Neurocomputing 97 (2012) 251–266

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

A new embedding quality assessment method for manifold learning

Peng Zhang [a,*], Yuanyuan Ren [b], Bo Zhang [c]

[a] Data Center, National Disaster Reduction Center of China, Beijing, PR China
[b] Career Center, Tsinghua University, Beijing, PR China
[c] Institute of Applied Mathematics, Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing, PR China

ARTICLE INFO

Article history:
Received 8 August 2011
Received in revised form
9 April 2012
Accepted 6 May 2012
Communicated by Xiaofei He
Available online 30 May 2012

Keywords:
Nonlinear dimensionality reduction
Manifold learning
Data analysis

ABSTRACT

Manifold learning is a hot research topic in the field of computer science. A crucial issue with current manifold learning methods is that they lack a natural quantitative measure to assess the quality of learned embeddings, which greatly limits their applications to real-world problems. In this paper, a new embedding quality assessment method for manifold learning, named as normalization independent embedding quality assessment (NIEQA) is proposed. Compared with current assessment methods which are limited to isometric embeddings, the NIEQA method has a much larger application range due to two features. First, it is based on a new measure which can effectively evaluate how well local neighborhood geometry is preserved under normalization, hence it can be applied to both isometric and normalized embeddings. Second, it can provide both local and global evaluations to output an overall assessment. Therefore, NIEQA can serve as a natural tool in model selection and evaluation tasks for manifold learning. Experimental results on benchmark data sets validate the effectiveness of the proposed method.

1. Introduction

Along with the advance of techniques to collect and store large sets of high-dimensional data, how to efficiently process such data issues a challenge for many fields in computer science, such as pattern recognition, visual understanding and data mining. The key problem is caused by "the curse of dimensionality" [1], that is, in handling with such data the computational complexities of algorithms often go up exponentially with the dimension.

The main approach to address this issue is to perform dimensionality reduction. Classical linear methods, such as principal component analysis (PCA) [2,3] and multidimensional scaling (MDS) [4], achieve their success under the assumption that data lie in a linear subspace. However, such assumption may not usually hold and a more realistic assumption is that data lie on or close to a low-dimensional manifold embedded in the high-dimensional ambient space. Recently, many methods have been proposed to efficiently find meaningful low-dimensional embeddings from manifold-modeled data, and they form a family of dimensionality reduction methods called *manifold learning*. Representative methods include locally linear embedding (LLE) [5,6], ISOMAP [7,8], Laplacian eigenmap (LE) [9,10], Hessian LLE (HLLE) [11], diffusion maps (DM) [12,13], local tangent space alignment (LTSA) [14], maximum variance unfolding (MVU) [15], and Riemannian manifold learning (RML) [16].

Manifold learning methods have drawn great research interests due to their nonlinear nature, simple intuition, and computational simplicity. They also have many successful applications, such as motion detection [17], sample preprocessing [18], gait analysis [19], facial expression recognition [20], hyperspectral imagery processing [21], and visual tracking [22].

Despite the above success, a crucial issue with current manifold learning methods is that they lack a natural measure to assess the quality of learned embeddings. In supervised learning tasks such as classification, the classification rate can be directly obtained through label information and used as a natural tool to evaluate the performance of the classifier. However, manifold learning methods are fully unsupervised and the intrinsic degrees of freedom underlying high-dimensional data are unknown. Therefore, after training process, we cannot directly assess the quality of a learned embedding. As a consequence, model selection and model evaluation are infeasible. Although visual inspection on the embedding may be an intuitive and qualitative assessment, it cannot provide a quantitative evaluation. Moreover, it cannot be used for embeddings whose dimensions are larger than three.

Recently, several approaches have been proposed to address the issue of embedding quality assessment for manifold learning, which can be cast into two categories by their motivations.

- Methods based on evaluating how well the rank of neighbor samples, according to pairwise Euclidean distances, is preserved within each local neighborhood.

* Corresponding author. Tel.: +86 10 528 11142.
E-mail addresses: paine.cheong@gmail.com, zhangpeng@ndrcc.gov.cn (P. Zhang).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.05.013

- Methods based on evaluating how well each local neighborhood matches its corresponding embedding under rigid motion or conformal mapping.

These methods are proved to be useful to isometric manifold learning methods, such as ISOMAP, MVU and RML. However, a large variety of manifold learning methods output normalized embeddings, such as LLE, HLLE, LE, and LTSA, just to name a few. In these methods, embeddings have unit variance up to a global scale factor. Then the distance rank of neighbor samples is disturbed in the embedding as pairwise Euclidean distances are no longer preserved. Meanwhile, anisotropic coordinate scaling (that is, separate scaling along each coordinate component) caused by normalization cannot be recovered by rigid motion or conformal mapping. As a consequence, existent methods would report false quality assessments for normalized embeddings.

In this paper, we first propose a new measure, named anisotropic scaling independent measure (ASIM), which can efficiently compare the similarity between two configurations under rigid motion and anisotropic coordinate scaling. Then based on ASIM, we propose a novel embedding quality assessment method, named normalization independent embedding quality assessment (NIEQA), which can efficiently assess the quality of normalized embeddings quantitatively. The NIEQA method owns three characteristics.

