Selección automática de parámetros en LLE

Automatic selection of parameters in LLE

Juliana Valencia Aguirre[*]*, Andrés Marino Álvarez Meza, Genaro Daza Santacoloma, Carlos Daniel Acosta Medina, Germán Castellanos Domínguez*

Universidad Nacional de Colombia sede Manizales. Grupo de Control y Procesamiento Digital de Señales, Manizales, Colombia

(Recibido el 10 de diciembre de 2009. Aceptado el 18 de mayo de 2010)

Resumen

Inmersión localmente lineal (LLE) es una técnica de reducción de dimensión no lineal que permite conservar la geometría local del espacio de alta dimensión, al realizar una inmersión de los datos a un espacio de baja dimensión. El algoritmo posee 3 parámetros libres que deben ser definidos por el usuario al momento de realizar la inmersión: el número de vecinos más cercanos k, la dimensión de salida de los datos m y el parámetro de regularización α. Este último sólo es necesario determinarlo cuando el número de vecinos elegido es mayor que la dimensión original de los datos, o cuando los puntos (datos) no están ubicados en posición general, pero juega un papel muy importante en el resultado de la inmersión. En este trabajo se propone un par de criterios que permiten encontrar el valor óptimo para los parámetros k y α, de manera que se obtenga una inmersión que represente de manera fiel los datos del espacio de entrada. Con el fin de comprobar la eficacia de los criterios propuestos, se realizaron pruebas sobre dos bases de datos artificiales y dos bases de datos reales. Además, se realiza una comparación de los resultados contra métodos encontrados en el estado del arte.

----- *Palabras clave:* Inmersiones localmente lineales, número de vecinos más cercanos, regularización automática, reducción de dimensión

Abstract

Locally Linear Embedding (LLE) is a nonlinear dimensionality reduction technique, which preserves the local geometry of high dimensional space performing an embedding to low dimensional space. LLE algorithm has 3 free parameters that must be set to calculate the embedding: the number of nearest neighbors k, the output space dimensionality m and the regularization

* Autor de correspondencia: teléfono: + 57 + 6 + 887 94 00 ext. 55712, fax: + 57 + 6 + 879 40 00 ext. 55713, correo electrónico: julyv51@gmail.com. (J. Aguirre)

parameter α. The last one only is necessary when the value of *k is* greater than the dimensionality of input space or data are not located in general position, and it plays an important role in the embedding results. In this paper we propose a pair of criteria to find the optimum value for the parameters *k* and α, to obtain an embedding that faithfully represent the input data space. Our approaches are tested on 2 artificial data sets and 2 real world data sets to verify the effectiveness of the proposed criteria, besides the results are compared against methods found in the state of art.

----- *Keywords:* Dimensionality reduction, locally linear embedding, number of nearest neighbors, automatic regularization

Introducción

Generalmente, en problemas de reconocimiento de patrones, se cuenta con etapas de caracterización que proveen una gran cantidad de datos acerca de los objetos que se analizan. Si los datos se encuentran en un espacio de dimensión alta es posible que sean redundantes y que escondan información importante para el problema en estudio. Con el fin de facilitar el análisis de datos, representar adecuadamente las observaciones y obtener la información más relevante de un fenómeno particular, surge la reducción de dimensión como una etapa del procesamiento de señales para el reconocimiento de patrones. Además, al disminuir la redundancia de los datos se mejora el desempeño en etapas de clasificación y se obtienen representaciones gráficas fiables, que permiten realizar un análisis visual de los datos [1].

El método más popular de reducción de dimensión es el análisis de componentes principales (PCA), el cual asume que los datos son lineales, y pretende encontrar una base de representación ortogonal, que permita proyectar los datos a un espacio de baja dimensión, en donde se conserve la mayor cantidad de variabilidad. Sin embargo, los métodos lineales de reducción de dimensión no son apropiados para descubrir correctamente estructuras subyacentes de datos que residen en variedades no lineales. Con el fin de solucionar este inconveniente, surgió una técnica no supervisada de reducción de dimensión no lineal, llamada Inmersión localmente lineal (LLE) [2, 3], cuyo objetivo principal es realizar un mapeo a un espacio de baja dimensión donde se preserve la estructura local del espacio original de los datos. Este método requiere de la sintonización manual de tres parámetros: la dimensión de salida *m*, el número de vecinos más cercanos *k* y el parámetro de regularización α; particularmente, *k* y α tienen una gran influencia en las inmersiones resultantes.

En este artículo se propone una metodología para la sintonización automática de α, seleccionándolo a partir de una optimización de problemas mal condicionados. Además, se propone una medida que permite cuantificar la calidad de una inmersión, la cual puede usarse como criterio para determinar un valor adecuado de *k*. Los resultados obtenidos son validados en bases de datos artificiales y reales, comparándolos contra técnicas de optimización de *k* y α encontradas en el estado del arte.

Inmersión localmente lineal - LLE

El algoritmo LLE realiza una inmersión de los datos a un espacio de baja dimensión, teniendo en cuenta que puntos cercanos en el espacio de alta dimensión, deben permanecen juntos y similarmente co-ubicados entre ellos en el espacio de baja dimensión [4].

Sea $\mathbf{X}$ la matriz de datos de entrada de tamaño n x p, donde se tienen los vectores observación $\mathbf{x}_i \in R^p$, $i = 1,...,n$. Se asume que los datos viven en una variedad no lineal o cerca a ella, n es lo suficientemente grande para asegurar que la variedad este bien muestreada, y cada punto y sus vecinos se encuentra ubicados sobre una región

lineal. 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 [5]. Como salida se dan *n* puntos $\mathbf{y}_i \in R^m$, $i = 1,\ldots,n;$ donde $m < p$.

