Tecno Lógicas
ISSN: 0123-7799
tecnologicas@itm.edu.co
Instituto Tecnológico Metropolitano
Colombia

Valencia-Aguirre, Juliana; Daza-Santacoloma, Genaro; Acosta, Carlos D.; Castellanos-Domínguez, Germán

Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades
Tecno Lógicas, núm. 25, diciembre, 2010, pp. 131-150
Instituto Tecnológico Metropolitano
Medellín, Colombia

Disponible en: http://www.redalyc.org/articulo.oa?id=344234320008

Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades

Juliana Valencia-Aguirre[1]
Genaro Daza-Santacoloma[2]
Carlos D. Acosta[3]
Germán Castellanos-Domínguez[4]

Resumen

En este trabajo se realiza una comparación de las principales técnicas de reducción de dimensión no lineal basadas en análisis por localidades, tales como: Locally linear embedding, Isometric feature mapping y Maximum variance unfolding. El estudio pretende determinar, bajo criterios objetivos, cuál de las técnicas consideradas conserva de mejor manera las propiedades locales de la variedad, y la estructura global de los datos de entrada al realizar un mapeo a un espacio de menor dimensión. Los métodos son especialmente analizados en aplicaciones de visualización. Las inmersiones obtenidas son evaluadas por medio de dos criterios: Error de Conservación de Vecindarios y Promedio de Vecinos Conservados. Para la validación experimental se utilizan bases de datos artificiales y reales que permiten confirmar visualmente la calidad de las inmersiones obtenidas. Con base en los resultados se observa que la técnica Maximum variance unfolding presenta inmersiones de mejor calidad, debido a que la técnica de optimización de este algoritmo preserva exactamente las

1 Grupo de Control y Procesamiento Digital de Señales, Universidad Nacional de Colombia sede Manizales, jvalenciaag@unal.edu.co
2 Grupo de Control y Procesamiento Digital de Señales, Universidad Nacional de Colombia sede Manizales, gdazas@unal.edu.co
3 Grupo de Control y Procesamiento Digital de Señales, Universidad Nacional de Colombia sede Manizales, cdacostam@unal.edu.co
4 Grupo de Control y Procesamiento Digital de Señales, Universidad Nacional de Colombia sede Manizales, cgcastellanosd@unal.edu.co

Fecha de recepción: 16 de Agosto de 2010
Fecha de aceptación: 04 de Noviembre de 2010

distancias entre puntos cercanos en el espacio de baja dimensión, conservando la estructura global de la variedad analizada.

Palabras clave

Análisis por localidades, isometric feature mapping, locally linear embedding, maximum variance unfolding, reducción de dimensión.

Abstract

In this paper, a comparison of methods for nonlinear dimensionality reduction is proposed in order to determine which technique preserves better the local properties, without losing the overall structure of the original data. We seek to establish which of these methods is the most appropriate for visualization tasks. The embeddings obtained with each technique are evaluated by two criteria Preservation Neighborhood Error and Preserved Neighbors Average. The methodologies were tested on artificial and real-world data sets which allow us to visually confirm the quality of the embedding. The results obtained show that Maximum variance unfolding computes high quality embeddings, because the optimization problem pretends to preserve exactly the local pair-wise distance between neighbors and conserve the global manifold structure.

Keywords

Dimensionality reduction, isometric feature mapping, local analysis, locally linear embedding, maximum variance unfolding.

1. INTRODUCCIÓN

La reducción de dimensión es una de las etapas más importantes en problemas de reconocimiento de patrones, pues permite revelar la estructura intrínseca de los datos y extraer la información más relevante del problema en estudio, mejorando el desempeño tanto en tareas de visualización como de clasificación.

Entre los métodos tradicionales de reducción de dimensión se encuentran el análisis de componentes principales – PCA (Jolliffe, 2002), y escalado multidimensional - MDS (Cox & Cox, 1994). En PCA se calculan las proyecciones lineales que posean la mayor varianza a partir de los vectores propios asociados a los valores propios más grandes de la matriz de covarianza de los datos. En MDS, se calcula la inmersión al espacio de baja dimensión que mejor conserve las distancias entre pares de puntos de los datos de entrada. No obstante, los métodos lineales de reducción de dimensión no son apropiados para descubrir estructuras subyacentes de datos que residen en variedades no lineales. Con el fin de solucionar este inconveniente, en la última década se han desarrollado nuevos métodos, cuyo objetivo principal es realizar un mapeo no lineal a un espacio de baja dimensión a partir de información local de los datos de alta dimensión, proyectándolos de manera que se preserve la geometría local y la estructura global de los datos originales. Entre los métodos propuestos se encuentran: Locally linear embedding - LLE (Saul & Roweis, 2003), Isometric feature mapping – Isomap (Tenenbaum, 1998) y Maximum variance unfoldin - MVU (Weinberger & Saul, 2006), etc.

