Ortega[3], L. Cruz-Cruz [2], J. Revelo-Fuelagán [3], G. Castellanos-Domínguez[1], and D. H. Peluffo-Ordóñez [4*]

[1] Universidad Nacional, sede Manizales, Colombia.
[2] Corporación Universitaria Autónoma de Nariño, Colombia.
[3] Universidad de Nariño, Colombia.
[4] Yachay Tech - Ecuador,

Abstract. Dimensionality reduction (DR) is a methodology used in many fields linked to data processing, and may represent a preprocessing stage or be an essential element for the representation and classification of data. The main objective of DR is to obtain a new representation of the original data in a space of smaller dimension, such that more refined information is produced, as well as the time of the subsequent processing is decreased and/or visual representations more intelligible for human beings are generated. The spectral DR methods involve the calculation of an eigenvalue and eigenvector decomposition, which is usually high-computational-cost demanding, and, therefore, the task of obtaining a more dynamic and interactive user-machine integration is difficult. Therefore, for the design of an interactive IV system based on DR spectral methods, it is necessary to propose a strategy to reduce the computational cost required in the calculation of eigenvectors and eigenvalues. For this purpose, it is proposed to use locally linear submatrices and spectral embedding. This allows integrating natural intelligence with computational intelligence for the representation of data interactively, dynamically and at low computational cost. Additionally, an interactive model is proposed that allows the user to dynamically visualize the data through a weighted mixture.

Keywords: Dimensionality reduction, generalized kernel, kernel PCA, multiple kernel learning.

1 Introduction

The aim of dimensionality reduction (DR) is to extract a lower dimensional, relevant information from high-dimensional data, being then a key stage within the design of pattern recognition and data mining systems. Indeed, when using adequate DR stages, the system performance can be enhanced as well as the data visualization can become more intelligible. The range of DR methods is diverse, including those classical approaches such as principal component analysis (PCA) and classical multidimensional scaling (CMDS), which are respectively based on variance and distance preservation

* This work is supported by Yachay Tech and SDAS research group www.sdas-group.com.

criteria [1]. Recent methods of DR are focused on the data topology preservation. Mostly such a topology is driven by graph-based approaches where data are represented by a non-directed and weighted graph. In this connection, the weights of edge graphs are certain pairwise similarities between data points, the nodes are data points, and a non-negative similarity (also affinity) matrix holds the pairwise edge weights. Spectral methods such as Laplacian eigenmaps (LE) [2] and locally linear embedding (LLE) [3] were the pioneer ones to incorporate similarity-based formulations. Also, given the fact that the rows of the normalized similarity matrix can be seen as probability distributions, divergence-based methods have emerged (i.e., stochastic neighbor embedding (SNE) [**?**]). Spectral approaches for DR have been widely used in several applications such as relevance analysis [**?**, **?**], dynamic data analysis [**?**, **?**] and feature extraction [**?**, **?**]. Because of being graph-driven methods and involving then similarities, spectral approaches can be easily represented by kernels [4], which means that a wide range of methods can be set within a Kernel PCA framework [5]. At the moment to choose a method, aspects such as nature of data, complexity, aim to be reached and problem to be solved should be taken into consideration. In this regard, as mentioned above, there exists a variety of DR spectral methods making the selection of a method a nontrivial task. Also, some problems may require the combination of methods so that the properties of different methods are simultaneously taken into account to perform the DR process and the quality of resultant embedded space is improved.

In this work, we explore the possibility to extend LE to ...

The experiments are carried out over well-known data sets, namely an artificial `Spherical shell`, a `Swiss roll` toy set, and `MNIST` image bank [6]. The DR performance is quantified by a scaled version of the average agreement rate between K-ary neighborhoods as described in [7].

The rest of this paper is organized as follows: Section ...

