# Improved Manifold Learning Algorithm for Data Dimension Reduction Based on KNN

**Liming Liang, Falu Weng, Zhaoyang Chen and Zhen Zhong**

School of Electric Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, 341000, P.R. of China
{495762632; 112983003; 16552303; 704674516}@qq.com

**Abstract -** In this paper, a new multi-manifold learning algorithm based on KNN algorithm is proposed in order to provide manifold learning model automatic parameters selection strategy. Basic ideas for such a algorithm is constructing a weighted norm as the variable of the intrinsic low dimensions expression function, and then optimizing the function's variables and getting a automatic selection of the size of the intrinsic low dimensions and the neighborhood in the manifold learning algorithm model. After a series of numerical experiments on simulated and experimental, results proves the feasibility and effectiveness of the algorithm.

**Index Terms** - Manifold learning, weighted norm, dimensionality reduction

## 1. Introduction

With the rapid development of information access and network technology, high dimensional data were generated so rapidly in those fields from scientific research, engineering application to social life, while these data often present characteristics like mass, high dimension, nonlinear, and contain the rules and knowledge that are difficult to directly observe. The characteristics of high-dimensional data have brought many problems, such as huge calculation and low efficiency problem caused by mass properties, 'the curse of dimensionality' caused by high dimensional feature and linear model failure problem caused by nonlinear characteristic. It is a difficult and hot topic to dig out the inner rules of these data effectively while keep the data information complete enough at the same time [1-2].

To analyze data based on the inner dimension of the data distribution is an important research direction on machine learning and multivariate data analysis. Many kinds of manifold learning algorithm emerged in recent years, such as isometric mapping (ISOMAP), embedded local linear method (LLE), local tangent space alignment (LTSA), etc, which call many scholars' attention, because of its high effect in restoring potential geometric structure of high-dimensional data sets. Manifold learning assumes that the data is uniform sampling in a low dimensional manifold while in the high dimensional Euclidean space, i.e. the points in the high-dimensional observation space is a combination of independent variables in the observation space expanded to a manifold. If the curly manifold in observation space is flattened effectively or the main variables of the internal is found out, then a low-dimensional manifold in the high-dimensional space can be found, the corresponding embedded map can be calculated, and dimensionality reduction and data visualization can be achieved in the end. This means that the manifold learning is better than traditional dimensionality reduction methods which can reflect the nature of things and more beneficial to the understanding of the data and further processing. However, in order to apply the manifold learning algorithm successfully, firstly the model parameters of the manifold learning need to be determined (i.e. the size of the intrinsic low dimensions and the neighborhood). Now these two parameters selection method have no uniform standards or methods. However the method how to choose the size of the intrinsic low dimensions $d$ and the neighborhood $k$ has a great influence to the application of algorithm , such as the size of the neighborhood is too large, linear structures consisting of the data and its neighborhood can seriously deviated from the essence of the manifold surface [3-5]. These are about four kinds of manifold learning model parameter selection methods: K-Nearest Neighbor (KNN), entropy estimation, the data's fractal dimension, and maximum likelihood estimate selection strategy used for dimensionality reduction.

Based on the KNN selection strategy, this is the information of the intrinsic low dimensions contained in the linear relationship between the distance of nearest neighbor and the neighborhood size, was used to construct the data's intrinsic low dimensions as the setting strategy. The advantage of this method lies in its computational efficiency and has the probability theory. Due to different size of neighborhood $k$ lead to different dimension in manifold data, the accuracy of the calculation results dependent on the primary choice of the neighborhood size parameter, and unrealized for automatic selection of the optimal neighborhood size $k$ [6-7].

