Information Sciences 270 (2014) 1–27

Contents lists available at ScienceDirect

Information Sciences

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

A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality

Antonio Gracia [*], Santiago González, Víctor Robles, Ernestina Menasalvas

Departamento de Arquitectura y Tecnología de Sistemas Informáticos, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660, Boadilla del Monte, Madrid, Spain

ARTICLE INFO

Article history:
Received 18 August 2013
Received in revised form 6 February 2014
Accepted 13 February 2014
Available online 20 February 2014

Keywords:
Manifold learning
Nonlinear dimensionality reduction
Linear dimensionality reduction
Loss of quality
Quality assessment criteria

ABSTRACT

Dimensionality Reduction (DR) is attracting more attention these days as a result of the increasing need to handle huge amounts of data effectively. DR methods allow the number of initial features to be reduced considerably until a set of them is found that allows the original properties of the data to be kept. However, their use entails an inherent loss of quality that is likely to affect the understanding of the data, in terms of data analysis. This loss of quality could be determinant when selecting a DR method, because of the nature of each method.

In this paper, we propose a methodology that allows different DR methods to be analyzed and compared as regards the loss of quality produced by them. This methodology makes use of the concept of preservation of geometry (quality assessment criteria) to assess the loss of quality. Experiments have been carried out by using the most well-known DR algorithms and quality assessment criteria, based on the literature. These experiments have been applied on 12 real-world datasets.

Results obtained so far show that it is possible to establish a method to select the most appropriate DR method, in terms of minimum loss of quality. Experiments have also highlighted some interesting relationships between the quality assessment criteria. Finally, the methodology allows the appropriate choice of dimensionality for reducing data to be established, whilst giving rise to a minimum loss of quality.

1. Introduction

The use of Dimensionality Reduction (DR) in recent decades has been motivated by the difficulties in analyzing very high dimensional data. Historically, the main DR applications have been, amongst others, the elimination of data redundancy and noise, the reduction in the number of features for minimizing the computational cost in data pre-processing, the identification of the most discriminative features and the reduction of features for visualization tasks.

However, the use of DR entails an inherent loss of quality that is likely to affect the understanding of the data, in terms of data mining. That is, patterns discovered and extracted from a dimensionally reduced data will probably be a small part of the patterns extracted from the original data. Furthermore, the meaning of these patterns may be altered by this reduction.

[*] Corresponding author. Tel.: +34 645469416.

E-mail addresses: antonio.gracia@upm.es (A. Gracia), sgonzalez@fi.upm.es (S. González), vrobles@fi.upm.es (V. Robles), emenasalvas@fi.upm.es (E. Menasalvas).

http://dx.doi.org/10.1016/j.ins.2014.02.068

On the other hand, each DR algorithm has been created to achieve a specific aim, which defines its specific nature. It is also true that, depending on its specification, a DR algorithm can give rise to more or less loss of quality at the time of reducing the data.

Different comparative studies comparing the different DR algorithms are currently being addressed in the literature [74,108,66]. Specifically, a set of quality assessment criteria, based on geometry-preservation concepts, have been used in several comparative research studies [77,37,119]. However, these studies are not sufficiently complete because of the lack of quality criteria and datasets used, as well as the fact that an exhaustive analysis of the geometry preservation is not carried out throughout the entire DR process (instead, it is carried out on a particular dimensionality, usually 2).

In this paper we propose a methodology for comparing DR algorithms based on the concept of loss of quality. Thus, the loss of quality could be strongly linked to the preservation of geometry. That is, the greater the loss of quality, the less the preservation of geometry. Hence, this methodology uses 11 quality assessment criteria to make a comparative analysis. Furthermore, this new approach attempts to address some of the shortcomings of the aforementioned studies.

The rest of this paper is structured as follows: Section 2 explains the basic concepts of a DR process and classification of DR algorithms. Quality assessment measures to calculate the preservation of geometry of data, used in the proposed methodology, are presented in Section 3. Previous comparative studies on DR, presented as related work, are detailed in Section 4. The proposed methodology for the comparison of DR methods is presented in Section 5. In Section 6 the environment for carrying out the experiments is described. The experimental results are also presented. Finally, Section 7 draws the main conclusions of the paper.

2. Dimensionality reduction methods

2.1. Basis

Based on the nomenclature stated in Table 1, Dimensionality Reduction (DR) can be defined as follows: X is made up of n datavectors $x_i (i \in 1, 2, \ldots, n)$ with dimensionality D. The DR techniques transform X with dimensionality D into a new dataset Y with a *target* dimensionality d' (where $d' < D$, often $d' \ll D$), while retaining the original geometric structure of high-dimensional data as much as possible [113]. The fundamental assumption that justifies the DR is that the original data actually lies, at least approximately, on a manifold (often nonlinear) of lower dimension than the original data space. The aim of DR is to find a representation of that manifold (a coordinate system) that will allow X to be projected on it and obtain Y, that is a low-dimensional and compact representation of the data.

Let d be the intrinsic dimensionality of the dataset. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [29,62]. Ideally, the reduced representation Y should have a dimensionality that corresponds to the intrinsic dimensionality of the data.

There are currently two canonical ways of dealing with data when carrying out a DR process. The first one does so in a linear way (Linear Dimensionality Reduction or LDR), while the second one is in a nonlinear way (Nonlinear Dimensionality Reduction or NLDR). LDR handles data containing linear dependencies. However, they are not powerful enough to deal with complex data. NLDR methods are assumed to be more powerful than linear ones, since the procedure to connect the latent variables (aka intrinsic dimensionality) to the observed ones (the dimensionality of the original space) may be much more

Table 1
Main nomenclature.

Notation	Description
D	Dimensionality of the high-dimensional data
d	Intrinsic dimensionality of the high-dimensional data
n	Total number of datapoints
M	Topological manifold
$\Re^D$	D-Dimensional Euclidean space where high-dimensional datapoints lie
$\Re^d$	d-Dimensional Euclidean space (low-dimensional space using d dimensionality)
x_i	the ith datapoint in $\Re^D$
y_i	the ith datapoint in $\Re^d$
X	Original dataset in $\Re^D$ ($X = x_1, x_2, \ldots, x_n$).
Y	Reduced dataset in $\Re^d$ ($Y = y_1, y_2, \ldots, y_n$).
Dg	Pairwise geodesic distance matrix in $\Re^D$
δ	Pairwise euclidean distance matrix in $\Re^D$
ζ	Pairwise euclidean distance matrix in $\Re^d$
Dg_{ij}	Pairwise geodesic distance between x_i and x_j
δ_{ij}	Pairwise euclidean distance between x_i and x_j
ζ_{ij}	Pairwise euclidean distance between y_i and y_j
k	Number of neighbors of a datapoint
Xi_k	Set of k nearest neighbors of x_i
Yi_k	Set of k nearest neighbors of y_i