El algoritmo LLE consta de 3 pasos principales. En primer lugar, se determina el vecindario para cada $\mathbf{x}_i$ identificando sus *k* vecinos más cercanos con base en la distancia Euclídea. Después de determinar los vecindarios, cada punto $\mathbf{x}_i$ se representa como una combinación lineal ponderada de sus vecinos, encontrando los pesos de reconstrucción **W** que minimicen

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{n} w_j \mathbf{x}_i \right\|^2, \qquad (1)$$

Se deben tener en cuenta dos restricciones: la restricción de dispersión, en donde $w_{ij} = 0$ si x_j no es vecino de x_i, y la restricción de invarianza $\sum_{j=1}^{n} w_{ij} = 1$. Ahora, suponiendo un punto cualquiera $\mathbf{x}_i \in R^p$ y el conjunto de sus *k* vecinos más cercanos η, se define la matriz de Gram **G** de tamaño *k* x *k*, cuyos elementos están dados como $G_{il} = \left\langle (\mathbf{x} - \boldsymbol{\eta}_j), (\mathbf{x} - \boldsymbol{\eta}_l) \right\rangle$, $j = 1,\ldots,k;\ l = j = 1,\ldots,k.$ Reescribiendo (1) se obtiene

$$\varepsilon = \mathbf{w}^T \mathbf{G} \mathbf{w} \quad \text{sujeto a} \quad \sum_{j=1}^{n} w_{ij} = 1. \qquad (2)$$

Con el fin de minimizar (2), se soluciona un problema de valores propios empleando el Teorema de Lagrange, por tanto se tiene que

$$\mathbf{w} = (\lambda/2)\mathbf{G}^{-1}\mathbf{1}, \ \ \lambda = 2/\mathbf{1}^T \mathbf{G}^{-1}\mathbf{1}, \qquad (3)$$

siendo **1** un vector de unos de tamaño *k* x 1 (a menos que se indique otra cosa). En el tercer paso, se calculan los vectores y_i mejor reconstruidos por los pesos W_{ij} que minimicen

$$\Phi(\mathbf{Y}) = \sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{j=1}^{n} w_j \mathbf{y}_i \right\|^2, \qquad (4)$$

sujeto a $\sum_{i=1}^{n} \mathbf{y}_i = 0$, y $\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I}_{m\times m}$. Sea $\mathbf{M} = (\mathbf{I}_{n\times n} - \mathbf{W}^T)(\mathbf{I}_{n\times n} - \mathbf{W})$, reescribiendo (4)

$$\Phi(\mathbf{Y}) = \mathrm{tr}(\mathbf{Y}^T \mathbf{M} \mathbf{Y}) \ \text{ sujeto a } \begin{cases} \mathbf{1}_{n\times 1}^T \mathbf{Y} = \mathbf{0}_{1\times m} \\ \frac{1}{n}\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_{m\times m} \end{cases} \qquad (5)$$

Para encontrar la matriz $\mathbf{Y}_{n\times m}$ con los datos mapeados al espacio de baja dimensión que minimizan la expresión (5), se pueden calcular los *m* + 1 vectores propios de **M**, asociados a los *m* + 1 valores propios más pequeños, pero se descarta el vector propio asociado al valor propio más pequeño.

Regularización automática

Cuando la matriz de Gram dada en (2) es singular o cerca a serlo, es decir, si **G** no es de rango completo, el problema de mínimos cuadrados que permite hallar la matriz de pesos **W**, no tiene solución única. En este sentido, se hace necesario regularizar **G** antes de encontrar **W**. Esta situación se presenta cuando el número de vecinos más cercanos es mayor que el número de dimensiones del espacio de entrada ($k > p$). El parámetro de regularización juega un papel muy importante al calcular la combinación lineal ponderada de los pesos que permiten representar cada punto a partir de sus vecinos. Es importante tener en cuenta, que el valor óptimo para este parámetro puede variar en un rango amplio y depende de la aplicación en particular, lo cual es parcialmente explicado por cambios en la escala de los datos de entrada [6]. En [2, 3] se propone calcular la versión regularizada de la matriz **G** como

$$G_{jl} \leftarrow G_{jl} + \alpha_1 \quad \text{donde} \qquad \alpha_1 = \delta_{ij}\left(\Delta^2/k\right)\mathrm{tr}(\mathbf{G}) \qquad (6)$$

donde δ_{jl} es 1 si $j = l$ y 0 en otro caso, y $\Delta^2 < 1$. Éste último parámetro debe ser ajustado de forma experimental y se recomienda emplear $\Delta = 0{,}1$ [4].

En [6] se propone regularizar **G** como

$$\mathbf{G} \leftarrow \mathbf{G} + \alpha_2 \mathbf{I}_{k\times k}, \quad (7)$$

siendo $\alpha_2 = \frac{1}{p-m}\sum_{i=m+1}^{p} \hat{\lambda}_i$ la media de los valores propios asociados a los vectores propios descartados; esto puede ser calculado a través de una descomposición espectral para cada vecindario usando PCA.

Por otra parte, en este trabajo se plantea una regularización de la forma

$$\mathbf{G} \leftarrow \mathbf{G} + \alpha_3^2 \mathbf{I}_{k\times k}, \quad (8)$$

donde, el parámetro de regularización α_3, se encuentra de forma automática tomando como base la formulación original del problema de optimización dado en (2) y las propiedades de estabilidad de la solución final calculada. Se busca un $\mathbf{w}$ que minimice el error ε, de esta manera es posible escribir (2) como

$$\begin{aligned} \varepsilon_{reg} &= \mathbf{w}^T \mathbf{G}\mathbf{w} + \alpha_3^2 \mathbf{w}^T \mathbf{w} \\ \text{sujeto a} \quad & \sum_{j=1}^{k} w_j = 1. \end{aligned} \quad (9)$$

La solución de (9), empleando multiplicadores de Lagrange, se expresa a través de un sistema lineal de ecuaciones

$$\begin{bmatrix} 2\left(\mathbf{G} + \alpha_3^2 \mathbf{I}\right) & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ -\lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \quad (10)$$

Cuya solución está dada por

$$\begin{aligned} \lambda_\alpha &= \frac{2}{\mathbf{1}^T \left(\mathbf{G} + \alpha_3^2 \mathbf{I}\right)^{-1} \mathbf{1}}, \\ \mathbf{w}_\alpha &= \frac{\lambda_\alpha}{2} \left(\mathbf{G} + \alpha_3^2 \mathbf{I}\right)^{-1} \mathbf{1} \end{aligned} \quad (11)$$

Además si se analiza la sensibilidad de la solución dada por (3) en presencia de ruido, suponiendo $\tilde{\mathbf{G}} = \mathbf{G} + \mathbf{E}$ como la versión con ruido, entonces se tiene la solución

$$\tilde{\lambda} = \frac{2}{\mathbf{1}^T \tilde{\mathbf{G}}^{-1} \mathbf{1}}, \quad \tilde{\mathbf{w}} = \frac{\tilde{\lambda}}{2} \tilde{\mathbf{G}}^{-1} \mathbf{1} \quad (12)$$

