J. Vis. Commun. Image R. 25 (2014) 740–747

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Intrinsic dimensionality estimation based on manifold assumption

Jinrong He [a,*], Lixin Ding [a], Lei Jiang [b], Zhaokui Li [a], Qinghui Hu [a]

[a] State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan, Hubei 430072, China
[b] Key Laboratory of Knowledge Processing and Networked Manufacture, Hunan University of Science and Technology, Xiangtan 411201, China

ARTICLE INFO

Article history:
Received 7 March 2013
Accepted 8 January 2014
Available online 24 January 2014

Keywords:
Intrinsic dimension estimation
Dimensionality reduction
Graph distance
Geodesic distance
Manifold assumption
Geometric relation
Local neighborhood
Data analysis

ABSTRACT

Dimensionality reduction is an important tool and has been widely used in many fields of data mining and machine learning. Intrinsic dimension of data sets is a key parameter for dimensionality reduction. In this paper, a new intrinsic dimension estimation method based on geometrical relationship between manifold intrinsic dimension and data neighborhood geodesic distances is presented. The estimator is derived by manifold sampling assumption. On a densely sampled manifold, the number of samples that fall into a ball is equal to the volume times the density of the ball. The radius of the ball is calculated by graph distance which is approximation of geodesic distance on manifold. Then the intrinsic dimension is estimated on each sample. Experiments conducted on synthetic and real world data set show that the performance of our new method is robust and comparable to other works.

1. Introduction

Many objects in our world can be electronically represented as high-dimensional data, such as speech signals, images, videos, digital text documents and so on. Due to the limitation of computational resources and storage space, it is too complicated or too unstable to be feasible to process those data directly by regular systems. In order to process high-dimensional data, dimensionality reduction techniques become crucial [1]. Many techniques have been proposed to reduce dimensionality. Basically these techniques could be classified into two groups: linear methods and nonlinear methods. Linear methods, such as principal component analysis (PCA) [2], multidimensional scaling (MDS) [3], linear discriminant analysis (LDA) [4] and random projection [5], perform dimensionality reduction by embedding the data into a subspace of lower dimensionality. Nonlinear methods, such as locally linear embedding (LLE) [6], ISOMAP [7], Laplacian Eigenmaps (LE) [8], maximum variance unfolding (MVU) [9], local tangent space alignment (LTSA) [10], Hessian locally linear embedding (HLLE) [11] and diffusion maps (DM) [12], assume that the data settled resides on a low-dimensional nonlinear manifold.

However, all of these techniques require the user to specify the dimension of the Euclidean space into which the data will be mapped, and this dimension is called target dimension. Choosing it too small may result in the loss of significant information, whereas choosing it too large may obscure the underlying structure. Perhaps the first fundamental question one would ask is "what is the true, or intrinsic, dimension of this data?" Knowing the intrinsic dimension (ID) of the data will clearly facilitate further applications of dimensionality reduction methods and thus obtain meaningful low-dimensional representations of the data. The role of this procedure in data processing system is shown in Fig. 1.

There are many definitions of the dimension of a set, some are more compelling than others, that have been used to study complex geometric shapes. Unfortunately, these definitions are not consistent with one another. Therefore, the selection of a single definition among these competing options is a crucial choice that must be made before any further progress. Intuitively, intrinsic dimensionality is the minimum number of parameters that is necessary in order to account for all the information in the data [13]. In more general terms, following Fukunaga [28], a data set $\Omega \subseteq R^D$ is said to have Intrinsic dimensionality (ID) equal to d if its elements lie entirely within an d-dimensional subspace of R^d (where $d < D$). Many estimators of the ID come from fractal geometry. In our work, the data set is modeled by the theory of manifold geometry. A connected topological manifold is locally homeomorphic to Euclidean d-space, and the number d is called the manifold's dimension which can be defined as intrinsic dimension of data set.

* Corresponding author.
E-mail address: hejinrong@whu.edu.cn (J. He).

http://dx.doi.org/10.1016/j.jvcir.2014.01.006
1047-3203/

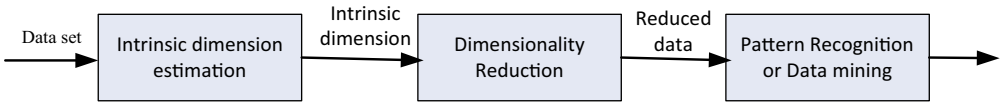

Fig. 1. The process of data analysis.

