Information Processing Letters 115 (2015) 492–501

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

An improved incremental nonlinear dimensionality reduction for isometric data embedding ☆

Xiaofang Gao [a,*], Jiye Liang [a,b]

[a] College of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, Shanxi 030006, China

ARTICLE INFO

Article history:
Received 10 March 2014
Received in revised form 16 November 2014
Accepted 8 December 2014
Available online 16 December 2014
Communicated by X. Wu

Keywords:
Manifold learning
Nonlinear dimensionality reduction
Incremental learning
ISOMAP
Design of algorithms

ABSTRACT

Manifold learning has become a hot issue in the field of machine learning and data mining. There are some algorithms proposed to extract the intrinsic characteristics of different type of high-dimensional data by performing nonlinear dimensionality reduction, such as ISOMAP, LLE and so on. Most of these algorithms operate in a batch mode and cannot be effectively applied when data are collected sequentially. In this paper, we proposed a new incremental version of ISOMAP which can use the previous computation results as much as possible and effectively update the low dimensional representation of data points as many new samples are accumulated. Experimental results on synthetic data as well as real world images demonstrate that our approaches can construct an accurate low-dimensional representation of the data in an efficient manner.

1. Introduction

Data from the real world is of a high-dimensional nature, and so it is very difficult to understand and analyze. Some linear dimensionality reduction techniques attempted to solve this problem by lowering the data dimensionality, such as PCA [8], MDS [18]. But they only can solve linear data. Since 2000, manifold learning has become a hot issue in the field of machine learning and data mining. Its main goal is to find a smooth low-dimensional manifold embedded in nonlinear high-dimensional data space. There are some algorithms proposed to extract the intrinsic characteristics of different types of high-dimensional data, such as ISOMAP [19], LLE [17] and so on. These algorithms aim to presever different geometrical properties of the data manifold, and formally transform the dimensionality reduction problem into an eigen-problem of matrices. Therefore, they are often mentioned as spectral embedding methods [21].

Most of these manifold learning algorithms operate in a batch mode, meaning that they have no incremental ability and all data points are need to be available during training [12]. However, in applications like video surveillance, and speech recognition, where data come sequentially, the batch methods seem clumsy: running them repeatedly is not only time consuming, but also wasteful to discard previous results [5]. So it is urgently necessary to develop incremental methods to efficiently find intrinsic properties of high-dimensional data. As more and more data points are obtained, the evolution of data manifold can reveal interesting properties of the data stream [12].

There have been some attempts to create incremental manifold algorithms, which can be roughly categorized into two groups. One group, known as out-of-sample

☆ This work was supported by the National Natural Science Foundation of China (Nos. 61303091, 61202018 and 61201453), Specialized Research Fund for the Doctoral Program of Higher Education (Nos. 20131401120004), the Natural Science Foundation of Shanxi (Nos. 2013021018-2 and 2012011014-4).

* Corresponding author.
E-mail addresses: gxfhtp@sxu.edu.cn (X. Gao), ljy@sxu.edu.cn (J. Liang).

http://dx.doi.org/10.1016/j.ipl.2014.12.004
0020-0190/

extension, attempts to parameterize new observations based on the assumption that all the known results are correct. Out-of-sample extensions for LLE, ISOMAP, LE are given by Bengio et al. [2], using kernel tricks to reformulate these algorithms. But the method may fail if the data manifold is non-uniformly sampled [16]. The another group tries to give more credible results, not only embedding new points but also updating the known results, such as incremental LLE [16], incremental Isomap [12], incremental LTSA [11], incremental LE [6], etc. All the recent methods in the latter group are restricted to dealing with only one new point per running, and thus they are forced to rerun as many times as the number of new data points. The total cost of the time complexity and memory requirement is high and even higher than those of re-running the original algorithms. As a further imperfection, the geometric structure of the manifold may be destroyed if the new sample does not lie in original sampled area.

In this paper, an improved incremental version of ISOMAP is proposed, which can use the previous computation results as much as possible and effectively update the low dimensional representation of data points as many new samples are collected simultaneously. The algorithm not only is more fit to the cognitive mechanism in our brain, but also improves the efficiency while the accuracy of the embedding results are not be decreased obviously. The experimental results on both synthetic "Swiss-roll" data set and two real images data sets show that the algorithm is feasible.