2 LE generalizado

Uno de los métodos RD espectrales del tipo no lineal es LE, que es un método basado en grafos. Este método se abordó ampliamente en la subsección **??** pero, en resumen, LE genera una representación de un espacio de alta dimensión en un espacio de menor dimensión, procurando conservar al máximo las relaciones de proximidad entre puntos cercanos. Para tal fin, LE realiza un mapeo de patrones de entrada cercanos en valores de salida cercanos [8]. La entrada para LE es una matriz $\boldsymbol{W}$ simétrica positiva (semi)definida de tamaño $N \times N$ que contiene la información del espacio original, pero, como vimos en la subsección **??**, este tipo de matrices puede ser reemplazada por una aproximación *kernel*. Por tanto, introducimos una metodología de RD espectral, al cual denominamos con las siglas LEK. Dicha modificación consiste en determinar el laplaciano $\boldsymbol{L}$ que utiliza LE pero en está metodología se cambia la matriz de similitud W por alguna matriz *kernel* $\boldsymbol{K}$, como la de los métodos espectrales de la subsección **??**, y modificando la matriz de grado $\boldsymbol{D} = Diag(\boldsymbol{W}\mathbf{1}_N)$ por $\boldsymbol{D} = Diag(\boldsymbol{K}\mathbf{1}_N)$, por tanto el laplaciano modificado sera:

One nonlinear dimensionality reduction method is known as Laplacian eigenmaps (LE) which is based on graphs. This method was widely explained in subsection **??**.

To sum up, LE generates a low-dimensionality representation from a high-dimensional data in order to preserve the highest relations of proximity between the nearest points. LE works through a mapping of close entry patterns on close output values [8]. The input for LE is a symmetric positive semidefinite matrix $\boldsymbol{W}$ of size $N \times N$ which contains the information of the original space. Nonetheless, as it was illustrated subsection **??** this kind of matrices are susceptible to be represented with kernel matrices. Hence, a new spectral methodology defined by the letters LEK is introduced. The modification made in the classical approach consist in finding the same laplacian $\boldsymbol{L}$ that is used in LE but changing the similarity matrix W by a kernel matrix $\boldsymbol{K}$ how it was depicted in spectral methods subsection **??**, then a modification of the grade matrix $\boldsymbol{D} = Diag(\boldsymbol{W}\mathbf{1}_N)$ by $\boldsymbol{D} = Diag(\boldsymbol{K}\mathbf{1}_N)$ is made, in consequence, the new Laplacian will be:

$$\boldsymbol{L} = \boldsymbol{K} - \boldsymbol{D}. \tag{1}$$

El planteamiento de la función objetivo del problema de minimización de LE (ver ecuación **??**) no se verá afectado y dará la misma solución vista en la subsección **??** ($\lambda \boldsymbol{D}\boldsymbol{f} = \boldsymbol{L}\boldsymbol{f}$). La figura 1 presenta el diagrama general de la metodología propuesta.

The objective function definition of the minimization problem of LE (equation **??**) is not affected and the solution will be the same demonstrated in subsection **??** ($\lambda \boldsymbol{D}\boldsymbol{f} = \boldsymbol{L}\boldsymbol{f}$). The figure 1 illustrates the general diagram of the proposed methodology.

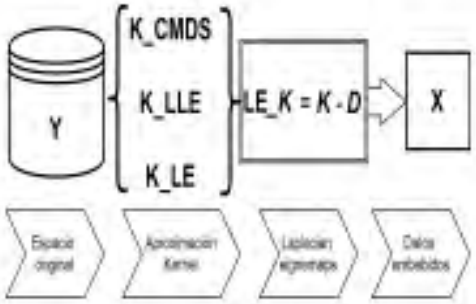


Fig. 1. Metodología para la generalización de RD espectral a partir de LE

3 Reducción del coste computacional a través de submatrices localmente lineales - LLKL

En [9] se introduce una solución aproximada al problema del costo computacional mediante la cual se reduce el tiempo en la estimación de los vectores y valores propios que realizan los métodos espectrales. El método que proponen se denomina submatrices localmente lineales (LLL). Dicho método consiste en seleccionar un número de submatrices $\widetilde{\boldsymbol{Y}}_{D\times L}$ tal que $L \ll N$, aproximando linealmente el espacio de entrada de forma local al rededor de las submatrices pero de forma no lineal desde la representación global. Ahora, considere a $\boldsymbol{Z}$ como una matriz de proyección que realiza un mapeo localmente lineal, como lo presenta la ecuación 2.