Todos los métodos que efectúan análisis por localidades tienen en común la sintonización de un parámetro libre: el número vecinos (usualmente denotado como k), el cual tiene gran influencia en la calidad de la inmersión resultante, por lo que es de gran importancia determinar un valor apropiado para dicho parámetro.

Ante la diversidad de métodos de reducción de dimensión no lineal que han surgido, elegir una técnica específica para analizar datos que residen en variedades no lineales y que presentan estructuras complejas, se ha convertido en una tarea difícil, pues

es necesario estudiar a fondo cada uno de los métodos para evaluar su efectividad y desempeño. Por tal motivo, en este trabajo se realiza una comparación de tres técnicas de reducción de dimensión no lineal: LLE, Isomap y MVU, evaluando los resultados obtenidos con cada una de ellas por medio de dos criterios que permiten determinar la calidad de la inmersión resultante, e identificar posibles traslapes y pérdida de la estructura global de los datos de alta dimensión. Se realizan pruebas con cada una de las técnicas sobre dos bases de datos artificiales conocidas en la literatura y sobre dos bases de datos de imágenes.

En la primera sección de este trabajo se exponen las técnicas de reducción de dimensión a utilizar, realizando una breve descripción de su fundamentación matemática. En la segunda sección se presentan los criterios utilizados para medir la calidad de las inmersiones obtenidas y se realiza una descripción de las bases de datos utilizadas en las pruebas. En la tercera y cuarta sección se presentan los resultados y la discusión. Finalmente se presentan las conclusiones de la comparación entre las técnicas de reducción de dimensión no lineal.

2. TÉCNICAS DE REDUCCIÓN DE DIMENSIÓN

2.1 Locally Linear Embedding – LLE

LLE es un algoritmo de aprendizaje no supervisado que realiza un mapeo de los datos a un subespacio de baja dimensión preservando la geometría local del espacio de datos de alta dimensión (Saul & Roweis, 2003).

Sea $\mathbf{X}$ la matriz de datos de entrada de tamaño $n \times p$ donde se tienen los vectores observación $\mathbf{x}_i \in R^p, i = 1, \ldots, n$. Se asume que los datos viven en una variedad no lineal, adecuadamente muestreada, donde cada dato y sus respectivos k-vecinos más cercanos se encuentran ubicados sobre una región lineal de la variedad. De esta manera, los puntos del espacio de alta dimensión pueden ser aproximados como combinaciones lineales ponderadas de sus vecinos más cercanos y posteriormente

mapeados a un espacio de menor dimensión m donde se conserve la geometría local de los datos (Polito & Perona, 2001). Como salida se tienen n puntos $\mathbf{y}_i \in R^m$, $i = 1,\ldots,n$ donde $m < p$.

El algoritmo LLE consta de 3 etapas. En primer lugar, para cada $\mathbf{x}_i$ se calculan los k vecinos más cercanos empleando distancia Euclídea. Después de determinar los vecindarios, se encuentran los pesos de reconstrucción $\mathbf{W}$ que minimicen

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{n} w_{ij}\mathbf{x}_j \right\|^2 \tag{1}$$

La ecuación (1) está sujeta a dos restricciones: la restricción de dispersión $w_{ij} = 0$ si $\mathbf{x}_j$ no es vecino de $\mathbf{x}_i$, y la restricción de invarianza $\sum_{j=1}^{n} w_{ij} = 1$.En el tercer y último paso se calculan los vectores $\mathbf{y}_i$ mejor reconstruidos por los pesos w_{ij} tal que minimicen

$$\Phi(\mathbf{Y}) = \sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j=1}^{n} w_{ij}\mathbf{y}_i \right\|^2 \tag{2}$$

sujeto a $\sum_{i=1}^{n} \mathbf{y}_i = 0$, y $\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i^T = \mathbf{I}_{m\times m}$. Donde $\mathbf{Y}_{n\times m}$ es la matriz de datos mapeados al espacio de baja dimensión. Ahora bien, sea $\mathbf{M} = \left(\mathbf{I}_{n\times n} - \mathbf{W}^T\right)\left(\mathbf{I}_{n\times n} - \mathbf{W}\right)$, se puede reescribir (2) como

$$\Phi(\mathbf{Y}) = \mathrm{tr}\left(\mathbf{Y}^T\mathbf{M}\mathbf{Y}\right) \text{ sujeto a } \begin{cases} \mathbf{1}_{n\times 1}^{\mathrm{T}}\mathbf{Y} = 0_{1\times m} \\ \frac{1}{n}\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_{m\times m} \end{cases} \tag{3}$$

Para encontrar $\mathbf{Y}$ que minimice la expresión (3) se calculan los *m+1* vectores propios de $\mathbf{M}$, asociados a los *m+1* valores propios más pequeños. El primer vector propio es excluido. Los restantes *m* vectores propios producen la inmersión final $\mathbf{Y}$.

