Automatic Choice of the Number of NearestNeighbors in Locally Linear Embedding

4 authors, including:

Genaro Daza-Santacoloma
Instituto de Epilepsia y P…
28 PUBLICATIONS **114** CITATIONS

SEE PROFILE

German Castellanos-Do…
National University of Co…
346 PUBLICATIONS **889** CITATIONS

SEE PROFILE

Automatic Choice of the Number of Nearest Neighbors in Locally Linear Embedding

Juliana Valencia-Aguirre, Andrés Álvarez-Mesa, Genaro Daza-Santacoloma, and Germán Castellanos-Domínguez

Control and Digital Signal Processing Group, Universidad Nacional de Colombia, Manizales, Colombia
{jvalenciaag,amalvarezme,gdazas,cgcastellanosd}@unal.edu.co

Abstract. Locally linear embedding (LLE) is a method for nonlinear dimensionality reduction, which calculates a low dimensional embedding with the property that nearby points in the high dimensional space remain nearby and similarly co-located with respect to one another in the low dimensional space [1]. LLE algorithm needs to set up a free parameter, the number of nearest neighbors k. This parameter has a strong influence in the transformation. In this paper is proposed a cost function that quantifies the quality of the embedding results and computes an appropriate k. Quality measure is tested on artificial and real-world data sets, which allow us to visually confirm whether the embedding was correctly calculated.

1 Introduction

In many pattern recognition problems the characterization stage generates a big amount of data. There are several important reasons for reducing the feature space dimensionality, such as, improve the classification performance, diminish irrelevant or redundancy information, find out underlying data structures, obtain a graphical data representation for visual analysis, etc [2]. Several techniques for dimensionality reduction have been developed, traditionally these techniques are linear [3] and they can not correctly discover underlying structures of data lie on nonlinear manifolds. In order to solve this trouble a nonlinear mapping method called locally linear embedding (LLE) was proposed in [4,5].

This method requires to manually set up three free parameters, the dimensionality of embedding space m, the regularization parameter α, and the number of nearest neighbors k for local analysis [6]. There are two previous approaches for choosing k. Kouropteva et. al [7] presented a hierarchical method for automatic selection of an optimal parameter value based on the minimization of the residual variance. Nonetheless, the residual variance can not quantify the local geometric structure of data. Besides, Goldberg and Ritov [8] display a novel measure based on Procrustes rotation that enables quantitative comparison of the output of manifold-based embedding algorithms, the measure also serves as a natural tool for choosing dimension-reduction parameters. The local procrustes

measure preserves local geometric structure but does not consider the global behavior of the manifold.

In this paper is proposed an automatic method for choosing the number of nearest neighbors, which is done by means of computing a cost function that quantifies the quality of embedding space. This function takes into account local and global geometry preservation. Proposed approach is experimentally verified on 2 artificial data sets and 1 real-world data set. Artificial data sets allow to visually confirm whether the embedding was correctly calculated, and real-world data set was used for visualization of multidimensional samples.

2 Locally Linear Embedding

Let $\mathbf{X}$ the input data $p \times n$ matrix, where the sample vectors $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ are had. Data live on or close to a non-linear manifold and that is well-sampled. Besides, each point and its neighbors lie on a locally linear patch. LLE algorithm has 3 steps. First, search the k nearest neighbors per point, as measured by Euclidean distance. If k is set too small, the mapping will not reflect any global properties; if it is too high, the mapping will lose its nonlinear character and behave like traditional PCA [6]. Second, each point is represented as a weighted linear combination of its neighbors [9], that is, we calculate weights $\mathbf{W}$ that minimize reconstruction error

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

subject to an sparseness constraint, $w_{ij} = 0$ if $\mathbf{x}_j$ is not $k-$neighbor of $\mathbf{x}_i$, and an invariance constraint $\sum_{j=1}^{n} w_{ij} = 1$. For a particular data point $\mathbf{x} \in \mathbb{R}^p$ and its k nearest neighbors $\boldsymbol{\eta}$. Let $\mathbf{G}$ the Gram $k \times k$ matrix, where its elements are $G_{jl} = \left\langle \left(\mathbf{x} - \boldsymbol{\eta}_j\right), \left(\mathbf{x} - \boldsymbol{\eta}_l\right)\right\rangle$, $j = 1, \ldots, k$; $l = 1, \ldots, k$. Then rewriting (1)

$$\varepsilon = \mathbf{w}^\top \mathbf{G} \mathbf{w} \quad \text{s.t.} \quad \sum_{j=1}^{n} w_j = 1. \tag{2}$$