The data's fractal dimension selection strategy is to use the self similar fractal structure characteristics of the data fractal structure approximately estimating its intrinsic low dimensions [8]. Within a certain range, the size of the neighborhood approximate logarithmic linear relationship between the corresponding fractal capacity and its slope corresponds to the fractal dimension under corresponding observation scale. When manifold intrinsic low dimension $d$ was fixed, then the linear fitting method can used to obtain appropriate dimension $d$ . Because different observation scales of data shows different fractal dimension of the data, the premise of method's validity is proper scale range should constructed to fitting the fractal dimension in a reasonable manner. Entropy estimate selection strategy is based on function relationship between topological entropy of the neighborhood graph and the intrinsic low dimensions in the entropy theory [9]. Its main

problem is that approximate estimate to geodesic distance matrix repeatedly. High computational complexity and the calculation results depend heavily on neighborhood size preselection. But its computation also depends on the primary neighborhood size and the distribution of original data has stronger statistical hypothesis. As a result, the current parameter selection method does not have comprehensive intrinsic low dimensions and neighborhood size automatic selection strategy. The maximum likelihood estimation selection strategy is to estimate the intrinsic low dimension by maximizing likelihood function, after the likelihood function of distance of neighborhood structured. The intrinsic low dimension estimation scheme in statistical framework by this method has high computing efficiency [10]. This paper proposes a manifold learning algorithm to do automatic selection of intrinsic low dimensions and neighborhood size. Its basic idea is to construct a function expression took the weighted norm of the k nearest distance of nearest neighbor as the variable of the intrinsic low dimensions, and then get an automatic selection of intrinsic low dimensions and neighborhood size for the model of manifold learning algorithm through the optimization function's variables. In the end, several numerical experiments are given to show the feasibility and effectiveness of the algorithms.

## 2. Manifold Learning Algorithm Based on Automatic Selection

### A. *Domain knowledge*

A given manifold dataset with independent distribution $X=\{x_i\}_{i=1}^{l}\subset R^n$ also has density function $p(\bullet)$, its density function is estimated by：

$$\hat{p}(x)=\frac{P\left(x\in B(x,r)\right)}{V_d r^d},\ x\in\Omega_n \tag{1}$$

$\Omega_n\subset R^n$ is manifold space with intrinsic low dimensions $d\ (d<n)$, $p(\bullet)$ is the density function of manifold space, $x$ is the center and $r$ is the radius in the open loop sphere area $B(x,r)$, $V_d$ is $d$ dimensions unit sphere volume, i.e.:

$$V_d=\frac{(\pi)^{d/2}}{\Gamma(d/2+1)} \tag{2}$$

$\Gamma(t)$ as the Gamma function, $\Gamma(t)=\int_0^{\infty}x^{t-1}e^{-x}dx$. When $l$ large enough, $r$ small enough and $p(\bullet)$ continuously on point $x$, then its can be proved that the estimated density $p(\bullet)$ convergence to the real density.

Generate limited samples $X=\{x_i\}_{i=1}^{l}$ in a given probability distribution $p(\bullet)$, usually adopt the expression (1) in the approximate probability value $p(\bullet)$:

$$\hat{p}\left(x_i\in B\left(x_i,r_{k,x_i}\right)\right)=k/l \tag{3}$$

$r_{k,x_i}$ is distance of nearest neighbor of $x$.

By the formula (3) plug into (1), and take the logarithm to it, then get this.

$$\log(k)=d\log\left(r_{k,x_i}\right)+\log\left(\eta\left(x_i,r_{k,x_i}\right)\right) \tag{4}$$

In the above formula, $\eta(x_i,r)$ is a function with two variable $x_i$ and $r$ .when $r$ is small enough, $\eta(x_i,r)$ can be approximately regarded as a constant and disrelated to $r$ .Due to the formula (4) only take distance of nearest neighbor of a sample $x_i$ into account，so its lack of statistical stability and robustness. In the practical application is regard $\log\left(\eta\left(x,r_{k,x}\right)\right)$ as a constant and the $r_{k,x}$ replaced by Variables in the data set on the average [11].

$$\bar{r}_k=\frac{1}{l}\sum_{i=1}^{l}\left(r_{k,x_i}\right) \tag{5}$$

Thus, Formula (4) turn into:

$$\log(k)=d\log\left(\bar{r}_k\right)+\delta \tag{6}$$

By the formula (6) can get this:

(1) Logarithmic linear relationship existed between neighborhood size $k$ and $r_{k,x_i}$ distance of nearest neighbor, and can be through linear fitting method to estimate the local intrinsic low dimension $d$ of the $x_i$ nearest neighbor.