2.2 Isometric Feature Mapping – Isomap

Isomap es una técnica de reducción de dimensión no lineal basada en MDS clásico. Su objetivo es preservar la geometría intrínseca de los datos reflejada en las distancias geodésicas de la variedad. La clave es encontrar una forma eficiente de calcular la verdadera distancia geodésica entre observaciones, dada solamente su distancia Euclídea en el espacio de alta dimensión (Tenenbaum et al., 2000).

Para puntos vecinos, la distancia Euclídea del espacio de alta dimensión proporciona una buena aproximación a la distancia geodésica. Para puntos lejanos, la distancia geodésica puede ser aproximada añadiendo una secuencia saltos entre puntos vecinos (Tenenbaum et al., 2000). Es otras palabras, Isomap asume que la distancia entre puntos en el espacio de características es una medida precisa de la distancia en la variedad sólo a nivel local y que estas distancias deben integrarse en caminos o trayectorias en la variedad con el fin de obtener distancias globales.

En Isomap se asume que los datos se encuentran en una variedad S desconocida, en un espacio de alta dimensión. Se busca realizar un mapeo del espacio original a un espacio de menor dimensión, que preserve lo mejor posible la estructura intrínseca de las observaciones.

El algoritmo de Isomap se compone de tres pasos fundamentales. En primer lugar se determina que puntos son vecinos en la variedad S. Con este propósito es posible utilizar dos criterios simples: el punto i se considera vecino del punto j si pertenece a los k vecinos más cercanos de j o si se encuentra a una distancia menor de un radio ρ. A partir de los vecindarios establecidos se construye un grafo G sobre todos los datos de entrada, y se establecen las longitudes de las aristas como $d_x(ij)$ entre puntos vecinos.

En segundo lugar se realiza el cálculo de los caminos más cortos, estimando las distancias geodésicas $d_S(i, j)$ entre todos los pares de puntos en S y calculando las distancias de los caminos o trayectorias más cortas $d_G(i, j)$ en el grafo G. Inicialmente

$d_G(i,j) = d_X(i,j)$ si i, j están conectados por una arista. En otro caso, es decir si el punto *i* y el punto *j* no se encuentran conectados $d_G = \infty$. La matriz $D_G = \{d_G(i,j)\}$ contiene las distancias de los caminos más cortos entre todos los pares de puntos en el grafo *G*.

Por último se aplica MDS clásico a la matriz de distancias del grafo $D_G = \{d_G(i,j)\}$ construyendo una inmersión de los datos a un espacio de dimensión *m* que preserva de mejor manera la geometría intrínseca de la variedad original. Los puntos $\mathbf{y}_i$ del espacio de salida se eligen de manera que se minimice la función de costo

$$E = \left\| \tau(\mathbf{D}_G) - \tau(\mathbf{D}_Y) \right\|_{L^2} \tag{4}$$

donde $\mathbf{D}_Y$ es la matriz de distancias Euclídeas $\left(d_Y(i,j) = \left\|\mathbf{y}_i - \mathbf{y}_j\right\|\right)$, y $\|A\|_{L^2}$ la norma L^2 de la matriz $\sqrt{\sum_{i,j} A_{ij}^2}$. El operador τ en (4) convierte las distancias en productos internos, el cual caracteriza de forma única la geometría de los datos y que permite una optimización eficiente. El único parámetro libre es el factor de vecindario *k* o ρ, el cual aparece en el primer paso del algoritmo. Dicho parámetro influye sobre el desempeño la técnica, por lo que es importante elegir un valor apropiado para el tamaño del vecindario.

2.3 Maximum Variance Unfolding – MVU

MVU es un algoritmo de reducción de dimensión no lineal cuyo objetivo principal es llevar a cabo la inmersión de los datos a un espacio de baja dimensión, realizando un análisis de localidades, de manera que se preserven exactamente las distancias y ángulos entre puntos cercanos (Weinberger & Saul, 2006). Este método se fundamenta en la noción de simetría, donde simetría se entiende como un mapeo suave e invertible que localmente luce como una rotación más traslación. Considere dos conjuntos de datos $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ y $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$ que tienen correspondencia uno a uno, y sea

Ω una matriz binaria de tamaño $n \times n$ que indica una relación de vecindarios entre **X** y **Y**, tal que $\mathbf{x}_j$ se considera vecino de $\mathbf{x}_i$ si y sólo si $\Omega_{ij} = 1$ (de igual forma para $\mathbf{y}_j$ y $\mathbf{y}_i$). Se dice que **X** y **Y** son localmente isométricos bajo la relación de vecindarios **Ω**, si para cada punto $\mathbf{x}_i$, existe una rotación, reflexión y/o traslación que mapea precisamente $\mathbf{x}_i$ y sus vecinos, en $\mathbf{y}_i$ y sus vecinos.