The solution of (2) is obtained by solving an eigenvalue problem. Employing Lagrange theorem $\mathbf{w} = (1/2)\lambda\mathbf{G}^{-1}\mathbf{1}$, where $\lambda = 2/\left(\mathbf{1}^\top\mathbf{G}^{-1}\mathbf{1}\right)$. When Gram matrix $\mathbf{G}$ is singular (or close), the result of least squares problem for finding $\mathbf{w}$ does not have unique solution. So, it is necessary to regularize $\mathbf{G}$ before finding $\mathbf{w}$. In [1,5] is proposed to calculate the regularization of $\mathbf{G}$ as $G_{jl} \leftarrow G_{jl} + \alpha$ where $\alpha = \delta_{jl}\left(\Delta^2/k\right)\text{tr}\left(\mathbf{G}\right)$, being $\delta_{jl} = 1$ if $j = l$ and 0 in other case, $\Delta^2 \ll 1$. However, Δ must be empirically tuned, in [1] is advised to employ $\Delta = 0.1$.

In third step low dimensional embedding is calculated. Using $\mathbf{W}$, the low dimensional output $\mathbf{Y}$ is found by minimizing (3)

$$\varPhi\left(\mathbf{Y}\right) = \sum_{i=1}^{n} \|\mathbf{y}_i - \sum_{j=1}^{n} w_{ij}\mathbf{y}_j\|^2, \tag{3}$$

subject to $\sum_{i=1}^{n} \mathbf{y}_i = \mathbf{0}$ and $\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^{\top}/n = \mathbf{I}_{m \times m}$, where $\mathbf{Y}$ is the embedding data $n \times m$ matrix (being $m \leq p$), and $\mathbf{y}_i \in \mathbb{R}^m$ is the output sample vector.

Let $\mathbf{M} = \left(\mathbf{I}_{n \times n} - \mathbf{W}^{\top}\right)\left(\mathbf{I}_{n \times n} - \mathbf{W}\right)$, and rewriting (3) to find $\mathbf{Y}$,

$$\Phi\left(\mathbf{Y}\right) = \text{tr}\left(\mathbf{Y}^{\top}\mathbf{M}\mathbf{Y}\right) \quad \text{s.t.} \quad \begin{cases} \mathbf{1}_{1 \times n}\mathbf{Y} = \mathbf{0}_{1 \times n} \\ \frac{1}{n}\mathbf{Y}^{\top}\mathbf{Y} = \mathbf{I}_{m \times m} \end{cases} \tag{4}$$

it is possible to calculate $m + 1$ eigenvectors of $\mathbf{M}$, which are associated to $m + 1$ smallest eigenvalues. First eigenvector is the unit vector with all equal components, which is discarded. The remaining m eigenvectors constitute the m embedding coordinates found by LLE.

3 Measure of Embedding Quality

When dimensionality reduction technique is computed is necessary to establish a criteria for knowing if its results are adequate. In LLE is searched a transformation that preserves the local data geometry and global manifold properties.

The quality of an output embedding could be judged based on a comparison to the structure of the original manifold. However, in the general case, the manifold structure is not given, and it is difficult to estimate accurately. As such ideal measures of quality cannot be used in the general case, an alternate quantitative measure is required [8].

In [7], the residual variance is employed as a quantitative measure of the embedding results. It is defined as

$$\sigma_R^2(D_{\mathbf{X}}, D_{\mathbf{Y}}) = 1 - \rho_{D_{\mathbf{X}} D_{\mathbf{Y}}}^2, \tag{5}$$

where ρ^2 is the standard linear correlation coefficient, taken over all entries of $D_{\mathbf{X}}$ and $D_{\mathbf{Y}}$; $D_{\mathbf{X}}$ and $D_{\mathbf{Y}}$ are the matrices for Euclidean distances in $\mathbf{X}$ and $\mathbf{Y}$, respectively. $D_{\mathbf{Y}}$ depends on the number of neighbors selected k. According to [7], the lowest residual variance corresponds to the best high-dimensional data representation in the embedded space. Hence, the number of neighbors can be computes as

$$k_{\sigma_R^2} = \arg\min_{k}(\sigma_R^2(D_{\mathbf{X}}, D_{\mathbf{Y}})). \tag{6}$$

