Pattern Recognition 42 (2009) 780 – 787

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Intrinsic dimension estimation of manifolds by incising balls

Mingyu Fan[a], Hong Qiao[b], Bo Zhang[c,*]

[a]Graduate School, Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China
[b]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[c]LSEC, Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 28 January 2008
Received in revised form 22 July 2008
Accepted 9 September 2008

Keywords:
Nonlinear dimensionality reduction
Manifold learning
Intrinsic dimension estimation
Data mining

ABSTRACT

Dimensionality reduction is a very important tool in data mining. Intrinsic dimension of data sets is a key parameter for dimensionality reduction. However, finding the correct intrinsic dimension is a challenging task. In this paper, a new intrinsic dimension estimation method is presented. The estimator is derived by finding the exponential relationship between the radius of an incising ball and the number of samples included in the ball. The method is compared with the previous dimension estimation methods. Experiments have been conducted on synthetic and high dimensional image data sets and on data sets of the Santa Fe time series competition, and the results show that the new method is accurate and robust.

1. Introduction

In modern world, scientists frequently encounter large volumes of data, and data can be converted into vectors in a high dimensional space such as the national statistics, snapshots of a moving object and the bio-current generated by millions of nerve sensors. Every digital image has three million pixels and could be converted into a vector in a three million dimensional space.

Due to the limitation of computational resources and storage space, it is difficult to process such data directly. There is a consensus that many types of data in a high dimensional space are not really high dimensional. Data with a few degrees of freedom could be regarded as points lying on a low dimensional manifold in a high dimensional space [1]. The hidden parameters of the data contain key information which is of pivotal importance for data visualization, storage and compression and so on. Dimensionality reduction finds compact representations of high dimensional samples which preserve the hidden structure for further processing. So dimensionality reduction plays a very important role in data analysis. Several classical methods have been proposed to recover parameters underlying the collected data points such as the principal components analysis (PCA) [2] and multidimensional scaling (MDS) [3]. PCA finds a linear space on which the projected data have maximum variance while MDS projects data points into a lower dimensional space by preserving pairwise Euclidean distances. But both methods can only discover linear parameters of the data. Naturally collected data points are frequently distributed on a nonlinear sub-manifold which is embedded in a high dimensional space. In the nonlinear case like the Swiss roll data set, linear dimensionality reduction methods will not work. Recently, several nonlinear dimensionality methods have been proposed to solve the problem, such as the isometric feature mapping (Isomap) [1], the locally linear embedding (LLE) [4], the Laplacian eigenmaps [5]and the Hessian LLE [6]. In recent years, Isomap and LLE have drawn great interests. They are simple to implement and avoid nonlinear optimization. However, both Isomap and LLE methods need the precise information of both the input parameters ε or k for the neighborhood identification and the intrinsic dimension d of the data set. The intrinsic dimension of a data set is very important since it is the fundamental information we need to know before any further analysis could be followed. If the intrinsic dimension d is set larger than what it really is, much redundant information will also be preserved; if it is set smaller, useful information of the data could be eliminated during the dimensionality reduction.

Many methods have been proposed to estimate the intrinsic dimension of data sets. Basically these methods could be classified into two groups [7]: projection methods and geometric methods. Projection methods, such as the method proposed by Fukunga and Olsen [8] (which will be referred to as local-PCA estimator), make use of locally linear projections of the data points, and the intrinsic dimension d is determined through comparing the largest variance in directions normal to the subspace with variation in directions of the subspace. Geometric methods, such as the fractal-based method [9] and the maximum likelihood estimation method [10], make assumptions on the geometrical distribution of data sets, which will be explained in detail in Section 2.

* Corresponding author. Tel.:+86 10 6265 1358.
E-mail addresses: fanmingyu@amss.ac.cn (M. Fan), hong.qiao@mail.ia.ac.cn (H. Qiao), b.zhang@amt.ac.cn (B. Zhang).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.09.016

2. Previous works on dimension estimation

As discussed above, the main methods to estimate the intrinsic dimension of a manifold in a high dimensional space can be classified into two groups [7]: projection methods and geometric methods.

Projection methods, such as PCA, have shown their capability in finding the dimension of a linear manifold in a n-dimensional space. The dimension is estimated by comparing the ratio of the sums of the top d eigenvalues and the remaining $n-d$ eigenvalues with a given threshold. However, if the data points distribute on a highly nonlinear manifold, PCA cannot give a correct intrinsic dimension. To improve the performance of PCA, kernel PCA [11] applies a Mercer kernel to project the data onto a feature space first and then performs PCA on the feature space. But selecting the appropriate Mercer kernel is a tricky problem. Under the assumption that the manifold is locally linear, local-PCA [8] performs PCA locally on the data set to estimate local dimension. Then the global dimension is determined by averaging the values of local dimension estimates. Nevertheless, the segmentation of the local regions is hard to decide. In order to solve this problem, the optimal topology preserving maps (OTPMs) was proposed in Ref. [12], which first constructs an optimal topology preserving map on the selected data centers and then performs PCA locally on the data centers. OTPMs works well on nonlinear manifolds but heavily depends on the choice of the data centers.

Geometric methods can be reviewed as below.