Considere que el mapeo local entre vecindarios existirá, si y sólo si se preservan exactamente las distancias y los ángulos entre los puntos y sus vecinos. Esta restricción se puede expresar en términos de producto punto. Sean $G_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$ y $K_{ij} = \mathbf{y}_i \cdot \mathbf{y}_j$, los elementos de las matrices de Gram de la entrada **G** y salida **K**, respectivamente. Expresando la condición de isometría local en término de las matrices de Gram se tiene que

$$K_{ii} + K_{jj} - K_{ij} - K_{ji} = G_{ii} + G_{jj} - G_{ij} - G_{ji} \tag{5}$$

Dados n puntos de entrada $\mathbf{x}_i \in R^p$, es posible encontrar $\mathbf{y}_i \in R^m$, donde $m < p$, de forma tal que los puntos de entrada y los puntos de salida sean localmente isométricos, y es posible plantear el problema en términos de matrices de Gram, es decir, se puede encontrar una matriz K_{ij} que satisfaga las restricciones de (5). Además de la condición de isometría local, las salidas $\mathbf{y}_i$ deben están centradas en el origen, de manera que se elimina un grado de libertad de traslación de la solución final, así $\left|\sum_i \mathbf{y}_i\right|^2 = \sum_{ij} \mathbf{y}_i \cdot \mathbf{y}_j = \sum_{ij} K_{ij} = 0$.

Debido a que las restricciones geométricas de las salidas $\mathbf{y}_i$ pueden expresarse en términos de la matriz de Gram K_{ij}, se considera el problema como una optimización sobre las matrices de Gram K_{ij} en lugar de los vectores $\mathbf{y}_i$. Sin embargo, sólo matrices simétricas, con valores propios positivos se pueden interpretar como matrices de Gram y se debe restringir la optimización para el cono de matrices semidefinidas (Weinberger

& Saul, 2004). De esta manera el problema de optimización de MVU se resume como

$$\max\left\{\mathrm{tr}\left(\mathbf{K}\right)\right\}$$
$$sujeto\ a.\begin{cases}\mathbf{K}\ \text{semidefinida positiva}\\ \sum_{ij}K_{ij}=0\\ K_{ii}+K_{jj}-K_{ij}-K_{ji}=G_{ii}+G_{jj}-G_{ij}-G_{ji}\quad \forall ij\quad \Omega_{ij}=1\end{cases} \tag{6}$$

A partir de la matriz de Gram **K** obtenida en el problema de optimización de (6) es posible recuperar las salidas $\mathbf{y}_i$, realizando una descomposición de valores singulares, es decir, a partir de los *m* valores propios más grandes de la matriz **K**.

3. CRITERIOS DE EVALUACIÓN DE CALIDAD DE LA INMERSIÓN

En tareas de visualización, el principal objetivo es realizar un mapeo de los datos a un espacio (de una, dos o tres dimensiones) que preserve lo mejor posible su estructura intrínseca. Por esta razón, cuando se aplica una técnica de reducción de dimensión es importante establecer un criterio que permita conocer la calidad de la transformación y evaluar si los resultados obtenidos son adecuados. La forma más simple de evaluar los resultados de una inmersión es por medio de confirmación visual, sin embargo, esta estrategia es subjetiva y poco fiable. Idealmente la calidad de una inmersión a la salida puede ser juzgada a partir de la comparación de la misma con la estructura de la variedad original, pero generalmente la estructura de dicha variedad no está dada y es difícil de establecer de forma precisa. Por tal motivo, una medida de calidad de inmersión ideal no se puede implementar en forma general, siendo necesario entonces establecer alguna medida alternativa.

En este trabajo se emplean dos criterios de evaluación con el fin de calificar objetivamente el desempeño de los resultados de inmersión obtenidos al aplicar tres técnicas diferentes de reducción de dimensión no lineal. A partir de los resultados de la

reducción se busca determinar cuál método preserva de mejor manera la estructura intrínseca de los datos originales.

3.1 Error de Conservación de Vecindarios – ECV

El ECV está basado en la conservación de la geometría local y la co-ubicación de los vecindarios logrando identificar posibles traslapes en el espacio de baja dimensión (Valencia-Aguirre et al., 2009).

$$ECV(\mathbf{X},\mathbf{Y}) = \frac{1}{2n}\sum_{i=1}^{n}\left\{\frac{1}{k}\sum_{j=1}^{k}\left(D_{(\mathbf{x}_i,\boldsymbol{\eta}_j)} - D_{(\mathbf{y}_i,\boldsymbol{\varphi}_j)}\right)^2 + \frac{1}{k_n}\sum_{j=1}^{k_n}\left(D_{(\mathbf{x}_i,\boldsymbol{\theta}_j)} - D_{(\mathbf{y}_i,\boldsymbol{\gamma}_j)}\right)^2\right\}. \quad (7)$$