(2) Neighborhood size $k$ and intrinsic low dimension $d$ influence each other. Caused by improper selection of $k$, the estimate of $d$ will not accurate.

(3) It is a kind of global estimation method, the formula (4) showed a better ability to estimate the statistical characteristics and intrinsic low dimension $d$.

However, in different dataset for manifold learning, the local intrinsic low dimension information must have certainly difference [12]. The dimensionality reduction method based on k nearest neighbor doesn't take the difference into account, so the estimated the nature intrinsic low dimension is a tradeoff value, yet regardless of the local distribution information of high dimensional data. For a non variable dimension data, an effective way to solve the manifold learning model parameter selection is to find a method to structure large sample statistical properties well, and extract data local intrinsic low dimension information.

### B. *A modified model of manifold learning algorithm based on KNN algorithm*

The basic idea of improved manifold learning algorithm is that there is a certain relationship between distance of nearest neighbor and local intrinsic low dimension in the data set, so through calculating the weighted norm of distance of k nearest neighbor distance about the data set to indirect access to the local intrinsic low dimension distribution information of the manifold learning.

Definition:

$$\bar{r}_{k,p} = \left( \frac{1}{l} \sum_{l=1}^{l} r_{k,x_i}^{p} \right)^{\frac{1}{p}} \tag{7}$$

Then get this:

$$\log(k) = d \log\left(\bar{r}_{k,p}\right) + \delta_0 \tag{8}$$

So a function to get local intrinsic dimension through calculating the weighted norm with as $p$ variable was constructed.

(1) When local intrinsic dimension of data set consistent, for any and all $p$ can be approximate to accurately estimate the manifold intrinsic low dimension;

(2) When p = 1 or p = 0, formula (8) degenerate into a distance of the nearest-neighbor estimation method for intrinsic low dimension;

(3) When $p \to -\infty$, Formula (8) approximate get data set with minimum intrinsic low dimension; when $p \to -\infty$, Formula (8) approximate get data set with maximum intrinsic low dimension.

Thus, the estimating function with $p$ weighted norm variable can estimate the local intrinsic dimension information of the data accurately.

So the formula about the estimating function with $p$ weighted norm variable:

$$\log(k) = d \log\left(\bar{r}_{k,p}\right) + \delta_0 \tag{9}$$

$$\log(k+1) = d \log\left(\bar{r}_{k+1,p}\right) + \delta_0 \tag{10}$$

$$d_k(p) = \log(k+1/k) \Big/ \log\left(\bar{r}_{k+1,p} / \bar{r}_{k,p}\right) \tag{11}$$

Improved manifold learning algorithm for data dimension reduction based on KNN

Step1 Set up neighborhood range $\{k_{\min}, k_{\max}\}$ and $\{p_{\min}, p_{\max}\}$;

Step2 To normalize $X = \{x_i\}_{i=1}^{l} \subset R^n$, make the norm of it less than 1;

Step3 Calculate $d_k(p)$ and $Var\left(d_k(\bullet)\right)$ in the formula (2-11), select the optimum neighborhood parameter $k_{opt}$ according to $k_{opt} = \arg\min\left(Var\left(d_k(\bullet)\right)\right)$;

Step4 According to $d_{opt} = [d_{kopt}(1)]$ or $d_{opt} = [d_{kopt}(0)]$ ($[\bullet]$ is integer arithmetic) as estimated parameter $d_{opt}$ of the intrinsic dimension of the manifold data.

Step5 Output the neighborhood size parameter $k_{opt}$ And the intrinsic dimension parameters $d_{opt}$.