In [9] an approximate solution for the computational cost problem is introduced in order to reduce the time estimation of eigen-values and eigen-vectors which are required by the spectral methods.The proposed method is called locally linear submatrices where the selection of a number of sub matrices $\bar{\boldsymbol{Y}}_{D\times L}$ with $L \ll N$ is done . This process does a linear approximation of the input space in a local way around the sub matrices in a nonlinear form from the global representation. Now $\boldsymbol{Z}$ is defined as a projection matrix which made a locally linear mapping illustrated in the equation 2.

$$\boldsymbol{Y} \approx \bar{\boldsymbol{Y}}\boldsymbol{Z}. \tag{2}$$

De esta forma, se puede reducir el costo computacional de la descomposición en vectores y valores propios sin una perdida significativa de información puesto que para la construcción de la matriz de afinidad se considera a todo el espacio de entrada. LLL es una mejora para los métodos RD, una solución al problema espectral, más no un nuevo método RD. Además, dado que los métodos RD se diseñan con criterios pre-establecidos, para el caso de LE por ejemplo, es necesario fijar parámetros como el ancho de banda σ de las afinidades gaussianas y, en general para la mayoría de métodos RD, el número de vecinos K_W como un nivel de dispersión.
Dado que LLL asume que existe una dependencia local entre los puntos de las submatrices del espacio de entrada 2 y del espacio embebido, entonces se puede utilizar la matriz de proyección $\boldsymbol{Z}$ nuevamente en un mapeo lineal en baja dimensión, así:

In this sense, the computational cost can be reduce by an alternative decomposition of eigen-values and eigen-vectors without a significan lost of information since the entire input space is considered for the construction of the similarity matrix. LLL is an improvement for the DR methods because it represents a solution for the spectral problem in spite of proposing new RD method. Moreover, due to the fact that DR methods are designed with pre-established criteria. For instance, for LE is necessary set up parameters as bandwidth σ of Gaussian affinities also in general the majority of DR methods take into account the number of neighbors K_W as a level of dispersion.

Since LLL assumed the existence of one local dependence between the sub-matrices points in the input space 2 and in the embedded space, the projection matrix $\boldsymbol{Z}$ can be used again in a linear mapping in low dimension as follows:

$$\boldsymbol{X} \approx \widetilde{\boldsymbol{X}}\boldsymbol{Z}. \tag{3}$$

Por lo tanto, se puede redefinir el problema espectral existente de forma aproximada, en relación a parámetros d y L, de la siguiente manera:

Thus, it is possible to define the current spectral problem through an approximation by the relation parameter d and L as it is described below:

$$\begin{aligned} \min_{X} \quad & tr(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{X}}^T) \\ \text{s. t.} \quad & \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{X}}^T = \boldsymbol{I}, \end{aligned} \tag{4}$$

where $\widetilde{\boldsymbol{A}} = \boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T$, $\widetilde{\boldsymbol{B}} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^T \in \mathbb{R}^{L\times L}$.

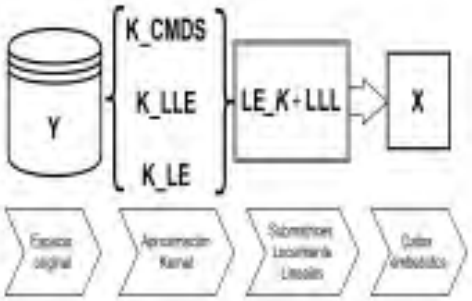


Fig. 2. Metodología de reducción del costo computacional basado en submatrices

Finalmente, dado que la solución es aplicada a LE, la matriz de afinidad $\boldsymbol{A}$ es reemplazada por el laplaciano $\boldsymbol{L}$ que contiene una relación de la matriz de afinidad $\boldsymbol{W}$ de LE y la matriz $\boldsymbol{B}$ es reemplazada por la matriz $\boldsymbol{D}$ del mismo método. Entonces, el problema espectral de LLL aplicada a LE (LLL + LE) se expresa de la forma:

Finally, whereas the solution is applied to LE, the affinity matrix $\boldsymbol{A}$ is replaced by the laplacian $\boldsymbol{L}$ which contains one relationship of the affinity matrix $\boldsymbol{W}$ of LE and matrix $\boldsymbol{B}$ is replaced by the matrix $\boldsymbol{D}$ of the same method. Then, the spectral problem of LLL applied to LE (LLL + LE) is expressed as follows:

$$\begin{aligned} \min_{X} \quad & tr(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T) \\ \text{s. t.} \quad & \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T = \boldsymbol{I}, \end{aligned} \tag{5}$$

and the approximate problem can be described by the next expression:

$$\begin{aligned} \min_{\widetilde{X}} \quad & tr(\widetilde{\boldsymbol{X}}\bar{\boldsymbol{L}}\widetilde{\boldsymbol{X}}^T) \\ \text{s. t.} \quad & \widetilde{\boldsymbol{X}}\bar{\boldsymbol{D}}\widetilde{\boldsymbol{X}}^T = \boldsymbol{I}, \end{aligned} \tag{6}$$

siendo $\bar{\boldsymbol{L}} = \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T, \bar{\boldsymbol{D}} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T \in \mathbb{R}^{L\times L}$ las matrices en el problema espectral reducido. La solución del problema aproximado esta dada por:

where $\bar{\boldsymbol{L}} = \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T, \boldsymbol{D} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T \in \mathbb{R}^{L\times L}$ are the matrices in the reduced spectral problem. The solution of the approximate problem is defined by:

$$\widetilde{\boldsymbol{X}} = \bar{\boldsymbol{U}}_d^T \bar{\boldsymbol{D}}^{-\frac{1}{2}}. \tag{7}$$

Es claro que una de las soluciones al problema es $\bar{\boldsymbol{L}}1 = 0$, donde $\lambda_1 = 0$ es la solución trivial y debe ser descartada.

Como se pudo apreciar en la sección anterior, el reemplazar la matriz $\boldsymbol{W}$ por $\boldsymbol{K}$ no afecta la función objetivo ni la solución a esta. Por tanto, es posible aplicar LLL en la metodología generalizada propuesta que toma como base fundamental a LE. De esta forma, la función objetivo aproximada tiene la forma:

It is well known that the trivial solution of the problem is $\bar{\boldsymbol{L}}1 = 0$, where $\lambda_1 = 0$ and it has to be discard.

The previous section illustrates that if the matrix $\boldsymbol{W}$ is replaced by $\boldsymbol{K}$, neither the objective function and its solution are affected. Thus, it is possible to apply LLL in the proposed generalized methodology which takes as fundamental base the DR method LE and the approximate objective function has the following form:

$$\begin{aligned} \min_{X} \quad & tr(\widetilde{\boldsymbol{X}}\bar{\boldsymbol{L}}\widetilde{\boldsymbol{X}}^T) \\ \text{s. t.} \quad & \widetilde{\boldsymbol{X}}\bar{\boldsymbol{D}}\widetilde{\boldsymbol{X}}^T = \boldsymbol{I}, \end{aligned} \tag{8}$$

siendo $\bar{\boldsymbol{L}} = \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T$, $\bar{\boldsymbol{D}} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T \in \mathbb{R}^{L\times L}$ las matrices en el problema espectral reducido, donde $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$ y $\boldsymbol{D} = Diag\boldsymbol{K}1_L$. La solución del problema aproximado esta dada por:

where $\bar{\boldsymbol{L}} = \boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T$, $\bar{\boldsymbol{D}} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T \in \mathbb{R}^{L\times L}$ are the matrices in the reduced spectral problem with $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$ y $\boldsymbol{D} = Diag\boldsymbol{K}1_L$. Therefore, The solution for the approximate problem is given by:

$$\widetilde{\boldsymbol{X}} = \bar{\boldsymbol{U}}_d^T \bar{\boldsymbol{D}}^{-\frac{1}{2}}. \tag{9}$$

3.1 Alternativas existentes

Así como LLL utiliza submatrices en la solución de problemas espectrales, existen otras propuestas para reducir el costo computacional. Entre estas destacan el uso de el método de Nȳstrom [10], submatrices que arrojan afinidades mediante una medida de distancias de desplazamiento [11], procesos de triangulación [12], uso de la distancia geodésica en vez de la distancia euclidiana [3] y una modificación a LLE usando ponderación local de vectores para la construcción de la matriz $\boldsymbol{Z}$ (MLLE) [13]. La ventaja que presenta LLL en relación a todas estas alternativas es que por una parte, los enfoques de Nȳstrom tienen en cuenta un número mucho más pequeño de submatrices, por lo que se podría perder información o incluso, no se lograría una buena reconstrucción del espacio original, es decir, en estos enfoques se proyecta la solución obtenida mediante las submatrices a puntos que están fuera de estas. Dado que LLL considera todo el espacio de entrada, el error en comparación al enfoque de Nȳstrom se reduce [9]. Además, utilizar una medida de distancia de desplazamiento para determinar la afinidad entre puntos es computacionalmente costosa al aumentar la dimensión, LLL no define nuevas afinidades puesto que las afinidades entre submatrices consideran la información en puntos fuera de ellas [9]. Finalmente, en MLLE se realiza un mapeo localmente lineal para producir un nuevo problema espectral, en cambio en LLL se utiliza un mapeo lineal pero para redefinir el problema espectral existente de forma aproximada [9].