En la ecuación (7) D corresponde a la distancia Euclídea estandarizada para obtener un valor máximo igual a 1 y η corresponde al conjunto de vecinos más cercanos de cada punto en el espacio de entrada, una vez realizada la inmersión, para cada $\mathbf{y}_i \in R^m$ se calcula el conjunto β de sus k vecinos más cercanos y se encuentra el conjunto φ correspondiente a las proyecciones de η. Los vecinos estimados en β que no son vecinos en η conforman un nuevo conjunto γ de tamaño k_n, el cual se define como $\boldsymbol{\gamma} = \boldsymbol{\beta} - (\boldsymbol{\beta} \cap \boldsymbol{\varphi})$ (Fig. 1). Además, las proyecciones de los elementos de γ en $\mathbf{X}$ generan el conjunto $\boldsymbol{\theta}$ con k_n elementos. En una inmersión ideal $ECV(\cdot) = 0$ (Daza-Santacoloma et al., 2010, Valencia-Aguirre et al., 2009).

3.2 Promedio de Vecinos Preservados – PVC

El segundo criterio utilizado para evaluar la calidad de la inmersión se basa en calcular el número de observaciones que se preservan como parte de los vecindarios tras la inmersión. De esta manera, teniendo en cuenta los conjuntos explicados anteriormente, es posible calcular el promedio de vecinos preservados como

$$PVC = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\boldsymbol{\varphi}_i \cap \boldsymbol{\beta}_i\right|}{k_i}, \tag{8}$$

En la ecuación (8) se denota la cardinalidad del conjunto como $|\cdot|$. Idealmente el *PVC* debe ser igual a 1.

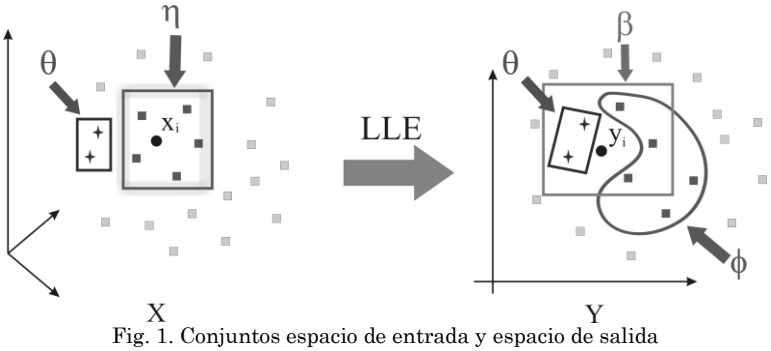


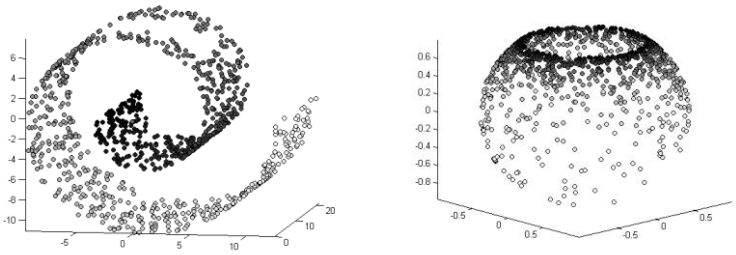
Fig. 1. Conjuntos espacio de entrada y espacio de salida

4. MARCO EXPERIMENTAL

Se realizaron pruebas sobre dos bases de datos artificiales, Rollo Suizo Hueco (Fig. 2a) y Pecera (Fig. 2b). Para las pruebas con datos reales se utilizó la base de datos Columbia Object Image Library (COIL-100), la cual contiene 100 objetos y 72 imágenes a color de cada uno de ellos. Se cuenta con imágenes de los objetos cada 5 grados de giro con una cámara fija. Las imágenes a color (RGB) son obtenidas en formato PNG, con una resolución de 128 × 128 píxeles. Para el proceso de reducción de dimensión se seleccionaron dos objetos de la base COIL-100: Maneki neko (Fig. 3a) y Tetera (Fig. 3b).

Para solucionar los problemas de regularización que presenta el método LLE en el caso de las bases de datos artificiales se utiliza el método propuesto en (Daza-Santacoloma et al., 2010). Además para todas las técnicas se fija la dimensión de salida *m=2* y el número de vecinos más cercanos es escogido usando la

metodología propuesta en (Álvarez-Meza et al., 2010), en la cual se calcula un número específico de vecinos para cada entrada $\mathbf{x}_i$.