Después de una sencilla manipulación algebraica de (12) se puede aproximar el error de calcular $\mathbf{w}$ como

$$\varepsilon = \begin{bmatrix} \mathbf{G}^{-1}\mathbf{E} & \frac{1}{2} \frac{\mathbf{1}^T \mathbf{G}^{-1} \mathbf{E} \mathbf{G}^{-1} \mathbf{1}}{\mathbf{1}^T \mathbf{G}^{-1} \mathbf{1}} \mathbf{G}^{-1} \mathbf{1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{w}} \\ \tilde{\lambda} \end{bmatrix} \quad (13)$$

De la expresión (13) se puede concluir que para evitar un error grande se debe calcular $\tilde{\mathbf{w}}$ y $\tilde{\lambda}$ acotando su norma. Por esta razón, en condiciones realistas, para la implementación de este tipo de regularización, se debe prestar especial atención al tamaño de la solución. Esta es una situación común en otros casos de regularización [7].

Por tales razones, se propone elegir el parámetro de regularización como

$$\alpha_{opt} = \arg\min_{\alpha} g(\alpha), \quad (14)$$

donde $g(\alpha) = \|\mathbf{x}_\alpha\|^2 = \|\mathbf{w}_\alpha\|^2 + |\lambda_\alpha|^2$. La función $g(\alpha)$ es el producto de las siguientes funciones: λ_α (creciente)y $\left\|\left(\mathbf{G} + \alpha_3^2 \mathbf{I}\right)^{-1} \mathbf{1}\right\|^2$ (no creciente).El valor de $g(\alpha)$ puede ser demasiado grande tanto para valores pequeños como para valores altos de α, por lo que se busca un compromiso entre la estabilidad y la precisión de la solución.

Selección del número de vecinos más cercanos

Este parámetro determina la preservación de la estructura tanto local como global de los datos en el espacio de baja dimensión. Si el valor de *k* es muy pequeño, el algoritmo puede calcular sub-espacios que no reflejen las propiedades globales de la variedad, mientras que si el número de vecinos es alto, el algoritmo puede perder las características no lineales y realizar una reducción de dimensión lineal como PCA [6]. En este sentido, es posible establecer una función de costo que permita cuantificar la calidad de la inmersión obtenida.

Idealmente la calidad de una inmersión a la salida se puede juzgar a partir de la comparación de la

misma con la estructura de la variedad original, sin embargo, generalmente la estructura de dicha variedad no está dada y es difícil de establecer. Por tal motivo, una medida de calidad de inmersión ideal no se puede implementar en forma general, siendo necesario entonces establecer alguna medida alternativa que puede ser utilizada como criterio para determinar el valor del parámetro *k* [8].

En el estado del arte, se encuentran algunas medidas de calidad de una inmersión, utilizadas como criterio para determinar el valor óptimo para el parámetro *k*. En [9] se plantea usar como medida cuantitativa la varianza residual, definida como

$$\sigma_R^2\left(D_{\mathbf{X}}, D_{\mathbf{Y}}\right) = 1 - \rho_{D_{\mathbf{X}}D_{\mathbf{Y}}}^2, \qquad (15)$$

donde $\rho_{D_{\mathbf{X}}D_{\mathbf{Y}}}$ es el coeficiente de correlación lineal estándar, tomado sobre todas las entradas $D_{\mathbf{X}}$ y $D_{\mathbf{Y}}$, siendo estas últimas las matrices de distancias Euclídeas entre cada par de puntos en **X** y **Y** respectivamente. Teóricamente, entre menor sea el valor de la varianza residual mejor representados están los datos de alta dimensión en el espacio de baja dimensión [9] y por tanto el número de vecinos más cercanos se calcula como

$$k_{\sigma_R^2} = \arg\min_k\left(\sigma_R^2\left(D_{\mathbf{X}}, D_{\mathbf{Y}}\right)\right). \qquad (16)$$

Como alternativa para determinar la calidad de una inmersión, en [8] se propone utilizar el estadístico de Procrustes. El estadístico de Procrustes determina la distancia entre dos configuraciones de puntos, calculando la suma de los cuadrados entre parejas de puntos correspondientes, después de que uno de ellos es rotado y trasladado, buscando mejorar la correspondencia con su respectiva pareja.

La medida se define como $P(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n}\left\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i - \mathbf{b}\right\|^2$, donde $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ y $b \in R^m$. **A** es la matriz de rotación y puede ser calculada a partir de $\mathbf{Z} = \mathbf{X}^T\mathbf{H}\mathbf{Y}$, donde $\mathbf{H} = \mathbf{I} - (1/k)\mathbf{1}^T$, **1** es un vector de unos de $n \times 1$ y **H** es la matriz de centralización. Ahora, calculando la descomposición singular de **Z**, se tiene que **A** es dada por $\mathbf{U}\mathbf{V}^T$, el vector de traslación de Procrustes **b** está dado por $\bar{x} - A\bar{y}$, donde $\bar{x}$ y $\bar{y}$ son las medias de **X** y **Y** respectivamente. Se puede definir la calidad de la preservación de los vecindarios locales al realizar la inmersión, calculando el estadístico de Procrustes local $P_L\left(\mathbf{X}_i, \mathbf{Y}_i\right)$ para cada par de vecindarios $\left(\mathbf{X}_i, \mathbf{Y}_i\right)$. Una inmersión global que preserva la estructura local es aquella que minimiza la sumatoria del estadístico de Procrustes para todos los pares de vecindarios de la inmersión [8], de acuerdo a

$$R_N\left(\mathbf{X},\mathbf{Y}\right) = \tfrac{1}{n}\sum_{i=1}^{n} P_L\left(\mathbf{X}_i, \mathbf{Y}_i\right) \Big/ \left\|\mathbf{H}_L\mathbf{X}_i\right\|_F^2, \qquad (17)$$

donde $\mathbf{H} = \mathbf{I} - (1/k)\mathbf{1}^T$, **1** es un vector de unos de *k* x 1 y $\|\cdot\|_F$ es la norma de Frobenius. Así, el número de vecinos más cercanos puede ser calculado como

$$k_{R_N} = \arg\min_k\left(R_N\left(\mathbf{X},\mathbf{Y}\right)\right) \qquad (18)$$

