Computers in Biology and Medicine 53 (2014) 48–54

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Locally linear representation Fisher criterion based tumor gene expressive data classification

Bo Li [a,b,c,*], Bei-Bei Tian [a,b], Xiao-Long Zhang [a,b], Xiao-Ping Zhang [c]

[a] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, Hubei 430065, China
[c] Department of Electrical and Computer Engineering, Ryerson University, Toronto, Ontario, Canada M5B 2K3

ARTICLE INFO

Article history:
Received 30 January 2014
Accepted 22 July 2014

Keywords:
Dimensionality reduction
Tumor gene expressive data
Feature extraction
Supervised learning

ABSTRACT

Tumor gene expressive data are characterized by a large amount of genes with only a small amount of observations, which always appear with high dimensionality. So it is necessary to reduce the dimensionality before identifying their genre. In this paper, a discriminant manifold learning method, named locally linear representation Fisher criterion (LLRFC), is applied to extract features from tumor gene expressive data. In LLRFC, an inter-class graph and an intra-class graph are constructed based on their genre information, where any tumor gene expressive data in the inter-class graph should select k nearest neighbors with different class labels and in the intra-class graph the k nearest neighbors for any tumor gene expressive data must be sampled from those with the same class. And then the locally least linear reconstruction is introduced to optimize the corresponding weights in both graphs. Moreover, a Fisher criterion is modeled to explore a low dimensional subspace where the reconstruction errors in the inter-class graph can be maximized and the reconstruction errors in the intra-class graph can be minimized, simultaneously. Experiments on some benchmark tumor gene expressive data have been conducted with some related algorithms, by which the proposed LLRFC has been validated to be efficient.

1. Introduction

With the emergence of tumor gene expressive data collected from DNA microarray, it comes true to simultaneously monitor expression of all genes in the genome, which contributes to make insight into biological processes and mechanisms of human diseases. However, how to interpret tumor gene expressive data still needs further demonstration. Up to now, many studies have been reported on tumor gene expressive data analysis [1–8], where key tumor genes selection and molecular classification of cancer are mainly concentrated on. It is the fact that tumor gene expressive data are always characterized by a large amount of variables (genes) with a small amount of observations (samples), thus before carrying out classification on them, some methods are recommended to reduce their dimensionality or extract features.

The popular linear methods involved in tumor gene expressive data analysis are principal component analysis (PCA) [41], partial least squares (PLS) [11,40] and independent component analysis (ICA) [9,10]. However, Pochet et al. systematically proved that nonlinear models are superior to those linear ones on many tumor gene expressive data sets in 2004 [12]. So how to nonlinearly mine the tumor gene expressive data has been attracting a lot of attention and some nonlinear models are presented. Alexandridis et al. put forward a nonlinear method with finite mixture distribution for tumor analysis [13]. Meanwhile, Martella et al. propose a nonlinear factor mixture model, where both factor factorization and normal mixture are integrated [14]. Moreover, other nonlinear feature extraction methods such as kernel methods and manifold learning have also been advanced for tumor gene expressive data analysis.

Unlike kernel methods, which nonlinearly extract features by a kernel transformation, manifold learning is straightforward to explore the inherent nonlinear structure hidden in the high dimensional space. Firstly manifold learning methods approach local manifold structures using k nearest neighbors (KNN), where any point and its k nearest neighbors will be viewed on a local super-plane. Then the locality can be well modeled by handling the linear computational rules in the local patch. At last, manifold learning pursues low dimensional embeddings of the original data by locality preserving. When mapping all the localities into a global framework, although the local geometry is linear, the corresponding global structure still shows its nonlinearity. In the last decade, some classical manifold learning algorithms have been presented. Among them, isometric feature mapping (ISOMAP) [15], Laplacian eigenmaps (LE) [18], locally linear

* Corresponding author at: School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei 430065, China. Tel.: +86 13006326136.
E-mail address: liberol@126.com (B. Li).

http://dx.doi.org/10.1016/j.compbiomed.2014.07.018
0010-4825/

embedding (LLE) [16,17] and their extensions are widely used for feature extraction or dimensionality reduction. They have yielded impressive results on artificial and real world datasets [19–21,42].