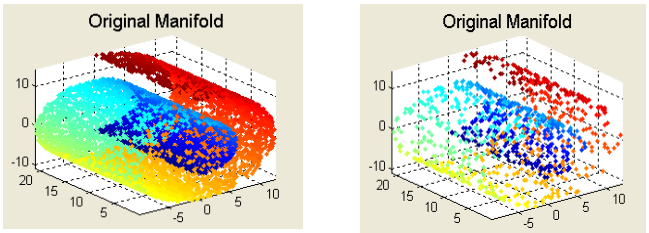
The above applications, the performance evaluation function $Var\left(d_k\right)$ as follows:

$$Var\left(d_k(\bullet)\right) = 1\text{-}R^2\left(D_G, D_Y\right) \tag{12}$$

where $D_Y$ is the Euclidean distance matrix of the $d$-space, $D_G$ is the shortest path distance and $R^2$ is correlation coefficient.

## 3. Simulation Experiments

**Example1** Use Swiss-roll structure data verify the validity of the improved manifold learning.




(a) (b)

Fig. 1 (a) The original manifold in Swiss-roll dataset (b) Take 1000 sample points in the data set

Towards to the scale of 1000 Swiss-roll manifold data set without noise (showed in the Fig. 1(b)), Neighborhood size k=8 and optimized intrinsic dimension $d_{opt} = 2$ (Showed in the Fig. 2, Arrow pointing in the direction of the optimum value of the neighborhood size)were calculated from the improved algorithm.

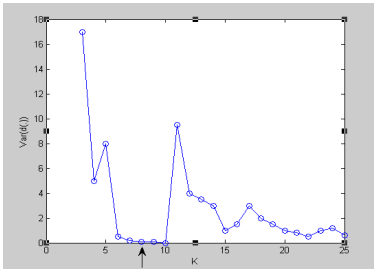


Fig. 2 The calculation of $Var(d_k)$ in Swiss-roll data set on the scale of 1000

**Example 2** Colon cancer data experiment. This experiment uses colon cancer data collected by Alon. In the colon cancer data, there are 62 samples, 2000 groups of genes, 40 lesions samples and 22 normal samples, the first 32 samples for training set before extraction, after 30 samples do the test set.

For verification the effect of the processing to the colon cancer based on the method is given in this paper, two different methods will using in experiments. One way is to choose the intrinsic dimensions and the size of the neighborhood based on experience; another way is to choose the parameters based on the method is given in this paper. The results are shown in Table I and Table II.

Introduction: According to the method that given in this paper, $k = 8$, $d_{opt} = 20$ is the calculated parameter of colon cancer data.

TABLE I The rate of colon cancer recognition after choose intrinsic dimension and neighborhood size empirical (%).

| | PCA | ISOMAP | LLE |
|---|---|---|---|
| k=7,d=10 | 53.3 | 80.3 | 79.4 |
| k=7,d=20 | 66.7 | 82.7 | 82.2 |
| k=7,d=50 | 73.3 | 77.8 | 77.5 |
| k=7,d=100 | 70.0 | 79.5 | 78.2 |

TABLE II The rate of colon cancer recognition after choose parameters based on improved Manifold learning model selection technique (%).

| | PCA | ISOMAP | LLE |
|---|---|---|---|
| k=8,d=10 | 57.3 | 85.6 | 82.4 |
| k=8,d=20 | 71.1 | 88.2 | 87.6 |
| k=8,d=50 | 77.2 | 82.5 | 81.9 |
| k=8,d=100 | 74.3 | 81.6 | 80.5 |

Comprehensive Table I and Table II shows:

(1) The dimensionality reduction effect between two kinds of method, the manifold learning Isomap and LLE is better than the traditional principal component analysis (PCA). It means that the data distribution of colon cancer is nonlinear manifold

(2) The rate of colon cancer recognition after choose parameters based on improved Manifold learning model selection technique is more higher than choose intrinsic dimension and neighborhood size empirical in the colon cancer data, that is also Verified the validity of the method;

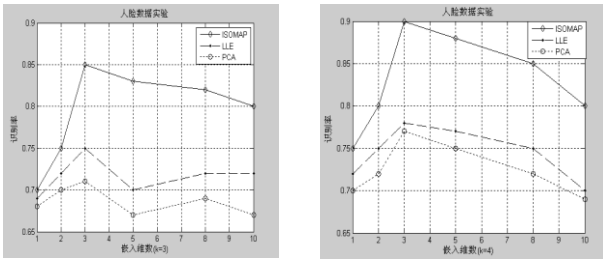
(3) When $k = 8$ , $d_{opt} = 20$ , the rate of colon cancer recognition is highest, shows that the manifold learning selection method based on weighted norm can choose intrinsic dimension $d$ and neighborhood size $k$ accurately.