En este trabajo se propone una medida alternativa como criterio para determinar el número de vecinos más cercanos en LLE, basada en la conservación de la geometría local y la co-ubicación de los vecindarios, identificando posibles traslapes en el espacio de baja dimensión. Se pretende medir la calidad de la transformación de cada vecindario, es decir, observaciones cercanas en el espacio de alta dimensión a cada $\mathbf{x}_i$, deben permanecer cercanas en el espacio inmerso para cada $\mathbf{y}_i$, teniendo en cuenta que observaciones que no pertenecían al vecindario seleccionado en el espacio de entrada para cada $\mathbf{x}_i$, no deben ser consideradas como parte del vecindario de la correspondiente inmersión $\mathbf{y}_i$. Por lo tanto, se cuantifica la calidad de la inmersión a partir del error cuadrático medio presentado a continuación

$$C_l(\mathbf{X},\mathbf{Y}) = \frac{1}{2n}\sum_{i=1}^{n} \left(\frac{1}{k}\sum_{j=1}^{k}\frac{\left(D_{(\mathbf{x}_i,\mathbf{n}_j)} - D_{(\mathbf{y}_i,\boldsymbol{\varphi}_j)}\right)^2}{k} + \frac{1}{k_{nv}}\sum_{j=1}^{k_{nv}}\frac{\left(D_{(\mathbf{x}_i,\boldsymbol{\theta}_j)} - D_{(\mathbf{y}_i,\boldsymbol{\gamma}_j)}\right)^2}{k_{nv}}\right), \qquad (19)$$

donde D es la distancia Euclídea estandarizada a uno. Una vez realizada la inmersión, para cada $\mathbf{y}_i \in R^m$, se calcula el conjunto $\boldsymbol{\varphi}$ de sus k vecinos más cercanos y se encuentra el conjunto $\boldsymbol{\beta}$ correspondiente a las proyecciones de $\boldsymbol{\eta}$. Los vecinos estimados en $\boldsymbol{\beta}$ que no son vecinos en $\boldsymbol{\eta}$ conforman un nuevo conjunto de tamaño $k_{nv,}$ definido como $\boldsymbol{\gamma} = \boldsymbol{\beta} - (\boldsymbol{\beta} \cap \boldsymbol{\varphi})$. Además, las proyecciones de los elementos de $\boldsymbol{\gamma}$ en $\mathbf{X}$ generan el conjunto $\boldsymbol{\theta}$ con k_{nv} elementos. En una inmersión ideal $C_I(\cdot) = 0$. El primer término de la medida pretende cuantificar la estabilidad y la conservación de la geometría local de la muestra $\mathbf{x}_i$ y sus k vecinos más cercanos, escogidos en el espacio original **X,** y posteriormente representados en el espacio de baja dimensión **Y**. El segundo término busca establecer el error generado por posibles traslapes de datos en el mapeo de los mismos, lo cual ocurre frecuentemente al incrementar demasiado el número de vecinos. En este segundo término solo se tiene en cuenta los k_{nv} vecinos más cercanos en el espacio de baja dimensión **Y** para cada $\mathbf{y}_i$, que no pertenecen al conjunto de los k vecinos más cercanos del espacio de entrada **X** para cada $\mathbf{x}_i$. Con base en esta medida el número de vecinos más cercanos puede hallarse como