On the other hand, in [8] is proposed to compare a neighborhood on the manifold and its embedding using the Procrustes statistic as a measure for qualifying the transformation. This measures the distance between two configurations of points and is defined as $P\left(\mathbf{X}, \mathbf{Y}\right) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_i - \mathbf{b}\|^2$, being $\mathbf{A}^{\top}\mathbf{A} = \mathbf{I}$ and $\mathbf{b} \in \mathbb{R}^m$. The rotation matrix $\mathbf{A}$ can be computed from $\mathbf{Z} = \mathbf{X}^{\top}\mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{I} - \frac{1}{k}\mathbf{1}\mathbf{1}^{\top}$, $\mathbf{1}$ is a $n \times 1$ column vector, and $\mathbf{H}$ is the centering matrix. Let $\mathbf{U}\mathbf{L}\mathbf{V}^{\top}$ be the singular-value decomposition of $\mathbf{Z}$, then $\mathbf{A} = \mathbf{U}\mathbf{V}^{\top}$, the translation vector $\mathbf{b} = \overline{\mathbf{x}} - \mathbf{A}\overline{\mathbf{y}}$, where $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are the sample means of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let $\|\cdot\|_F$ the Frobenius norm, so $P\left(\mathbf{X}, \mathbf{Y}\right) = \|\mathbf{H}(\mathbf{X} - \mathbf{Y}\mathbf{A}^{\top})\|_F^2$.

In order to define how well an embedding preserves the local neighborhoods using the Procrustes statistic $P_L(\mathbf{X}_i, \mathbf{Y}_i)$ of each neighborhood-embedding pair

$(\mathbf{X}_i, \mathbf{Y}_i)$. $P_L(\mathbf{X}_i, \mathbf{Y}_i)$ estimates the relation between the entire input neighborhood and its embedding as one entity, instead of comparing angles and distances within the neighborhood with those within its embedding. A global embedding that preserves the local structure can be found by minimizing the sum of the Procrustes statistics of all neighborhood-embedding pairs [8], taking into account an scaling normalization, which solves the problem of increased weighting for larger neighborhoods, so

$$R_N(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} P_L(\mathbf{X}_i, \mathbf{Y}_i) / \|\mathbf{H_L}\mathbf{X}_i\|_F^2, \tag{7}$$

where $\mathbf{H_L} = \mathbf{I} - \frac{1}{k}\mathbf{1}\mathbf{1}^\top$; $\mathbf{1}$ is a $k \times 1$ vector. Therefore, the number of nearest neighbors can be calculated as

$$k_{R_N} = \arg\min_{k}(R_N(\mathbf{X}, \mathbf{Y})). \tag{8}$$

In this work, we propose an alternative measure for quantifying the embedding quality. This measure attempts to preserve the local geometry and the neighborhood co-location, identifying possible overlaps on the low dimensional space. And it is defined as

$$C_I(\mathbf{X}, \mathbf{Y}) =$$

$$\frac{1}{2n} \sum_{i=1}^{n} \left\{ \frac{1}{k} \sum_{j=1}^{k} \left(D_{(\mathbf{x}_i, \boldsymbol{\eta}_j)} - D_{(\mathbf{y}_i, \boldsymbol{\phi}_j)} \right)^2 + \frac{1}{k_n} \sum_{j=1}^{k_n} \left(D_{(\mathbf{x}_i, \boldsymbol{\theta}_j)} - D_{(\mathbf{y}_i, \boldsymbol{\gamma}_j)} \right)^2 \right\}, \tag{9}$$

where D is an standardized Euclidean distance to obtain a maximum value equal to one. For example $D_{(\mathbf{x}_i, \boldsymbol{\eta}_j)}$ is the distance calculated between the observation $\mathbf{x}_i$ and each one of its k neighbors on the input space.

Once the embedding, for each point $\mathbf{y}_i \in \mathbb{R}^m$ a set $\boldsymbol{\beta}$ of k nearest neighbors is calculated, and the projection $\boldsymbol{\phi}$ of $\boldsymbol{\eta}$ is found. The neighbors computed in $\boldsymbol{\beta}$ that are not neighbors in $\boldsymbol{\eta}$ conform a new set $\boldsymbol{\gamma}$, that is $\boldsymbol{\gamma} = \boldsymbol{\beta} - (\boldsymbol{\beta} \cap \boldsymbol{\phi})$. The size of $\boldsymbol{\gamma}$ is k_n. Besides, the projections of the elements of $\boldsymbol{\gamma}$ in $\mathbf{X}$ conform the set $\boldsymbol{\theta}$ of k_n neighbors. In an ideal embedding $C_I(\cdot) = 0$.