There are others approaches proposed in order to reduce the computational cost apart from LLL which takes into account the use of sub-matrices for the solution of spectral problems. Among others approaches, it is possible to highlight some of them.

Algorithm 1 Pseudo-código para implementación de LLKL

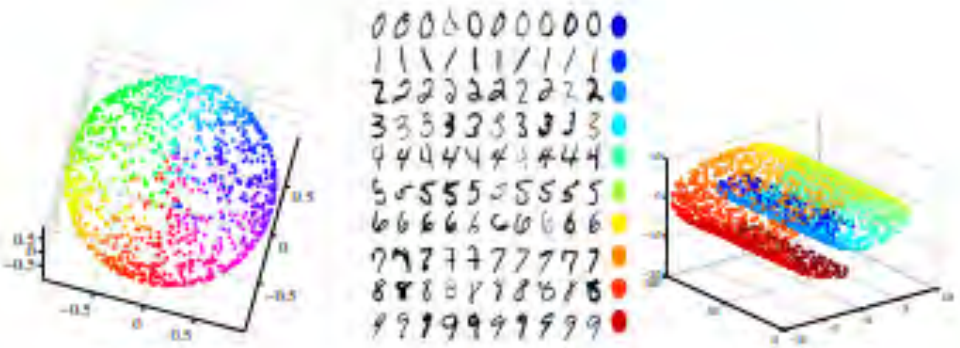
1. **Verificación 1:** Verificar si los datos consisten de distancias en pares, si es así saltar al paso 3.
2. **Verificación 2: Si** los datos consisten de vectores, **hacer**:
   Calcular todas las distancias en pares
   Fin
3. **Determinar:** Escoger para determinar los k vecindarios o los ϵ vecindarios esféricos.
4. **Construcción del grafo:** Construir el grafo correspondiente y su matriz de proximidad $\boldsymbol{P}$.
5. **Construcción del kernel $\boldsymbol{K}$:** Construir la matriz $\boldsymbol{K}$ mediante algún tipo de *kernel*, para este caso los presentados en la sección **??**.
6. **Construcción de $\boldsymbol{D}$:** Construir la matriz $\boldsymbol{D}$ mediante $\boldsymbol{D} = Diag(\boldsymbol{K}\mathbf{1}_N)$.
7. **Calcular el laplaciano:** Calcular $\boldsymbol{L} = \boldsymbol{K} - \boldsymbol{D}$.
9. **Descomposición en valores propios:** Aplicar DVP a la matriz del laplaciano $\boldsymbol{L}$.
10. **Representación en baja dimensión: Realizar**
    a) Multiplicar los vectores propios por $\boldsymbol{D}^{1/2}$.
    b) Trasponer el resultado.
    b) Determinar los valores asociados a los d valores propios más pequeños sin considerar el ultimo valor propio.
    Fin

For instance, the N*ij*strom explained in [10], the application of sub-matrices that provide affinities through a measurement of displacement distances [11], triangulation processes [12], use of geodetic distance instead of Euclidean distance [3] and a modification of LLE method with local weighting of vectors for the matrix $\boldsymbol{Z}$ (MLLE) construction [13]. The benefit of use LLL is that the methods previously mentioned use a reduced number of sub matrices, in consequence, a lost of information and even an inaccurate reconstruction of the original space can be expected. Due to the fact that LLL consider the entire input space the error rate is reduce compared to N*ij*strom approach [9]. In addition, if a measure of displacement distance is used in order to obtain the affinity between points, the computational cost increase since the dimensionality increase. LLL method do not define new affinities becasuse the affinities between submatrices consider the information in points outside of them [9]. Finally, in MLLE a locally linear mapping is made in order to produce a new spectral problem. Nevertheless, LLL uses a linear mapping to redefine approximately the existing spectral problem [9].