1. NIEQA can be applied to both isometric and normalized embeddings. Since NIEQA uses ASIM to assess the similarity between patches in high-dimensional input space and their corresponding low-dimensional embeddings, the distortion caused by normalization can be eliminated. Then even if the aspect ratio of a learned embedding is scaled, NIEQA can still give faithful evaluation on how well the geometric structure of data manifold is preserved.
2. NIEQA can provide both local and global assessments. NIEQA consists of two components for embedding quality assessment, a global one and a local one. The global assessment evaluates how well the skeleton of a data manifold, represented by a set of landmark points, is preserved, while the local assessment evaluates how well local neighborhoods are preserved. Therefore, NIEQA can provide an overall evaluation by combining the both.
3. NIEQA can serve as a natural tool for model selection and evaluation tasks. Using NIEQA to provide quantitative evaluations on learned embeddings, we can select optimal parameters for a specific method and compare the performance among different methods.

In order to evaluate the performance of NIEQA, we conduct a series of experiments on benchmark data sets, including both synthetic and real-world data. Experimental results on these data sets validate the effectiveness of the proposed method.

The rest of the paper is organized as follows. A literature review on related works is presented in Section 2. The anisotropic scaling independent measure (ASIM) is described in Section 3. Then the normalization independent embedding quality assessment (NIEQA) method is depicted in Section 4. Experimental results are reported in Section 5. Discussions and concluding remarks are given in Sections 6 and 7, respectively.

2. Literature review on related works

In this section, the current state-of-the-art works on embedding quality assessment methods are reviewed. For convenience and clarity of presentation, main notations used in this paper are summarized in Table 1. Throughout the whole paper, all data samples are in the form of column vectors. The superscript of a data vector is the index of its component.

Table 1
Main notations.

$\mathbb{R}^n$	n-Dimensional Euclidean space where high-dimensional data samples lie
$\mathbb{R}^m$	m-Dimensional Euclidean space, $m<n$, where low-dimensional embeddings lie
x_i	The i-th data sample in $\mathbb{R}^n$, $i=1,2,\ldots,N$
$\mathcal{X}$	$\mathcal{X}=\{x_1,x_2,\ldots,x_N\}$
X	$X=[x_1\ x_2\ \cdots\ x_N]$, $n\times N$ data matrix
$\mathcal{X}_i$	$\mathcal{X}_i=\{x_{i_1},x_{i_2},\ldots,x_{i_k}\}$, local neighborhood of x_i
X_i	$X_i=[x_{i_1}\ x_{i_2}\ \cdots\ x_{i_k}]$, $n\times k$ data matrix
$\mathcal{N}_k(x_i)$	The index set of the k nearest neighbors of x_i in $\mathcal{X}$
y_i	Low-dimensional embedding of x_i, $i=1,2,\ldots,N$
$\mathcal{Y}$	$\mathcal{Y}=\{y_1,y_2,\ldots,y_N\}$
Y	$Y=[y_1\ y_2\cdots y_N]$, $m\times N$ data matrix
$\mathcal{Y}_i$	$\mathcal{Y}_i=\{y_{i_1},y_{i_2},\ldots,y_{i_k}\}$, low-dimensional embedding of $\mathcal{X}_i$
Y_i	$Y_i=[y_{i_1}\ y_{i_2}\ \cdots\ y_{i_k}]$, $m\times k$ data matrix
$\mathcal{N}_k(y_i)$	The index set of the k nearest neighbors of y_i in $\mathcal{Y}$
e_k	$e=[1\ 1\ \cdots\ 1]^T$, k dimensional column vector of all ones
I_k	Identity matrix of size k
$\|\cdot\|_2$	L_2 norm for a vector
$\|\cdot\|_F$	Frobenius norm for a matrix

According to motivation and application range, existent embedding quality assessment methods can be categorized into two groups: local and global approaches. Related works in the two categories are reviewed respectively as follows.

2.1. Local approaches

Goldberg and Ritov [23] proposed the procrustes measure (PM) that enables quantitative comparison of outputs of isometric manifold learning methods. For each $\mathcal{X}_i$ and $\mathcal{Y}_i$, their method first uses procrustes analysis [24–26] to find an optimal rigid motion, consisting of a rotation and a translation, after which $\mathcal{Y}_i$ best matches $\mathcal{X}_i$. Then the local similarity is computed as

$$L(X_i,Y_i)=\sum_{j=1}^{k}\|x_{i_j}-Ry_{i_j}-t\|_2^2,$$

where R and t are the optimal rotation matrix and translation vector, respectively. Finally, the assessment is given by

$$M_P=\frac{1}{N}\sum_{i=1}^{N}L(X_i,Y_i)/\|X_iH_k\|_F^2, \tag{1}$$

where

$$H_k=I_k-\frac{1}{k}e_ke_k^T.$$

An M_P close to zero suggests a faithful embedding. Reported experimental results show that the PM method provides good estimation of embedding quality for isometric methods such as ISOMAP. However, as pointed out by the authors, PM is not suitable for normalized embedding since the geometric structure of every local neighborhood is distorted by normalization. Although a modified version of PM is proposed in [23], which further eliminates global scaling of each neighborhood and is suitable for conformal mappings, it still cannot address the issue of separate scaling of coordinates in the low-dimensional embedding.

Besides the PM method, a series of works follow the line that a faithful embedding would yield a high degree of overlap between the neighbor sets of a data sample and of its corresponding embedding. Several works are proposed by using different ways