a) Rollo Suizo Hueco b) Pecera

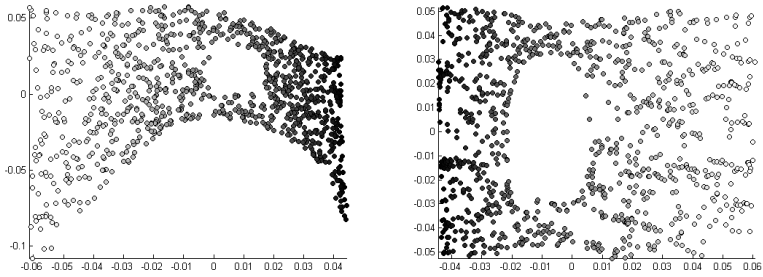
Fig.2. Bases de datos artificiales

a) Maneki neko b) Tetera

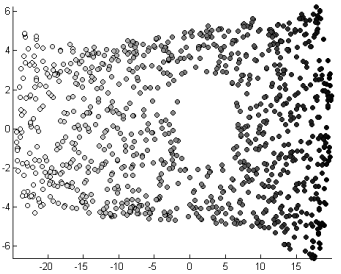
Fig.3. Objetos

5. RESULTADOS

5.1 Base de Datos Artificiales

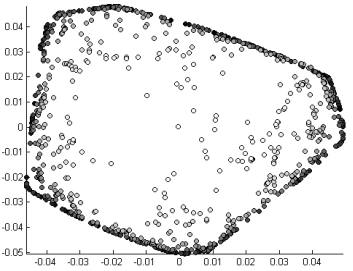


a) LLE b) Isomap

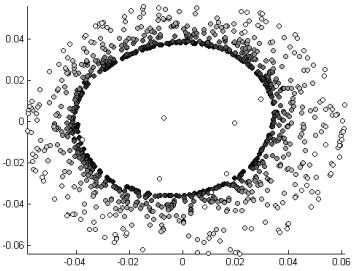


c) MVU

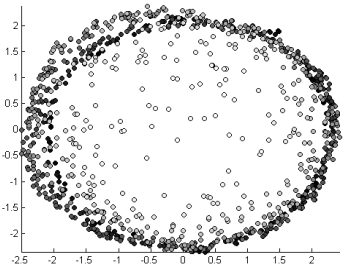
Fig. 4. Resultados de inmersión Rollo Suizo Hueco