4 Experimental setup

Databases: 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 a randomly selected subset of the `MNIST` image bank [6], 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 $D = 24^2$). The third 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.



(a) *Sphere* (b) *MNIST* (c) *Swiss Roll*

Fig. 3. The three considered datasets.

Kernels for DR: Three kernel approximations for spectral DR methods [4] are considered. Namely, classical multidimensional scalling (CMDS), locally linear embedding (LLE), and graph Laplacian eigenmaps (LE). CMDS kernel is the double centered distance matrix $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ so

$$\boldsymbol{K}^{(1)} = \boldsymbol{K}_{CMDS} = -\frac{1}{2}(\boldsymbol{I}_N - \boldsymbol{1}_N \boldsymbol{1}_N^\top)\boldsymbol{D}(\boldsymbol{I}_N - \boldsymbol{1}_N \boldsymbol{1}_N^\top), \tag{10}$$

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 $\boldsymbol{\mathcal{W}}$ holding linear coefficients that sum to 1 and optimally reconstruct observed data. Define a matrix $\boldsymbol{M} \in \mathbb{R}^{N \times N}$ as $\boldsymbol{M} = (\boldsymbol{I}_N - \boldsymbol{\mathcal{W}})(\boldsymbol{I}_N - \boldsymbol{\mathcal{W}}^\top)$ and λ_{max} as the largest eigenvalue of $\boldsymbol{M}$. Kernel matrix for LLE is in the form

$$\boldsymbol{K}^{(2)} = \boldsymbol{K}_{LLE} = \lambda_{max}\boldsymbol{I}_N - \boldsymbol{M}. \tag{11}$$

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

$$\boldsymbol{K}^{(3)} = \boldsymbol{K}_{LE} = \boldsymbol{L}^\dagger. \tag{12}$$

where $\boldsymbol{L} = \boldsymbol{\mathcal{D}} - \boldsymbol{S}$, $\boldsymbol{S}$ is a similarity matrix and $\boldsymbol{\mathcal{D}} = \text{Diag}(\boldsymbol{S}\boldsymbol{1}_N)$ is the degree matrix. All previously mentioned kernels are widely described in [4]. The similarity matrix $\boldsymbol{S}$

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 [**?**]. 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 entry are given by $\exp(-0.5||\boldsymbol{y}_i - \boldsymbol{y}_j||/\sigma^2)$ with $\sigma = 0.1$. For all methods, input data is embedded into a 2-dimensional space, then $d = 2$.

Accordingly, the MKL approach is performed considering $M = 4$ kernels. The generalized kernel provided $\bar{\boldsymbol{K}}$ here as well as the individual kernels $\boldsymbol{K}^{(1)}, \ldots, \boldsymbol{K}^{(M)}$ are tested on kernel PCA as explained in [5].

Performance measure: To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [7] 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

Para demostrar la aplicabilidad de nuestra metodología, la figura x1 a x3 presenta una comparación de la representación obtenida por los métodos convencionales de RD espectral, entre la representación obtenida con KPCA y, finalmente, con la representación obtenida por KLE.

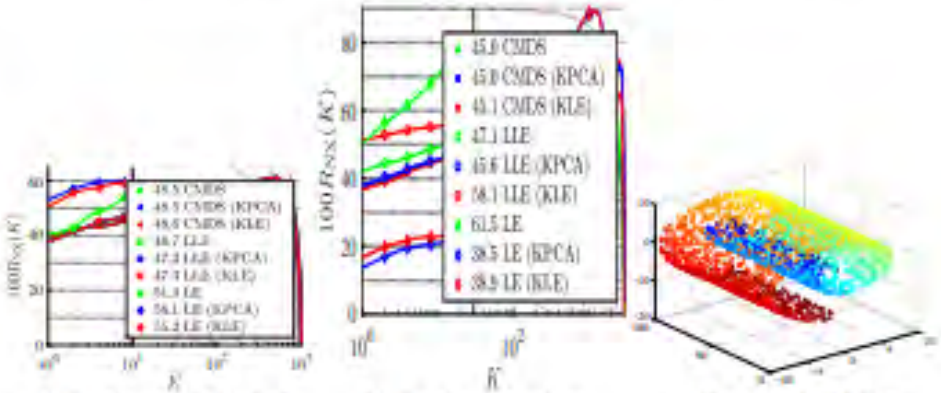
Como se puede apreciar en las imágenes, los tres métodos permiten obtener una representación en dos dimensiones de los datos originales. De estas representaciones se puede apreciar que nuestra metodología genera una representación en 2D de los datos originales, demostrando que los resultados en baja dimensión son separables y conservan su topología.

