An improved locally linear embedding method for feature extraction

Zhang Wei\textsuperscript{1,a} \quad Zhou Weijia\textsuperscript{1,b}

\textsuperscript{1}Shenyang Institute of Automation Chinese Academy of Sciences, State key Laboratory of Robotics, Nanta Street 114#, Shenyang, Liaoning Province, China, 110016E
\textsuperscript{a}zhangwei@sia.cn, \textsuperscript{b}zwj@sia.cn

Key words: Locally Linear Embedding; tangent space distance; feature extraction

Abstract. In this work, a feature extraction approach based on improved Locally Linear Embedding (LLE) is proposed. In the algorithm, tangent space distance is introduced to LLE, which overcomes the shortcoming of original LLE method based on Euclidean distance. It can satisfy the requirement of locally linear much better and so can express the I/O mapping quality better than classical method. Simulation results are given to demonstrate the effectiveness of the improved LLE method.

1 Introduction

To enhance the autonomous capability, systems are equipped with more and more sensors to detect operating status. Therefore, to deal with these massive data the feature extraction becomes extremely critical in order to keep the data interested and reduce its dimension.

By far, the most frequently-used method is Principle Component Analysis (PCA), which was proposed by Pearson in 1901\cite{1}, when he was studying regression analysis, and was mathematically developed by Hotelling in 1933\cite{2}. PCA has been widely used to monitor the industrial processes with multiple variables and diagnose system fault\cite{3,4}. But conventional PCA cannot reveal the exact correlations among the variables of dynamic processes. For this reason, dynamic PCA has been proposed, which takes serial correlation into account by augmenting the observation vector with lagged variables\cite{5,6}. However, all these methods were developed to target linear data systems. When data system is high-dimensional and severely non-linear, the method becomes ineffective. Therefore, in recent years, some new methods were proposed for non-linear systems. Kernel principle component analysis (KPCA) is derived from PCA, which is a widely used nonlinear feature extraction method\cite{7,8}. But it can not reveal the intrinsic topological of the data, and it is very hard to find a suitable kernel function, which is very important for the approach.

In 2000, Roweis proposed a manifold learning algorithm called Local Linear Embedding (LLE), which is an unsupervised non-linear technique that analyzes the high-dimensional data sets and reduces their dimensionalities with preserved local topology\cite{9}. Today, LLE has been widely used in cluster analysis, image processing, biological informatics, etc\cite{10,11}. But the original LLE method usually uses the Euclidean distance to compute neighbors, which can not express the locally linear very well. So in this work, tangent space distance is introduced to LLE algorithm which overcomes the shortcoming.

2 Improved locally linear embedding algorithm

The basic concept of LLE is to find a weight vector between a sample and its neighbors, and to keep this relationship in a feature space\cite{9}. It assumes that even if the manifold embedded in a
high-dimensional space is non linear, it still can be considered locally linear if each data point and its neighbors lie on or close to a locally linear patch of the manifold, i.e., the manifold can be covered with a set of locally linear patches which, when analyzed together, can yield information about the global geometry of the manifold. The weight vector expressing the intrinsic geometrical properties of the local patch can be obtained in three steps: (a) to find the neighbors of every sample in the high-dimensional space, (b) to obtain the reconstruction weight and a sparse matrix of the weight vectors, and (c) to compute the low dimensional embedding -- the bottom nonzero eigenvectors of the sparse matrix are the low dimensional embeddings of high dimensional samples.

It can be seen from the definition of LLE that the point and its neighbors must lie on or close to a locally linear patch of the manifold. Usually the correlation of data is computed after neighbors have been decided in Euclidean distance, but sometimes the nearest neighbors in Euclidean distance do not lie on the approximate linear curved face, as shown in Fig.1,

![Fig 1](image)

**Fig 1** Euclidean distance and tangent space distance from neighbor to point

Where $X_0$ is the point and lies on a curved face, $X_1$ and $X_2$ are its neighbors. $TV_1$ and $TV_2$ are tangent vectors of $X_0$, and construct the tangent space. $d_1$, $d_2$ are Euclidean distance from neighbors to $X_0$, and $D_1, D_2$ are distance from neighbors to tangent space of $X_0$. According to the definition of LLE we know that shorter $D$ means better of the locally linear quality of point and its neighbors. From Fig. 1 it can be seen that although $X_1$ is the nearest neighbor of point $X_0$ in Euclidean distances, but in fact $X_2$ is more suitable to be the nearest neighbor in locally linear sense. So we introduce tangent space into the algorithm. Firstly, tangent space of every point was determined, and then distances between neighbors and the tangent spaces were computed. And so nearest neighbors who can satisfy the hypothesis of a locally linear patch better can be achieved.

Suppose $F$ is a $d$ dimensional manifold in $m$ dimensional space mapped by an unknown function $f(\tau), \tau \in \mathbb{R}^d$, and date $X = (x_1, x_2, \cdots, x_n), x_i \in \mathbb{R}^m$ is the image of the unknown function, $x_i = f(\tau_i), i = 1, 2, \cdots, n$. To get the tangent space of $x_i$, it is equivalent to get the differential of function $f(\tau)$ at $\tau_i$. Suppose $f$ is smooth, then the first Taylor expansion at $\tau$ can be described as follow,

$$f(\tau) = f(\bar{\tau}) + J_f(\bar{\tau}) \cdot (\tau - \bar{\tau}) + O(\|\tau - \bar{\tau}\|^2),$$

where $J_f(\tau) \in \mathbb{R}^{m \times d}$ is the Jacobi matrix of $f$ at $\tau$. 