$$k_{C_I} = \arg\min_k (C_I(\mathbf{X}, \mathbf{Y})) \qquad (20)$$

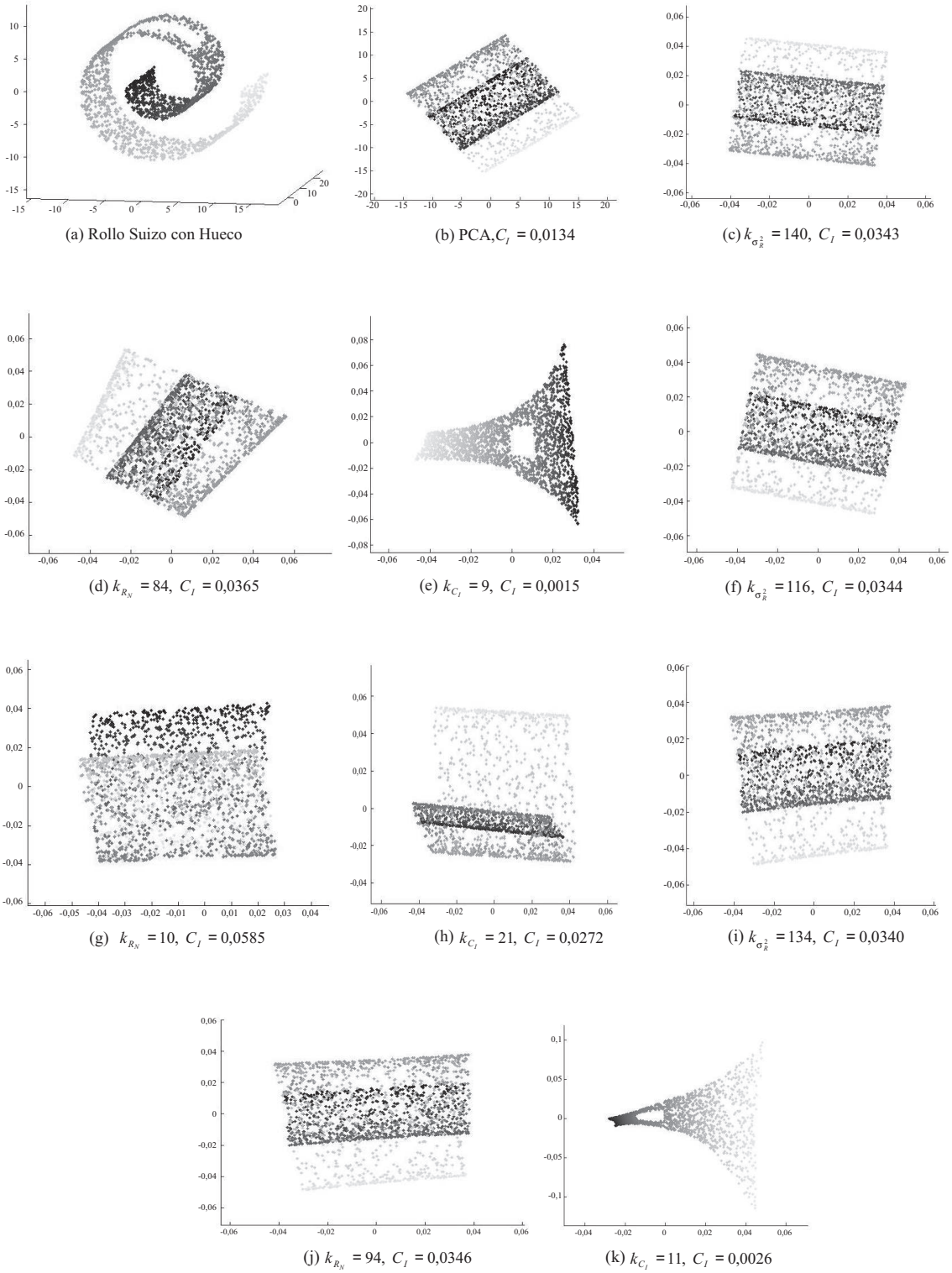
Marco experimental

Bases de datos artificiales

Se realizaron pruebas sobre dos bases de datos artificiales, rollo suizo hueco con n = 2.000 y pecera con n = 1.500. Estas variedades permiten identificar visualmente la calidad del resultado obtenido en la inmersión al seleccionar de forma automática el número de vecinos más cercanos y el parámetro de regularización, comparándolo con el desdoblamiento ideal de la variedad, pues pertenecen a un espacio de entrada tridimensional (p = 3). Con el fin de cuantificar la calidad de la inmersión y determinar el número de vecinos más cercanos necesarios para una correcta reducción de dimensión, se varía el parámetro k para todos los posibles valores desde 3 hasta 250. La calidad de la inmersión se mide de acuerdo a las expresiones (15), (17) y (19), y el número de vecinos más cercanos se determina por medio de (16), (18) y (20). Además el parámetro de regularización se selecciona de acuerdo a (6), (7) y (8), y se establece la dimensión de salida m = 2. Cada técnica de selección de número de vecinos se combina con los tres métodos de regularización mencionados anteriormente, de forma que sea posible comparar la eficacia de los métodos propuestos contra los procedimientos encontrados en el estado del arte. Las inmersiones obtenidas utilizando los diferentes métodos para calcular el número de vecinos más cercanos y el parámetro de regularización se presentan en las figuras 1 y 2.

Bases de datos reales

En este caso se analizan dos bases de datos diferentes, una base de datos de imágenes y una base de datos meteorológica. Para las pruebas con imágenes se utilizó la base de datos Columbia Object Image Library (COIL-100) [10], la cual contiene 100 objetos y 72 imágenes a color de cada uno de ellos sobre una mesa giratoria, que permite una rotación de 360 grados. Se capturan 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 no lineal empleando LLE, se seleccionó un objeto de la base COIL-100: Pato de hule (Figura 3). Estas imágenes se transformaron a escala de grises y se submuestrearon al 50%, obteniendo espacios de entrada con $n = 72$ y p = 8,192. Para determinar la calidad de la inmersión y el tamaño de los vecindarios se varía el parámetro k para los posibles $k \in \{2,4,5,...,50\}$. En el caso de las imágenes no es necesario regularizar, por esta razón se realizan experimentos exclusivamente para validar el método propuesto para el cálculo automático del parámetro k.