Para comparar la calidad de la representación se utiliza una medida de AUC dada por la curva $R_{NX}(\mathbf{k})$, estas curvas se realizan para cada método aplicado a cada base de datos utilizada. Los resultados de aplicar cada método de forma convencional se muestran en la figura Xa, los resultados de KPCA en la figura Xb y los resultados de figura Xc.

Estos resultados permiten determinar que la calidad de nuestro enfoque es aproximadamente igual que KPCA, lo que demuestra que nuestra metodología es comparable en representación y resultados. Para demostrar la reducción del costo computacional se seleccionaron valores para L = XXX y kZ = xxxx. Dichos valores fueron obtenidos después de correr pruebas para diferentes valores de L entre $[1, N]$ y $kZ = [1, 50]$. Finalmente, la figura XX presenta los resultados obtenidos para KPCA en relación a los obtenidos para KLE.

With the objective to depict the methodology usability, the figures between x1 and x3 represent a comparison of the embedded spaces obtained with KPCA and KLE methods, and its conventional representation obtained with each method. The graphics illustrate, a two-dimensional representation of the original data which is obtained through the application of three different dimensionality reduction techniques.

These results allow to measure the quality of the proposed approach in both, in a visual way where it is possible to highlight some patterns and trends that is not observable in the raw data and in a numeric way with the area under the curve given by the function $R_{NX}(\mathbf{k})$. This curves are created for each method which is applied to the databases mentioned previously. On the one hand the results of AUC of classical approaches are showed in figure Xa, on the other hand the results of AUC of use the approximate methods KPCA and KLE are illustrate in the figures xb and xc, respectively.



(a) *Results of a Swiss Roll* (b) *Results of a sphere* (c) *Results of a MNIST*

Fig. 4. The three considered datasets.

The information presented in the results allow to determine the quality of this approach due to the approximation of KPCA. Thus, the methodology proposed is comparable in results and representation.

In order to demonstrate the computational cost reduction, values of L = XXX and kZ = xxxx were selected. These values were obtained after running tests for different values of L between $[1, N]$ and $kZ = [1, 50]$. Finally, Figure XX presents a comparison between the results obtained for KPCA and the results obtained for KLE.

6 Conclusions and future work

In this work, a multiple kernel learning approach for dimensionality reduction tasks is presented. The core of this approach is the generalized kernel that is calculated by means of a linear combination of kernel matrices representing spectral dimensionality reduction methods, where the coefficients are obtained from a variable ranking based on a variance criterion. Proposed approach improves both data visualization and preservation by exploiting the representation ability of every single technique.

As future work, new multiple kernel learning approaches will be explored by combining kernel representations arising from other dimensionality reduction methods, aimed at reaching a good trade-off between preservation of data structure and intelligible data visualization.

References

1. Borg, I., Groenen, P.J.: Modern multidimensional scaling: Theory and applications. Springer Science & Business Media (2005)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15**(6) (2003) 1373–1396
3. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. science **290**(5500) (2000) 2323–2326
4. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the twenty-first international conference on Machine learning, ACM (2004) 47
5. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, IEEE (2014) 171–177
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
7. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing **112** (2013) 92–108
8. Saul, L.K., Weinberger, K.Q., Ham, J.H., Sha, F., Lee, D.D.: Spectral methods for dimensionality reduction. Semisupervised learning (2006) 293–308
9. Vladymyrov, M., Carreira-Perpinán, M.Á.: Locally linear landmarks for large-scale manifold learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (2013) 256–271
10. Drineas, P., Mahoney, M.W.: On the Nyström method for approximating a Gram matrix for improved kernel-based learning. Journal of Machine Learning Research **6**(Dec) (2005) 2153–2175
11. Luxburg, U.V., Radl, A., Hein, M.: Getting lost in space: Large sample analysis of the resistance distance. In: Advances in Neural Information Processing Systems. (2010) 2622–2630
12. De Silva, V., Tenenbaum, J.B.: Sparse multidimensional scaling using landmark points. Technical report, Technical report, Stanford University (2004)
13. Zhang, Z., Wang, J.: Mlle: Modified locally linear embedding using multiple weights. In: Advances in neural information processing systems. (2007) 1593–1600