Record $f(r) = \begin{pmatrix} f_1(r) \\ \vdots \\ f_m(r) \end{pmatrix}$, so $J_f(r) = \begin{pmatrix} \frac{\partial f_1}{\partial r_1} & \cdots & \frac{\partial f_1}{\partial r_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial r_1} & \cdots & \frac{\partial f_m}{\partial r_d} \end{pmatrix}$, and the tangent space $\Gamma_r$ of $f$ at $r$ is expanded by $d$ column vectors of $J_f(r)$, $\Gamma_r = \text{span}(J_f(r))$. Because function $f$ is unknown, so $J_f(r)$ can not be computed directly. It has been known that it is a $d$ dimensional space, so $J_f(r) \cdot (\tau - r) = Q_r \theta^*_\tau$ can be achieved if there is a set of orthogonal basis of $J_f(r)$. Where $\theta^*_\tau$ is the coordinate of $\tau - r$ at tangent space corresponding to $Q_r$. Because $f(\tau) - f(r) \approx J_f(r) \cdot (\tau - r)$, so $f(\tau) - f(r) \approx Q_r \theta^*_\tau$, that is $f(\tau) = Q_r \theta^*_\tau + f(r)$.

For a certain point $x_i$, its neighbors are $x_{ij} \approx x_i + Q_r \theta_i$. Construct following optimum function,

$$
\min_{x_i, Q_r, \theta} \sum_{j=1}^{k} \| x_{ij} - (x_i^* + Q_r \theta_j) \|^2 = \min_{x_i, Q_r, \theta} \| X_i - (x_i^* e^T + Q_r \theta) \|^2.
$$

Where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$, $Q_r \in \mathbb{R}^{d \times d}$, $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik}) \in \mathbb{R}^{d \times k}$, so $Q_r$ can be seen as $d$ approximate orthogonal basis of tangent space of $x_i$. The question could be resolved by singular value decomposition (SVD) to $X_i (I - \frac{1}{k} e e^T)$, where $Q_r$ is the $d$ eigenvectors corresponding to $d$ maximum eigenvalues of $X_i (I - \frac{1}{k} e e^T)$.

To evaluate the performance of dimension reduction, some researchers proposed residual variance according to Input/Output mapping quality, and that is the description effectiveness of the original data in higher dimensional space. It is defined as $1 - \rho^2_{D_x, D_y}$, where $\rho$ is the standard linear correlation coefficient, taken over all entries of $D_x$ and $D_y$, where $D_x$ and $D_y$ are the matrices of Euclidean distances (between pairs of points) in the high-dimensional and corresponding low-dimensional spaces, respectively. According to the definition, it can be seen that the lower the residual variance is, the better high-dimensional data is represented in the embedded space.

S-curve dataset is uniform sampled from noiseless three dimensional S-curve curved face[12]. In this work it is utilized to test the effectiveness of neighbor selection algorithm based on tangent space distance and compared with Euclidean distance.

There were 30 neighbors in the simulation, and the nearest 30 neighbors of the first point was shown as Fig. 2. Where the yellow points are neighbors in tangent space distance and green stars are neighbors in Euclidean distance. The red line is the tangent space of the first point. According to the character of S-curve, it can be seen that yellow points are distributed in tangent space of the point, and green points are distribute in curve face in relative sense.
To test the performance of the algorithm, residual variances were computed under 1 to 50 neighbors, and the result was shown in Fig.3.

It can be seen from Fig.3 that when there are few neighbors, the locally structure could not be expressed well, so the residual variance is much bigger. With more neighbors, both performances of two methods increase, but with neighbors keep increased, the method based on Euclidean distance could not guarantee the local linear characteristic, so the result was not good as in tangent space distance method obviously.

3. Experiment results

To evaluate the feature extraction performance of the algorithm, an example is presented and the results were compared with that using the KPCA and LLE approach.

Iris is a database contains 150 informations of Setosa, Versicolour and Virginica. In which, every style contains 50 informations.

The kernel of KPCA is radial basis kernel function, \( k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \), where \( 2\sigma^2 = 10 \). The projection result and which based on LLE and improved LLE are shown in figure 4,
According to separability criterion, a good feature extraction algorithm should have small inner-class distance and big inter-class distance after projection. So it can be concluded from Fig.4 that the approach based on improved LLE method has much better performance than KPCA and LLE.

4. Conclusion

According to the limitation of LLE algorithm based on Euclidean distance, the concept of tangent space distance is introduced to the algorithm, which can satisfy the requirement of locally linear much better than Euclidean distance, and so can express the I/O mapping quality better. Experiment result certifies the effectiveness of the improved LLE method, and so provides a new nonlinear feature extraction method.

References


Locally Linear Embedding for dimensionality reduction in QSAR[J]. Journal of