The corresponding works (main contributions) of our approaches include

- An effective method to update the neighborhood graph and geodesic distances matrix. Different from ISOMAP [19] and its incremental versions [12], the method does not re-compute and update k-NN neighborhood graph. It keeps the previous neighborhood relations as much as possible, only adds the new neighborhood relations related to some new points and deletes the original links leading to short circuits. And the method also does not update the geodesic distances one by one. It only updates the distances of two kinds of paths: the paths leading to the conflicting predecessor matrix; the paths including short circuits.
- A simple method to detect the short circuits in the neighborhood graph. The method re-estimates all weights of the edges in the original neighborhood graph in view of the newest "geodesic distance" between new points and all points (including all the original points and new points). At the same time, the thresholds of the weights are also estimated by computing the maximum distances of two neighborhood pitches. If its weight is larger than its threshold, the edge can be marked as a short circuits edge.
- A better solution of the incremental eigen-decomposition problem with increasing matrix size, which computes eigen-values and eigenvectors by subspace iteration with Rayleigh–Ritz acceleration. This differs from previous incremental ISOMAP version [12] where only one new sample is increased and its coordinate is directly estimated.

The rest of the paper is organized as follows: Section 2 reviews the related works. Section 3 describes the proposed incremental version of ISOMAP. Section 4 shows the complexity of the proposed algorithm and compares it with those of ISOMAP and law-IISOMAP. Section 5 presents the experimental results and finally Section 6 gives a conclusion.

2. Related works

Suppose that $M \subset R^D$ is a smooth manifold. A set of data points $\{x_1, ..., x_n\}$ is sampled from it. ISOMAP assumes that the data lie on a (Riemannian) manifold and maps x_i to its d-dimensional representation y_i in such a way that the geodesic distance between x_i and x_j is as close to the Euclidean distance between y_i and y_j in R^d as possible.

ISOMAP algorithm has three steps:

i. Constructing the neighborhood graph. ISOMAP requires specifying a parameter of the neighborhood: k-nearest neighbors (k-NN) or ε-hyper sphere. The k-NN version is more common since the sparseness of the resulting structures is guaranteed. The weighted undirected neighborhood graph $NG = (V, E)$ is constructed with the vertex $v_i \in V$ corresponding to x_i. An edge $e(i, j)$ between v_i and v_j exists if x_i is a neighbor of x_j. The weight of $e(i, j)$, denoted by w_{ij}, is the value of the Euclidean distance. If the set of the k-NN neighborhood of x_i is denoted by $knn(i)$ and the set of indices of the vertices adjacent to v_i in G is denoted by $adj(i)$, then $adj(i)$ is corresponding to $knn(i) \bigcup \{v_j | v_i \subset knn(j)\}$.

ii. Estimating the geodesic distances. The key assumption is that the geodesic between two points on the manifold can be approximated by the shortest path between the corresponding vertices in the neighborhood graph. Let g_{ij} denote the length of the shortest path $sp(i, j)$ between v_i and v_j. The shortest paths can be found by the Dijkstra's algorithm with different source vertices. The shortest paths can be stored efficiently by the predecessor matrix Π, where $\pi_{ij} = k$ if v_k is immediately before v_j in $sp(i, j)$. Since g_{ij} is the approximate geodesic distance between x_i and x_j, we shall call g_{ij} the geodesic distance. So the geodesic distance matrix $G = \{g_{ij}\}$ is symmetric.

iii. Recovering the embedding results $\{y_1, ...y_n\}$ by using the classical MDS on the geodesic distances. Let B be the target inner product matrix, i.e., the matrix of the target inner products between different y_i. If restricting $\sum_i y_i = 0$, B is recovered as $B = -HAH/2$, where $a_{ij} = g_{ij}^2$, $H = I_n - J_n/n$, I_n is an identity matrix and J_n is a matrix with $n \times n$ ones. We seek $Y^T Y$ to be as close to B as possible in the least square sense. Then the embedding result $Y = diag(\sqrt{\lambda_1} ... \sqrt{\lambda_d})[u_1 ... u_d]^T$ is achieved, where $\lambda_1, ..., \lambda_d$ are the d largest eigenvalues of B, with corresponding eigenvectors $u_1, ..., u_d$.

3. Incremental ISOMAP

According to the original ISOMAP algorithm, the main works in incremental algorithms involve three steps: up-