ANEXO 6. PÁGINA WEB

Dentro del desarrollo de este proyecto, se contempla la creación de una página web en Google Sites, donde, se puede encontrar información general acerca de la herramienta implementada, así como, el código fuente y el archivo ejecutable, para probar la aplicación en cualquier ordenador que cuente con Windows. Un manual de usuario y un video tutorial que explica el funcionamiento de la herramienta son incluidos con el fin de dar un mejor entendimiento acerca del proyecto y fomentar la divulgación de los resultados obtenidos.

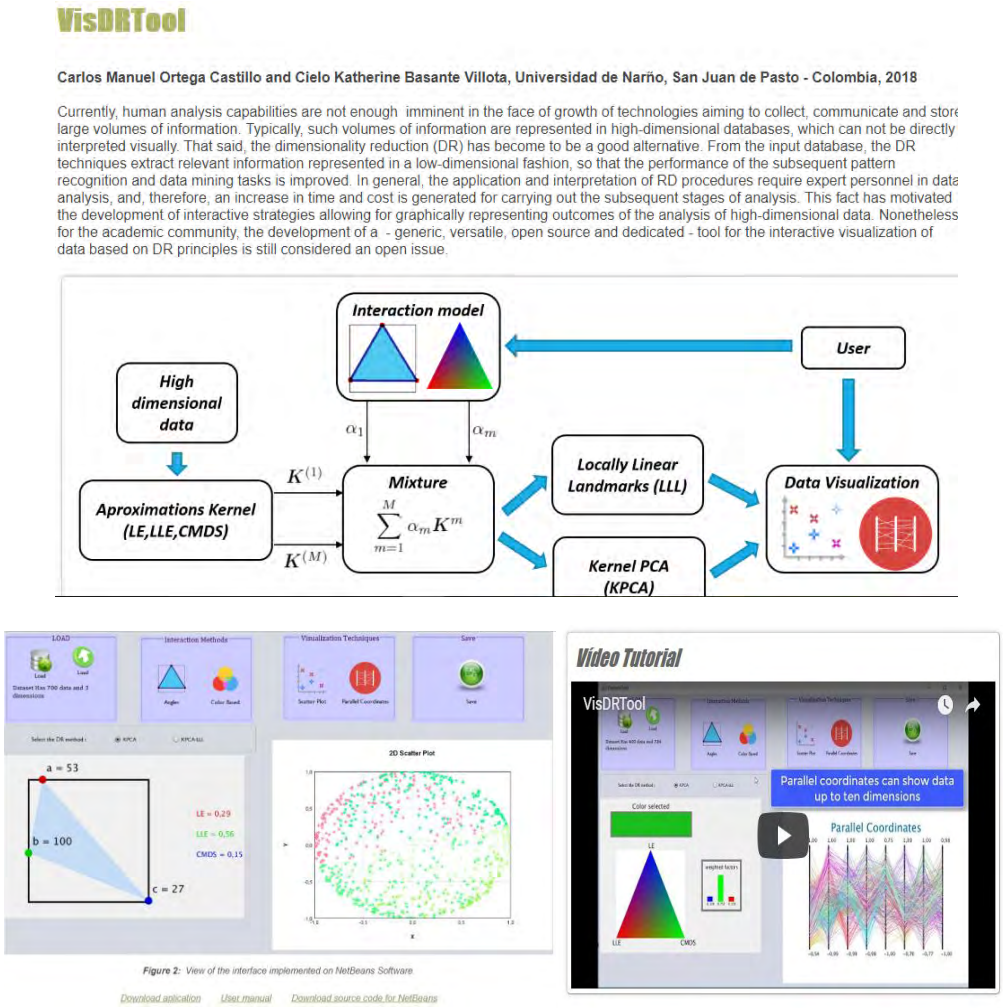


Figura 37. Diseño de la página web[3] en donde se encuentra la información acerca de la herramienta asi como el código fuente, ejecutable, manuales y tutoriales. **Fuente:** Esta investigación.

[3] https://sites.google.com/site/degreethesisdiegopeluffo/datavistool