The first term in (9) quantifies the local geometry preservation. The k nearest neighbors of $\mathbf{x}_i$ chosen on the high dimensional space $\mathbf{X}$ are compared against their representations in the embedded space $\mathbf{Y}$. The second term computes the error produced by possible overlaps in the embedded results, which frequently occurs when the number of neighbors is strongly increased, and global properties of the manifold are lost. The number of nearest neighbors can be found as

$$k_{C_I} = \arg\min_{k}(C_I(\mathbf{X}, \mathbf{Y})). \tag{10}$$

4   Experimental Background

4.1   Tests on Artificial Data Sets

Two different manifold are tested, which allow to visually confirm whether the embedding was correctly calculated. The Swiss Roll with Hole data set with

(a) Swiss Roll with Hole   (b) Fishbowl
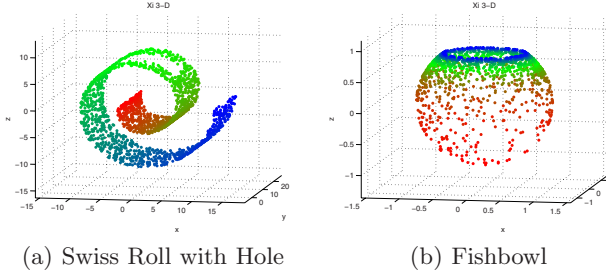
Fig. 1. Artificial Data Sets

2000 samples (Fig. 1(a)) and the Fishbowl data set with uniform distribution in embedding space and 1500 samples (Fig. 1(b)).

In order to quantify the embedding quality and find the number of nearest neighbors needed for a faithful embedding, LLE is computed by varying k in the subset $k \in \{3, 4, 5, \ldots, 250\}$, fitting the dimensionality of the embedded space to $m = 2$. The embedding quality is computed according to (5), (7) and (9). The number of nearest neighbors is found by means of (6), (8) and (10).
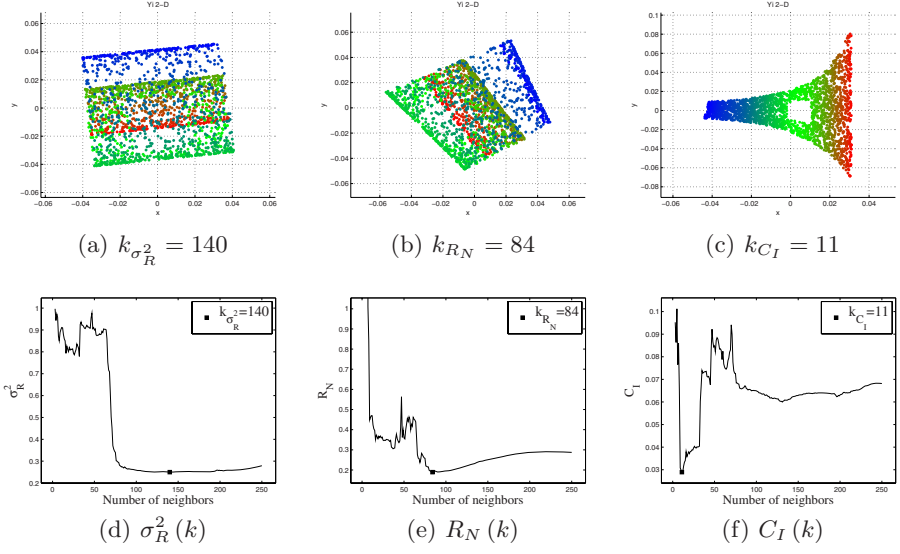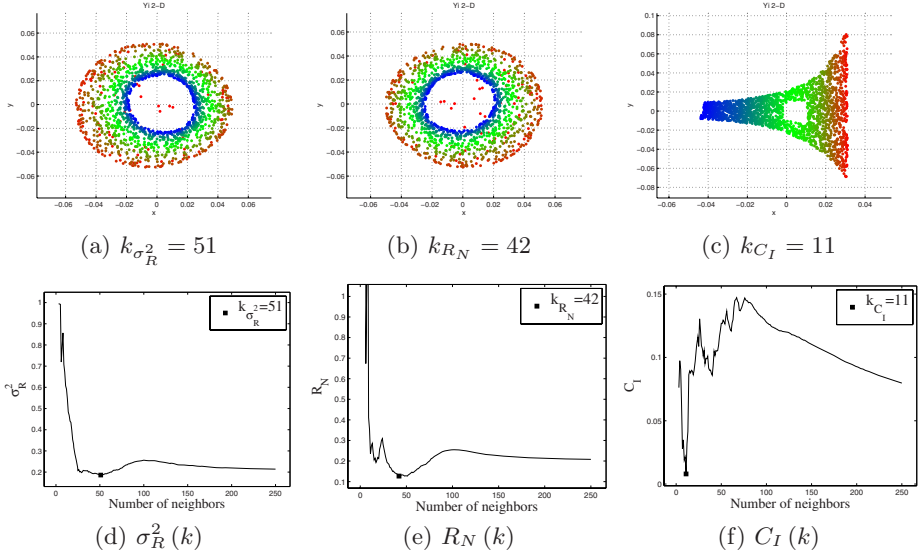
In Figures 2(a), 2(b), and 2(c) are presented the embedding results for the Swiss Roll with Hole data set using an specific number of neighbors in accordance with each one of the criteria above pointed out. Similarly, in Figures 3(a), 3(b), and 3(c) the embedding results for the Fishbowl data set are displayed. For these artificial data sets only our approach (10) finds appropriate embedding results preserving local and global structure. In the case of the Swiss Roll with Hole, the criteria (6) and (8) produce overlapped embeddings. Besides, in Fishbowl, the unfolding results obtained by means of (6) and (8) are wrong, those are similar to PCA and then local structure is lost.