2.1. Fractal-based methods and related work

Fractal-based methods introduce a popular dimension definition, that is, the box-counting dimension, which is based on the fractal theory. However, for computational simplicity, the correlation dimension is used to replace the box-counting dimension.

Assume that the observations $x_1, x_2, \ldots, x_n$ lie on the manifold M in $\mathbb{R}^D$ and that the low dimensional embedding samples are in $\mathbb{R}^d$, where d is unknown and $d \ll D$. The correlation integral $C(r)$ is defined as

$$C(r) = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I(\|x_i - x_j\| \leqslant r),$$

where I is an indicator function. The correlation dimension D is then defined as

$$D = \lim_{r\to 0} \frac{\log(C(r))}{\log r}. \tag{1}$$

Based on Eq. (1), Grassberger and Procaccia [13] proposed a direct estimation method of the correlation dimension by measuring the slope of $\log(C(r))$ versus $\log r$. Camastra and Vinciarelli [14] improved the Grassberger–Procaccia (GP) method using an empirical procedure. However, for a data set with a finite number of samples, the limit $r \to 0$ in Eq. (1) will not be achieved. To overcome this problem, a scale-dependent correlation dimension is introduced in Ref. [15] based on geometric heuristics and L′ Hospital's rule. Assume that $S_n = \{x_1, \ldots, x_n\}$ is a finite data set, the scale-dependent capacity dimension is defined as

$$D_{\mathrm{cap}}(r_1, r_2) = -\frac{\log M(r_2) - \log M(r_1)}{\log(r_2) - \log(r_1)},$$

where $M(r)$ is the minimal number of boxes with size r covering S_n. This method avoids taking limit on a finite data set and does not require input parameters. However, finding $M(r)$ for S_n is an NP-hard problem.

A kernel version of the fractal-based method was introduced in Ref. [16]. This method works by replacing the indicator function I in the correlation integral $C(r)$ with a generalized kernel function $K(x, y)$.

2.2. Dimension estimation based on distances between samples and their neighbors

Assume that the data points are locally uniformly distributed. Then many authors have used distances between samples and their nearest neighbors to estimate intrinsic dimensions. Pettis et al. [17] proposed an iterative algorithm with intrinsic dimension estimator

$$d = \frac{r_k}{(r_{k+1} - r_k)k}, \tag{2}$$

where r_k is the mean distance between each sample and its k-nearest neighbors. However, their own experiments show that estimator (2) is biased. In Ref. [18], Verveer et al. pointed out that the bias cannot be corrected even with an iterative version. They proposed a non-iterative algorithm (nearest neighbor estimator):

$$d = \left[\sum_{k=k_{\min}}^{k_{\max}-1} \frac{(r_{k+1} - r_k)r_k}{k}\right] \times \left[\sum_{k=k_{\min}}^{k_{\max}-1} (r_{k+1} - r_k)\right]^{-1}.$$

Instead of using a single neighborhood size k, this method uses a range of neighborhood sizes, $k = k_{\min}, k_{\min}+1, \ldots, k_{\max}$. In Ref. [7], it is argued that this non-iterative version of the estimator is sensitive to noisy samples and has edge effects.

Costa et al. [19,20] proposed to estimate the intrinsic dimension by calculating the length of the k-nearest neighbors (k-NN) graph. The length of the k-NN graph is defined as

$$L_{\gamma,\kappa} = \sum_{i=1}^{n} \sum_{x\in N_{\kappa,i}} |x - x_i|^{\gamma},$$

where $N_{\kappa,i}$ is the set of the neighbors of x_i. In Refs. [19,20], Costa et al. also proved the following relationship between $L_{\gamma,\kappa}$ and the intrinsic dimension d:

$$\log L_{\gamma,\kappa} = a \log n + b + \varepsilon_n,$$

where $a = (d-\gamma)/d$, b is an unknown constant to be determined, and ε_n is an error residual going to zero with probability 1 as $n \to \infty$. The estimator works by sub-sampling subsets $X_1, X_2, \ldots, X_Q$ from the data set X. Assume that

$$l = [\log L_1, \ldots, \log L_Q]^{\mathrm{T}}, \quad A = \begin{bmatrix} \log n_1 & \cdots & \log n_Q \\ 1 & \cdots & 1 \end{bmatrix}^{\mathrm{T}}$$

where n_i is the cardinality of the subset X_i and L_i is the total length of the k-NN graph for X_i. Then applying the linear least square strategy yields $\hat{a}$ and the estimate $\hat{d} = \mathrm{round}\{\gamma/1 - \hat{a}\}$. However, a suitable selection of the neighborhood size k is difficult.

2.3. Dimension estimation based on distribution assumptions

Recently, in Ref. [10], a maximum likelihood estimation method was proposed and proved to have a better performance than the previous dimension estimation methods. Maximum likelihood estimation of intrinsic dimensions provides a maximum likelihood dimension estimator (MLE-estimator) to the nearest neighbors distances. Assume that x is an arbitrary data point and $S_x(t)$ is a sphere of radius t and centered at x. Consider the binomial function

$$N(t, x) = \sum_{i=1}^{n} I(X_i \in S_x(t)).$$