(a) Rollo Suizo con Hueco

(b) PCA, $C_I = 0{,}0134$

(c) $k_{\sigma_R^2} = 140,\ C_I = 0{,}0343$

(d) $k_{R_N} = 84,\ C_I = 0{,}0365$

(e) $k_{C_I} = 9,\ C_I = 0{,}0015$

(f) $k_{\sigma_R^2} = 116,\ C_I = 0{,}0344$

(g) $k_{R_N} = 10,\ C_I = 0{,}0585$

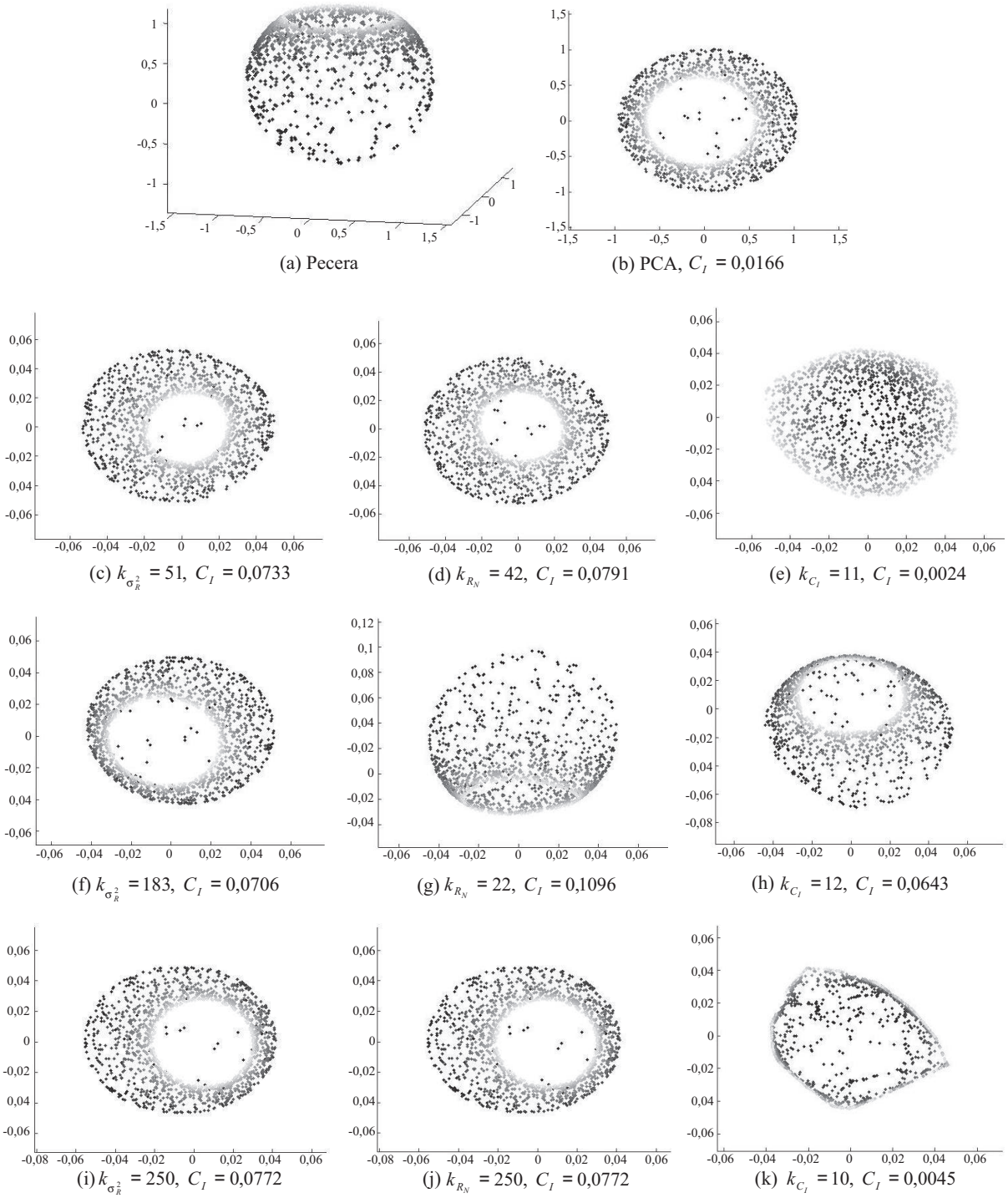
(h) $k_{C_I} = 21,\ C_I = 0{,}0272$

(i) $k_{\sigma_R^2} = 134,\ C_I = 0{,}0340$

(j) $k_{R_N} = 94,\ C_I = 0{,}0346$

(k) $k_{C_I} = 11,\ C_I = 0{,}0026$

Figura 1 Resultados Rollo Suizo Hueco, figura (a) Rollo Suizo Hueco, figura (b) Resultado PCA, figuras (c), (d), (e) Regularización α_1 (f), (g), (h) Regularización α_2, (i), (j) ,(k) Regularización α_3



(a) Pecera (b) PCA, C_I = 0,0166

(c) $k_{\sigma_R^2}$ = 51, C_I = 0,0733 (d) k_{R_N} = 42, C_I = 0,0791 (e) k_{C_I} = 11, C_I = 0,0024

(f) $k_{\sigma_R^2}$ = 183, C_I = 0,0706 (g) k_{R_N} = 22, C_I = 0,1096 (h) k_{C_I} = 12, C_I = 0,0643

(i) $k_{\sigma_R^2}$ = 250, C_I = 0,0772 (j) k_{R_N} = 250, C_I = 0,0772 (k) k_{C_I} = 10, C_I = 0,0045

Figura 2 Resultados Pecera, figura (a) Pecera, figura (b) Resultado PCA, figuras (c), (d), (e) Regularización α_1 (f), (g), (h) Regularización α_2, (i), (j) ,(k) Regularización α_3

La base de datos meteorológica es utilizada con el fin de verificar el desempeño de la técnica propuesta para determinar automáticamente tanto el parámetro de regularización como el número de vecinos más cercanos. Esta base de datos fue desarrollada en el proyecto ECA&D (European Climate Assessment and Dataset) [11], los datos se componen de un resumen diario del clima de Berlín-Alemania del año 2004. Se cuenta con 366 muestras y seis variables medidas: 1) temperatura media del aire (ºC), 2) presión barométrica media (hPa), 3) humedad relativa media (%), 4) cantidad de lluvia (mm), 5) radiación solar (horas/día), y 6) profundidad de la nieve (cm). Además, cada muestra fue etiquetada como día helado, día frío, día templado y día cálido, en las figuras 5 y 6 se presentan los resultados obtenidos al aplicar LLE a la base de datos de imágenes y a la base de datos meteorológica, respectivamente.

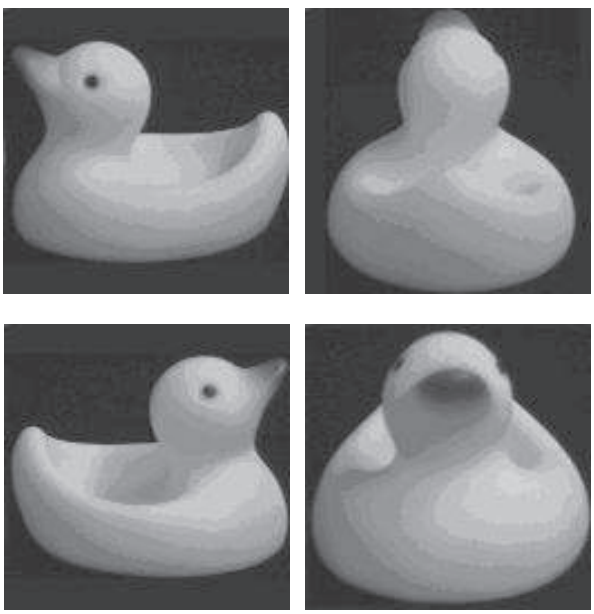


Figura 3 Pato de hule en 0º, 90º, 180º y 270º

Discusión y resultados

De las pruebas realizadas sobre las bases de datos artificiales es posible observar que el método empírico de regularización (α_1) produce mejores resultados de inmersión que los métodos automáticos de regularización (α_2 y α_3), lo cual se evidencia en las figuras 1 y 2. Sin embargo, este método requiere establecer de forma manual el valor de la variable Δ en la expresión (6), lo cual puede ser una tarea difícil cuando los datos poseen una dimensión de entrada mayor a 3, ya que no existe una referencia sobre los resultados de la inmersión. Esto se comprueba con los resultados obtenidos de la base de datos meteorológica, donde el método de regularización empírico presenta un desempeño inferior a los métodos de regularización automáticos

Figura 4 Etiquetas gráficas para base de datos de meteorología

El método de regularización propuesto en [6] (α_2), exhibe errores en las inmersiones resultantes, ya que las variedades utilizadas para realizar las pruebas presentan discontinuidades. En estos casos la media de los valores propios descartados calculada a través de (7) no es una técnica de regularización apropiada. Por otra parte, el método de regularización propuesto en este trabajo (α_3) fue efectivo al encontrar valores adecuados para el parámetro de regularización para los datos artificiales, obteniendo resultados correctos que corresponden con el desdoblamiento esperado. En el caso de la base meteorológica, α_3 exhibe el mejor desempeño de los tres métodos utilizados, de acuerdo a la figura 6 y a la medida de calidad propuesta en este trabajo (19).