a) LLE



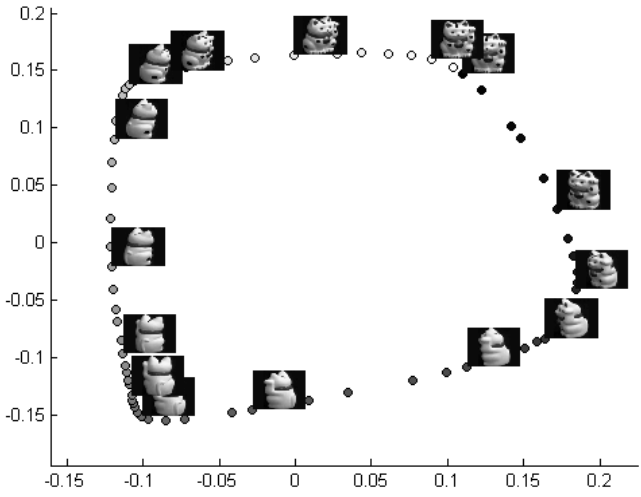
b) Isomap



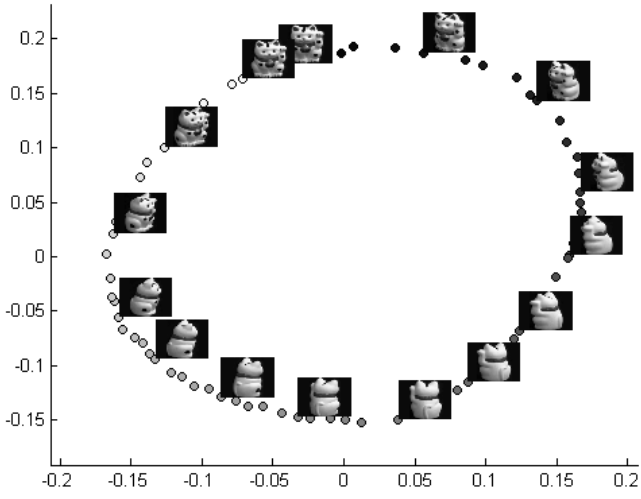
c) MVU

Fig. 5. Resultados de inmersión Pecera

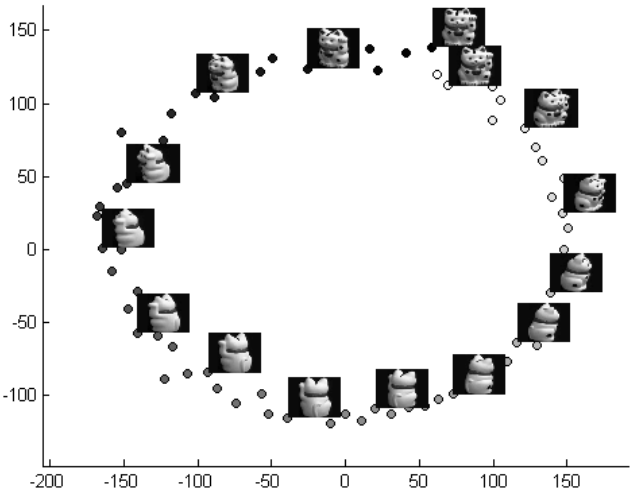
5.2 Base de Datos Reales



a) LLE

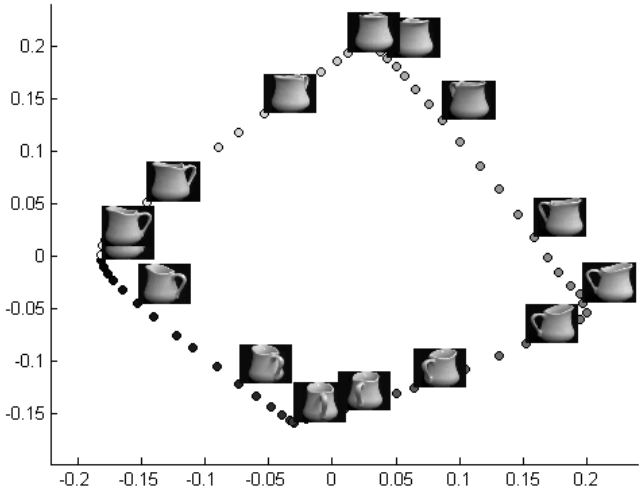


b) Isomap

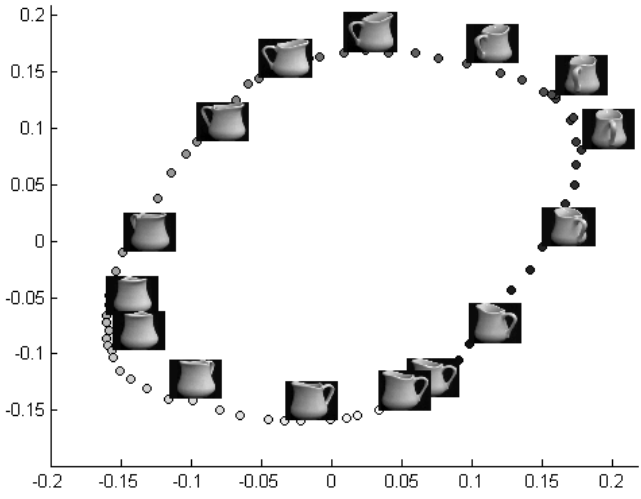


c) MVU

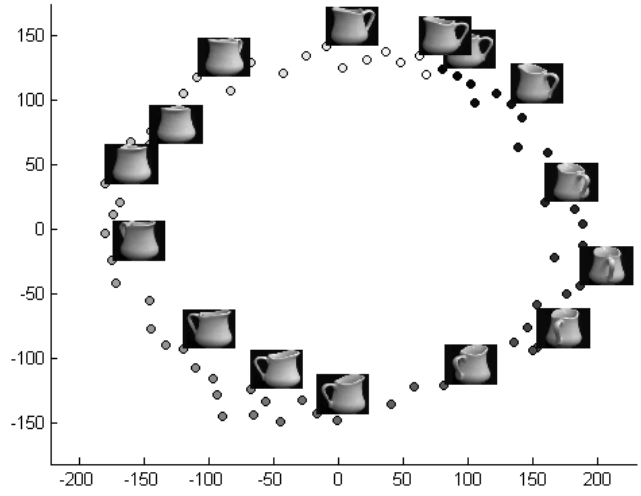
Fig. 6. Resultados de inmersión Maneki neko



a) LLE



b) Isomap



c) MVU

Fig. 7. Resultados de Inmersión Tetera

Tabla 1. Error de conservación de vecindarios, valor ideal ECV=0

Base de datos	LLE	Isomap	MVU
Rollo Suizo Hueco	0,0044	0,0035	0,0031
Pecera	0,0062	0,014	0,0121
Maneki Neko	0,098	0,0863	0,046
Tetera	0,0412	0,0405	0,0226

Tabla 2. Promedio de vecinos preservados, valor ideal PVC=1

Base de datos	LLE	Isomap	MVU
Rollo Suizo Hueco	0,649	0,673	0,864
Pecera	0,6066	0,7103	0,4957
Maneki Neko	0,8426	0,8239	0,9863
Tetera	0,8768	0,847	0,988