**Example 3** Face data experiments. The experimental data is the face data download from www.datatang.com. There are 698 samples, 4096 d. First 398 samples as training set, after 300 samples as test set, using support vector machine (SVM) classification model to classify the experimental data.

According to the manifold learning selection method, The calculation of intrinsic dimension is $d_{opt} = 4$ and the neighborhood size is $k = 4$ . The results are plotted in the Fig. 3, which shows

(1) Model selection method based on weighted norm is better than the traditional principal component analysis (PCA) in the rate of face recognition.

(2) when $k = 4$ , $d_{opt} = 4$ , the rate of face recognition is highest. It can be shown that the manifold learning selection method based on weighted norm can choose intrinsic dimension $d$ and neighborhood size $k$ accurately.




(a) (b)

Fig. 3 (a) Choose empirical-based (b) model selection method based on weighted norm.


## Acknowledgment

This work was supported by National Natural Science Foundation (No. 5136501), and Jiangxi Provincial Natural Science Foundations (Nos. 20132BAB203020 and GJJ13430) of China.


## References


[1] Z. Yan and X. LIU, "Manifold Learning and Research of Algorithm," *Computer Technology and Development*, vol. 21, no. 5, pp. 99-102, 2011.

[2] X. GAO, "Problems and Analysis in Manifold Learning," *Computer Science*, vol. 36, no. 4, pp. 25-28, 2009.

[3] Z. Zheng, X. Chang and J. Yang, "ISO-Neighborhood Projection," *Journal of Computer Research and Development*, vol. 47, no. 7, pp. 1286-1293, 2010.

[4] L. He, J. Zhang and Z. Zhou, "Investigating Manifold Learning Algorithms Based on Magnification Factors and Principal Spread Directions," Chinese Journal of Computers, vol. 28, no. 12, pp. 2000-2009, 2005.

[5] Z. Zhang, H. Zha, "Linear low rank approximation and the nonlinear dimension reduction," *Science in China, Ser: A*, vol. 35, no. 3, pp. 273-285, 2005.

[6] G. Wen, L. Jiang and J. Wen, "Dynamically Determining Neighborhood Parameter for Locally Linear Embedding," *Journal of Software*, vol. 19, no. 7, pp. 1666-1673, 2008.

[7] A. Farahmand, C. Szepesvari, JY Audibert, "Manifold-adaptive dimension estimation," *Internation Conference on Machine Learning*, ACM Press, 2007, pp. 265-272.

[8] G. Wang, L. Huang, X. Zhao, "Nonlinear noise reduction method based on fractal dimension and the local tangent space mean reconstruction," *Journal of Electronic Measure and Instrument*, vol. 24, no. 8, pp. 699-704, 2010.

[9] J. Costa and A Hero, "Learning Intrinsic Dimension and Entropy of High-Dimensional Shape Spaces," *Statistic and Analysis of Shapes*, 2006, pp. 231-252.

[10] E. Levina and P. Bickel, "Maximum likelihood estimation of Intrinsic Dimension," *Neural Information Processing Systems*, 2005, pp.777-784.

[11] D. Meng, "A Research on Several Fundamental Problems and Core Algorithms of Manifold Learning," *Doctoral Dissertation of Xi'an Jiaotong University*, 2008.

[12] Z. Wang, X. Qian and M. Kong, "Survey on manifold learning algorithms," *Computer Engineering and Applications*, vol. 44, no. 8, pp. 9-12, 2008.