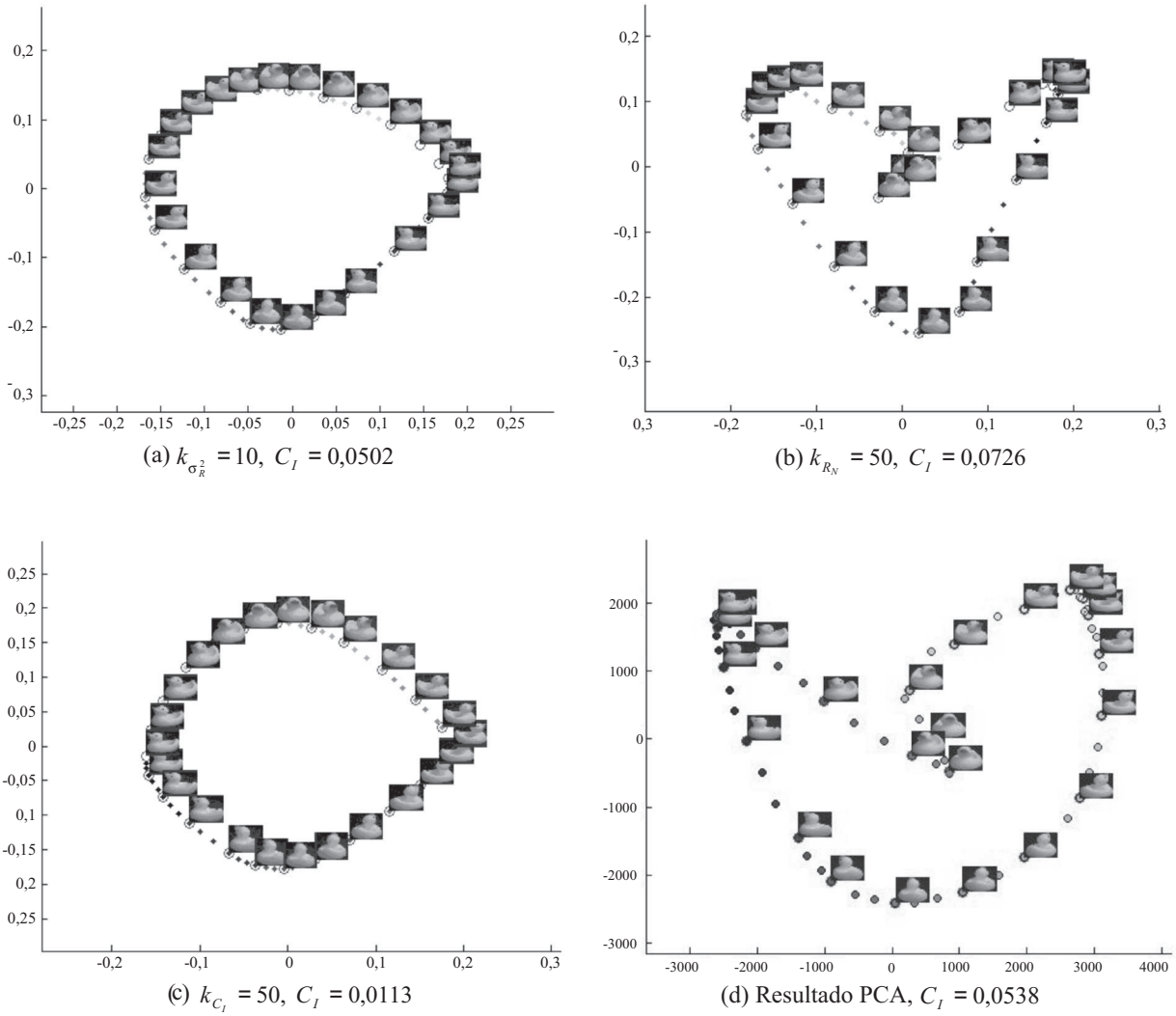
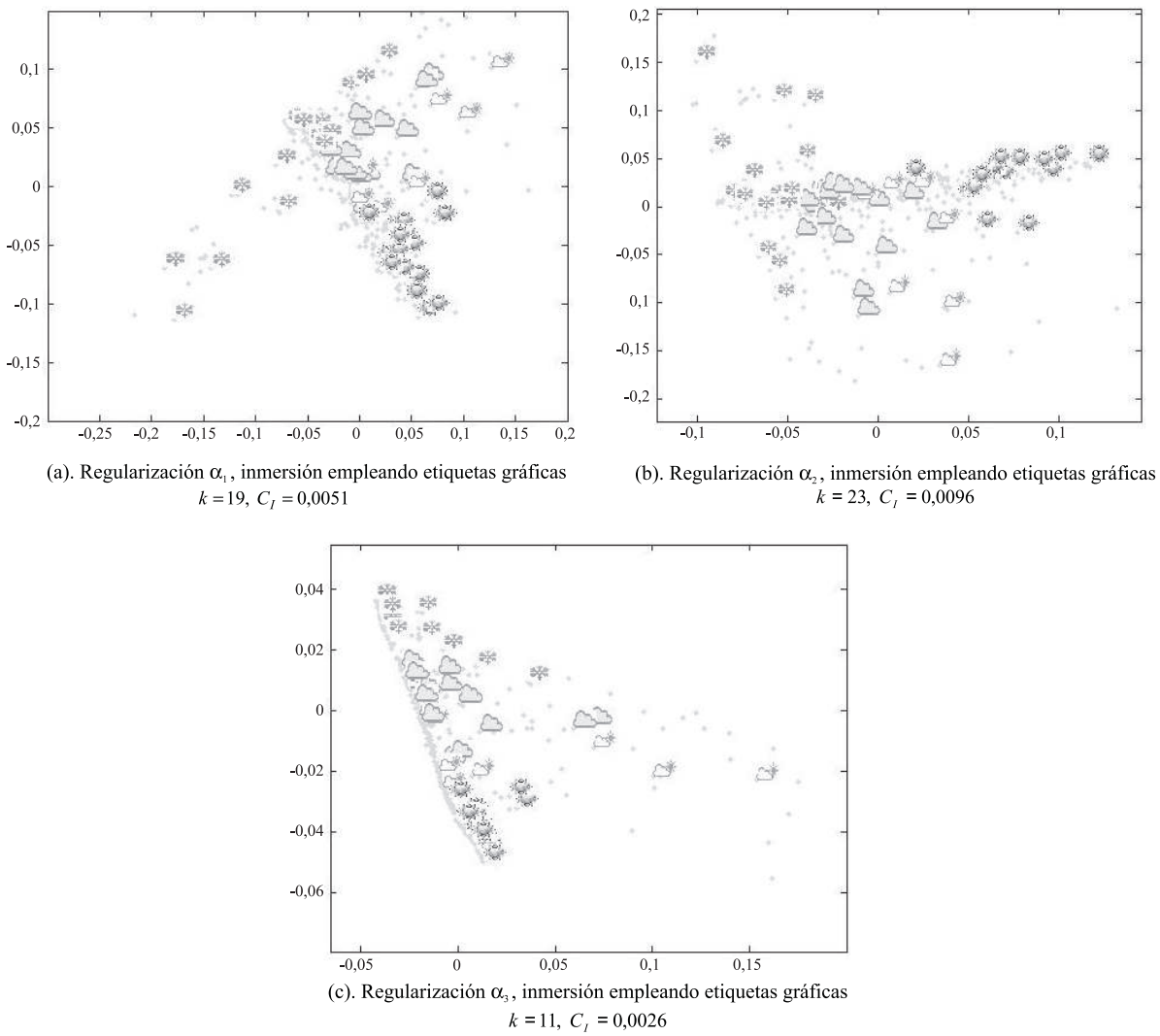