6. DISCUSIÓN

En los resultados obtenidos para el Rollo Suizo Hueco (Fig. 4), se observa que los tres métodos realizan un desdoblamiento adecuado de la variedad, originando pocos traslapes en el espacio de salida, conservando la geometría local del espacio original. Sin embargo, las inmersiones calculadas con Isomap (Tenenbaum, 1998) y MVU (Weinberger & Saul, 2006) exhiben una estructura global de los datos más simétrica y consistente, como se ratifica en los resultados de la Tabla 1. Por otra parte, de acuerdo a las transformaciones calculadas para la Pecera (Fig. 5), se puede apreciar que Isomap no logra mapear correctamente los datos de entrada. A pesar de que la técnica Isomap tiende a preservar los vecindarios (Tabla 2), genera traslapes en el espacio inmerso, y pierde totalmente la estructura global de la variedad. Asimismo, MVU no logra mantener la estructura global ni local de la variedad (Tabla 1 y Tabla 2). En este caso, la única técnica que realiza un mapeo apropiado para la Pecera es el algoritmo LLE (Saul & Roweis, 2003), no obstante, la inmersión obtenida exhibe algunos traslapes, aunque conserva de mejor manera la estructura

global de los datos originales. Es importante tener en cuenta que esta variedad presenta regiones con poca densidad de muestras, lo cual afecta el desempeño de los algoritmos.

Ahora, a partir de los resultados obtenidos sobre las bases de datos reales se puede apreciar que todas las técnicas generan inmersiones apropiadas, siendo posible identificar de forma clara la rotación de los objetos, Fig. 6 y Fig. 7. Sin embargo, es evidente que MVU proporciona mayor simetría y suavidad de los datos en el espacio de salida, lo cual se manifiesta en un menor error de conservación de vecindarios y un promedio mayor de vecinos conservados.

7. CONCLUSIONES

A partir de los resultados obtenidos se puede concluir que la técnica de reducción de dimensión no lineal que preserva de mejor manera la estructura intrínseca de los datos y al mismo tiempo conserva su estructura global es Maximum Variance Unfolding. De acuerdo a los resultados obtenidos, este método presenta menos traslapes y conserva la geometría local al realizar la inmersión de los datos. Este algoritmo genera inmersiones de mejor calidad al preservar las distancias locales de manera exacta y al maximizar la varianza de los datos en el espacio de menor dimensión. Así, obtiene inmersiones en donde se conserva la escala original de los datos y concuerda con el desdoblamiento visual esperado. La gran desventaja de esta técnica es el tiempo requerido para resolver el problema con programación semidefinida (SDP), siendo intratable en grandes bases de datos (Weinberger et al., 2005). Por otra parte, aunque el algoritmo LLE tiene en cuenta la información local de la variedad para realizar la inmersión, su desempeño puede verse afectado al elegir de manera incorrecta el parámetro de regularización, el cual influye en la calidad de los resultados. Finalmente, a pesar de que Isomap utiliza la distancia geodésica con el fin de identificar de mejor manera la estructura de la variedad de entrada, presenta dificultades en variedades con regiones poco densas que pueden

generar desconexiones entre los datos y por lo tanto una inmersión inapropiada.

8. AGRADECIMIENTOS

Esta investigación fue financiada gracias a una beca para estudios de maestría y al proyecto DIMA Nº 20201005601 de la Universidad Nacional de Colombia.

9. REFERENCIAS

Álvarez-Meza, A., Valencia-Aguirre, J., Daza-Santacoloma, G., & Castellanos-Domínguez, G., (2010); Global and Local Choice of the Number of Nearest Neighbors in Locally Linear Embedding. Pattern Recognition Letters . Submited.

Cox, T. F. & Cox, M. A. (1994); Multidimensional Scaling, Chapman and Hall,London.

Daza-Santacoloma, G., Acosta-Medina, C.D., & Castellanos-Domínguez, G., (2010); Regularization parameter choice in locally linear embedding. Neurocomputing , 73, 1595–1605.

Jolliffe, I.T., (2002); Principal Component Analysis, 2ª edición, Springer, NY, USA.

Polito, M., & Perona, P., (2001); Grouping and Dimensionality Reduction by Locally Linear Embedding. NIPS.

Saul, L.K., & Roweis, S.T., (2003); Think Globally Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. Machine Learning Research , 4, 119-155.

Tenenbaum, J.B., de Silva, V., & Langford, J.C., (2000); A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science , 290, 2319-232.

Valencia-Aguirre, J., Álvarez-Mesa, A., Daza-Satacoloma, G., & Castellanos-Domínguez, G., (2009); Automatic choice of the number of nearest neighbors in locally linear. 14th Iberoamerican Progress on Pattern Recognition - CIARP 2009, 77-84, Guadalaja, México.

Weinberger, K.Q., & Saul, L.K., (2004); Unsupervised learning of image manifolds by semidefinite programming. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR- 04), 2, 988–995, Washington D.C., USA.

Weinberger, K.Q., Packer, B.D., & Saul, L.K., (2005); Nonlinear Dimensionality Reduction by Semidefinite Programming and Kernel Matrix Factorization. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 381--388.

Weinberger, K.Q., & Saul, L.K., (2006); An introduction to nonlinear dimensionality reduction by maximum variance unfolding. 21st National Conference on Artificial Intelligence. Boston, USA.