2. Previous work on intrinsic dimension estimation

A great deal of research work has been devoted to the development of algorithms to estimate the intrinsic dimensionality of a data set [14]. Traditional methods for intrinsic dimension estimation include PCA, which determines the ID by the number of eigenvalues greater than a given threshold. However, if the data points concentrate around a nonlinear manifold, PCA may not obtain the intrinsic dimension correctly. Recently, there has been a surge of interest in geometric methods, which focus on the fractal-based techniques. Of these techniques, correlation dimension and packing dimension are particularly prominent. If one has n identically independent distributed samples $x_1, x_2, \ldots, x_n$ from some domain $\Omega \subseteq R^D$, then the correlation dimension d_C of a data set is defined as

$$d_C = \lim_{r\to 0} \frac{\ln(C(r))}{\ln(r)}$$

where $C(r)$ is correlation integral which defined as

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

Here, r is radius to be set as a positive number. $I(\cdot)$is an indicator function which can be defined as:

$$I(\cdot) = \begin{cases} 1 & \|x_j - x_i\| \le r \\ 0 & \|x_j - x_i\| > r \end{cases}$$

Then the ID can be obtained by plot $\ln r$ versus $\ln C(r)$ and estimate the slope of the linear portion of the plot. This approach is first used by Grassberger and Procaccia [15]. Since the Grassberger–Procaccia algorithm performs badly on sets of high dimensionality, Camastra and Vinciarelli developed an empirical method which is the modification of the Grassberger–Procaccia algorithm [16]. The advantage of the correlation dimension is that it can be straightforwardly and quickly calculated. However, it also has disadvantage, that is, it will severely underestimate the ID under non-uniform distribution of the manifold. To tackle this drawback, Kégl defined packing dimension, and using packing numbers rather than covering numbers to improve the computational efficiency [17]. The packing dimension d_P of a data set is defined as

$$d_P = -\lim_{r\to 0} \frac{\ln(M(r))}{\ln(r)}$$

where $M(r)$ is the maximum number of data points that can be covered by a single hyper sphere with radius r.

In 2004, Costa and Hero [26,27] estimate the intrinsic dimension of a manifold using a global method based on minimal spanning trees of geodesic graphs (GMST). A similarity matrix based on the geodesic distances between all points is constructed and then a minimal spanning subgraph is computed. Then the intrinsic dimension is estimated from the subgraph using a method-of-moments technique. Define $N_k(x_i)$ to be the set of k nearest neighbors of x_i, They prove that the total edge length of the K-NN graph is then given by

$$L = \sum_{i=1}^{n} \sum_{x_j \in N_k(x_i)} ||x_j - x_i||^{\alpha} = n^{\frac{d-\alpha}{d}} c + \varepsilon$$

where $0 < \alpha < d$ is a parameter controlling locality, c is a constant not depending on the data distribution, and ε is an error residual that goes to zero a.s. as $n \to \infty$. By applying this technique in local neighborhoods, they extend the method to give local dimension estimates.

There are other estimating methods in literature. For example, Levina and Bickel used a Poisson process to model the number of samples landing in a ball around data point x_i, and then the ID can be estimated by maximum likelihood principle [18]. The estimator proceeds as follows:

$$d_{MLE} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{k-2} \sum_{j=1}^{k-1} \log \frac{L_k(x_i)}{L_j(x_i)} \right]^{-1}$$

where $L_k(x_i)$ is the Euclidean distance from the fixed point x_i to its kth nearest neighbor.

Farahmand et al. estimated the dimension locally around the data points using the nearest neighbor method and then combined these local estimates together [19]. By decomposing the sample data into locally linear low-dimensional patches, Brand [29] proposed a heuristic estimate of dimension based on the local geometric relations in the manifold. Raginsky and Lazebnik [30] introduced a technique for dimensionality estimation based on the notion of quantization dimension, which connects the asymptotic optimal quantization error for a probability distribution on a manifold to its intrinsic dimension. Haro et al. extended the maximum likelihood method by developing a translated Poisson mixture model [20]. Fan et al. proposed an ID estimator which derived by finding the exponential relationship between the radius of an incising ball and the number of samples included in the ball [21]. Mordohai and Medioni estimated geometric relationships by tensor voting from the perspective of instance-based learning, then the ID at each point can be found as the maximum gap in the eigenvalues of the tensor [22]. Most existing ID estimators generally underestimated ID when its value is sufficiently high, so Rozza et al. presents two ID estimators based on the statistical properties of manifold neighborhoods to reduce this effect [23].

In this paper the data points can be viewed as points constrained to lie on a low-dimensional manifold embedded in a higher dimensional space. The intrinsic manifold dimension will be estimated based on the geometric properties of the data distribution. This paper is organized as follows. Section 3 introduces a new intrinsic dimension estimation method based on graph distance neighborhood system under manifold assumption. Section 4 illustrates the behavior of the proposed approach in different situations and Section 5 makes some concluding remarks.

3. Proposed method

3.1. Geometric property based on manifold assumption

The intrinsic dimension of a data set depends on the global manifold on which samples of the data set locates. Assume that the data set $X = \{x_1, x_2, \ldots, x_n\} \subset R^D$ locally and uniformly distributed on an underlying d-dimensional manifold $M \subset R^D (d << D)$. If $x_1, x_2, \ldots, x_n$ are independent identically distributed samples from a density $f(x)$ in R^d, and $L_k(x_i)$ is the Euclidean distance from a fixed point x_i to its kth nearest neighbor in the sample, then the following geometric property can be inferred