On the other hand, Figures 2(d), 2(e), 2(f), and 3(d), 3(e), 3(f), show curves of the cost function value versus the number of neighbors. The full fill square on the curves is the global minimum of the function and corresponds to the number of nearest neighbors chosen for the embedding.
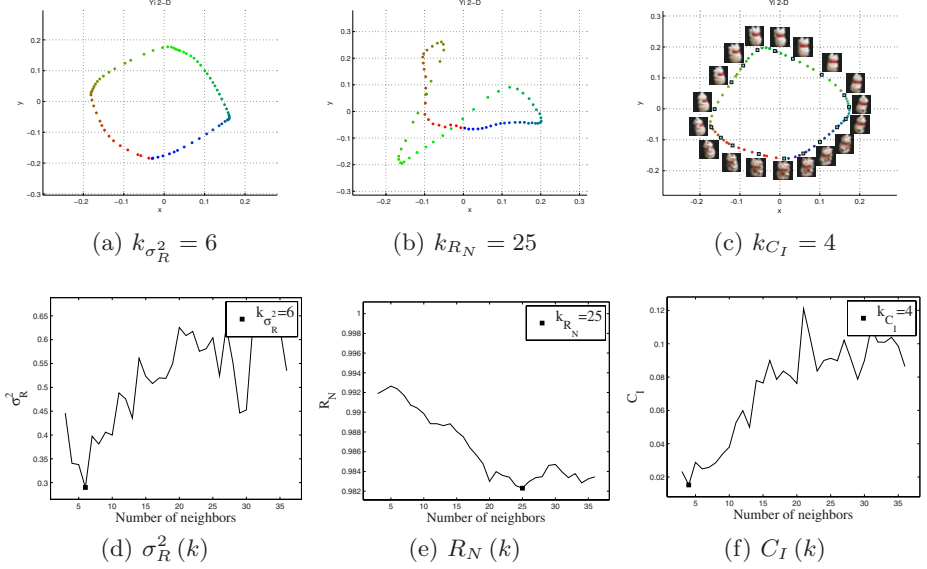
4.2   Tests on Real-World Data Sets

We use Maneki Neko pictures, which is in the Columbia Object Image Library (COIL-100) [10]. There are 72 RGB-color images for this object in PNG format. Pictures are taken while the object is rotated 360 degrees in intervals of 5 degrees. The image size is 128×128. In Fig. 4 are shown some examples. We transform these color images to gray scale, next the images were subsampled to 64×64 pixels. Then, we have input space of dimension $p = 8192$ and $n = 72$.