Figura 5 Resultados imágenes Pato de hule, figuras (a), (b), (c) LLE con *k* automático, (d) Resultado PCA

No obstante, es importante resaltar que el parámetro *k* también influye de gran manera en los resultados de LLE. Para las bases de datos artificiales, el método de la varianza residual (16) y el estadístico de procrustes (18) presentan traslapes en el espacio de salida, por lo cual se pierde la estructura global del espacio original. Esto se debe a que dichos métodos eligen un valor alto de *k*, y es claro que un número grande de vecinos implica una transformación lineal de los datos (similar a PCA). Dichos resultados pueden ser visualizados en las figuras 1 y 2 ((b), (c), (d), (f), (g), (i) y (j)). Por otra parte, con el método propuesto en este trabajo (20), se obtiene un valor adecuado de *k* para las dos bases de datos artificiales, dando como resultado inmersiones que corresponden con el desdoblamiento visual esperado de, logrando conservar la geometría local del espacio original y al mismo tiempo la estructura global de los datos, lo cual se evidencia en las figuras 1 y 2 ((e) y (k)). En las figuras 1 y 2 (h), es posible observar que el resultado del algoritmo es afectado de forma negativa al utilizar α_2, por tal motivo la calidad de la inmersión no es la apropiada.



(a). Regularización α_1, inmersión empleando etiquetas gráficas $k = 19,\ C_I = 0{,}0051$

(b). Regularización α_2, inmersión empleando etiquetas gráficas $k = 23,\ C_I = 0{,}0096$

(c). Regularización α_3, inmersión empleando etiquetas gráficas $k = 11,\ C_I = 0{,}0026$

Figura 6 Resultados LLE para base de datos de meteorología

Por otro lado, de la base de datos de imágenes, se observa que el método propuesto para determinar el numero de vecinos da como resultado una inmersión que permite identificar de forma clara la rotación del objeto (figura 5 (c)), validando así la eficacia de la metodología propuesta sobre datos reales.

Conclusiones

En este artículo se propone criterio para determinar de forma automática el valor del número de vecinos necesarios para una correcta reducción de dimensión, a través de una medida que permite cuantificar la calidad de la inmersión realizada por medio del algoritmo LLE. Este criterio fue comparado contra otros métodos presentados en la literatura. Los resultados obtenidos demuestran que el método propuesto es superior a los otros dos, porque considera tanto la preservación de la geometría local como la conservación de la estructura global de los datos originales, por lo cual se convierte en un criterio confiable para determinar adecuadamente el valor del parámetro *k*.

También se propone una metodología para calcular de forma automática el parámetro de regularización de LLE. Los resultados comprueban que el parámetro de regularización obtenido permite calcular una inmersión adecuada para las bases de datos utilizadas, presentando mejor desempeño que el método empírico en las bases de datos reales. Además al utilizar el método propuesto no es necesario que un experto sintonice el algoritmo, lo cual es una tarea compleja cuando se cuenta con datos en espacios de dimensión mayor a 3.

Agradecimientos

Esta investigación se llevó a cabo gracias a los recursos del proyecto "Representación y discriminación de datos funcionales empleando inmersiones localmente lineales" y a una beca para estudios de maestría financiada por la Universidad Nacional de Colombia. Además al proyecto "Desarrollo de un sistema piloto de mantenimiento predictivo en la línea de propulsión de las lanchas patrulleras de la armada nacional mediante el análisis de vibraciones mecánicas e imágenes termográficas", a una beca de doctorado otorgada por Colciencias y al proyecto DIMA No 20201005253.

Referencias

1. M. A. Carreira-Perpiñan. "A review of dimension reduction techniques". Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09. 1997. pp 1-69.
2. S. T. Roweis, L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding". *Science*. Vol. 290. 2000. pp. 2323-2326.
3. L. K. Saul, S. T. Roweis. "An introduction to locally linear embedding". *AT&T Labs and Gatsby Computational Neuroscience Unit*. Tech. Rep. 2000. pp 1-16.
4. L. K. Saul, S. T. Roweis. "Think globally, fit locally: Unsupervised learning of low dimensional manifolds". *Machine Learning Research*. Vol. 4. 2003. pp. 119-155.
5. M. Polito, P. Perona. "Grouping and dimensionality reduction by locally linear embedding". *NIPS.* Vol. 14. 2001. pp. 1255-1262.
6. D. de Ridder, R. P. W. Duin. *Locally linear embedding for classification*. Pattern Recongnition Group. Delft University of Technology. Netherlands. Tech. Rep. 2002. pp. 1-15.
7. P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM. Philadelphia. 1998.
8. Y. Goldberg, Y. Ritov. "Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms". *Machine learning*. Vol. 77. 2009. pp 1-25.
9. O. Kouropteva, O. Okun, M. Pietikäinen. *Selection of the optimal parameter value for the locally linear embedding algorithm*. The 1st ICFSKD. 2002. pp. 359-363.
10. S. A. Nene, S. K. Nayar, H. Murase. *Columbia object image library: Coil-100*. Department of Computer Science, Columbia University. NY.Tech. Rep. 1996. pp. 1-16.
11. A. Tank. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int. Jour. Climatology*. Vol. 22. 2002. pp. 1441-1453.