In order to quantify the embedding quality and find the number of neighbors needed for a faithful embedding, LLE is computed by varying k in the subset $k \in \{3, 4, 5, \ldots, 36\}$, fitting the dimensionality of the embedded space to $m = 2$. In Fig. 5 are shown the embedding results for the Maneki Neko data set and

(a) $k_{\sigma_R^2} = 140$　　(b) $k_{R_N} = 84$　　(c) $k_{C_I} = 11$

(d) $\sigma_R^2(k)$　　(e) $R_N(k)$　　(f) $C_I(k)$

Fig. 2. Results for the Swiss Roll with Hole Data Set

(a) $k_{\sigma_R^2} = 51$　　(b) $k_{R_N} = 42$　　(c) $k_{C_I} = 11$

(d) $\sigma_R^2(k)$　　(e) $R_N(k)$　　(f) $C_I(k)$

Fig. 3. Results for the Fishbowl Data Set

its corresponding cost function curves, which allow to establish the number of neighbors employed in the transformation, according to (6), (8) and (10).

(a) 0°   (b) 45°   (c) 90°

Fig. 4. Examples from Maneki Neko Data Set

(a) $k_{\sigma_R^2} = 6$ (b) $k_{R_N} = 25$ (c) $k_{C_I} = 4$

(d) $\sigma_R^2(k)$ (e) $R_N(k)$ (f) $C_I(k)$

Fig. 5. Results of Maneki Neko Data Set

5 Discussion

From the obtained results on artificial data sets (Fig. 2(d) and 3(d)) employing the cost function (6) proposed in [7], as a global trend, it is possible to notice if the size of the neighborhood is augmented, the value of σ_R^2 is diminished. Because a transformation employing a large number of neighbors results in a linear transformation and the residual variance does not quantify the local geometric structure, then this measure can not identify a suitable number of neighbors k. Figures 2(a) and 3(a) show some examples of this situation. In this case, data in low-dimensional space are overlapped and the cost function (6) does not detect it. Nevertheless Fig. 5(d) does not display the trend above pointed out and allows to calculate an appropriate embedding (Fig. 5(a)). The inconsistent results obtained by using the residual variance make of it an unreliable measure. In [7] the ambiguous results are attributed to the fact that Euclidean distance becomes an unreliable indicator for proximity.

On the other hand, the measure proposed in [8] takes into account the local geometric structure but does not consider the global behavior of the manifold.

Then, far neighborhoods can be overlapped in the low-dimensional space and this measure shall not detect it, which can be seen in Fig. 2(b), 3(b), and 5(b). The measure of embedding quality presented here (10) computes a suitable number of neighbors for both artificial and real-world manifolds. Besides, the embedding results calculated using this criterion are in accordance to the expected visual unfolding (Fig 2(c), 3(c), and 5(c)). The proposed measure preserves the local geometry of data and the global behavior of the manifold.

6   Conclusion

In this paper a new measure for quantifying the embedding quality using LLE is proposed. This measure is employed as a criterion for choosing automatically the number of nearest neighbors needed for the transformation. We compare the new cost function against two methods presented in the literature. The best embedding results (visually confirmed) were obtained using the approach exposed, because it preserves the local geometry of data and the global behavior of the manifold.

References

1. Saul, L.K., Roweis, S.T.: Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Machine Learning Research 4 (2003)
2. Carreira-Perpiñan, M.A.: A review of dimension reduction techniques (1997)
3. Webb, A.R.: Statistical Pattern Recognition, 2nd edn. Wiley, USA (2002)
4. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
5. Saul, L.K., Roweis, S.T.: An introduction to locally linear embedding. AT&T Labs and Gatsby Computational Neuroscience Unit, Tech. Rep. (2000)
6. de Ridder, D., Duin, R.P.W.: Locally linear embedding for classification. Delft University of Technology, The Netherlands, Tech. Rep. (2002)
7. Kouropteva, O., Okun, O., Pietikäinen, M.: Selection of the optimal parameter value for the locally linear embedding algorithm. In: The 1st International Conference on Fuzzy Systems and Knowledge Discovery (2002)
8. Goldberg, Y., Ritov, Y.: Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. Machine learning (2009)
9. Polito, M., Perona, P.: Grouping and dimensionality reduction by locally linear embedding. In: NIPS (2001)
10. Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library: Coil-100. Columbia University, New York, Tech. Rep. (1996)