LLE is an effective method for data visualization. However, it exposes some limitations when applied to data classification. One is out-of-sample problem [22]. Another limitation is that the classical LLE does not take into account class information of the training samples, which displays negative impacts on the recognition accuracy.

In order to avoid the problem mentioned above, more and more supervised versions of LLE have been presented to deal with data classification. In the original LLE, the manifold local geometry is usually explored using KNN, where Euclidean distance is involved. In most cases, some points with different labels may also have a shorter Euclidean distance than those with the same class, which results in wrong neighborhoods for classification because some nearest neighbors are from those data with different classes. To address the problem, a method is brought forward to adjust neighborhood weights using class information, where the distance between any two points belonging to different classes is defined to be relatively larger than its Euclidean distance while those distances between points with the same label are preserved. The work is first presented by de Ridder et al. [23]. Instead of enlarging the between-class distances, Wen et al. utilize a nonlinear function to shrink the within-class distances [24], which shows similar impacts on recognition performance. These methods either enlarge between-class distances or shrink within-class distances. Thus Zhang poses an enhanced supervised model of LLE by reducing within-class distance and expanding between-class distance simultaneously [25]. Later, Zhang and Zhao define a probability-based distance that can enlarge the Euclidean distance for labeled and unlabeled points [29,30]. Combining to the class information, these methods endeavor to increase the accuracy of LLE by adjusting the distances between neighborhood points rather than by selecting the neighborhoods points. Thus Hui et al. [26] and Zhao et al. [27] impose a strict constraint that only points with the same class can be considered to be k nearest neighbors. But it must be noted that the neighborhoods points determined by the method mentioned above will be not enough to explore the manifold geometry structure when they are not densely sampled. Therefore, Han et al. propose a method to make a supplement [28]. According to the ascending Euclidean distances, the same class samples are predefined as neighborhood points, and then the remaining neighbors are searched from those with different classes. Moreover, to overcome out-of-sample problem, Kokiopouloua et al. propose an orthogonal neighborhood preserving projection (ONPP) method, which introduces a linear transformation to minimize the reconstruction errors in low dimensional space [43]. Later, Kokiopouloua et al. define a repulsion graph to extract supervised features, where an objective function is constructed to minimize the weighted difference of the reconstruction errors to distances between any two points with different labels in low dimensional space [44]. Similar to Kokiopouloua, Zhang et al. also design an intra-class graph and expect to explore a subspace with the minimum weighted difference of the reconstruction errors in the intra-class graph to distances between any two differently labeled points [45]. On the basis of ONPP, some other methods are presented to set the weights between nodes adaptively [46,47]. However, these modified versions mainly take advantage of class information to adjust the distances between points or to select the neighborhood points in KNN graph, where more parameters are introduced with the augment of the application difficulty.

In addition, some other supervised LLE algorithms combined with LDA have also been boomed. Based on the projection distances of the preprocessed points in LDA subspace, Pang et al. select the k minimum-distance points as the neighbors for each data point and then apply LLE [31]. This method can be viewed as the mode of LDA+LLE because LLE is introduced to extract features from those data handled by LDA. Zhang et al. present a unified framework of LLE and LDA [32,33]. This framework essentially equals to LLE+LDA, where LLE is firstly used to project the original data into a subspace and then LDA is employed to extract features discriminatively. Pang et al. also bring forward an integrated model which is linearly constructed by the objective functions of LLE and LDA under some constraints [34]. The model can be changed into LLE or LDA when the coefficient is one or zero, respectively. Furthermore, a local Fisher embedding (LFE) is put forward by de Ridder et al. [35], where local geometry and global class information are absorbed into a Fisher formulation. Li et al. also propose a supervised LLE algorithm named local linear discriminant embedding (LLDE) based on the fact that the embeddings cost function is invariant to translation and rescaling under sum-to-one constraint to the reconstruction weights in LLE, where the translations and the rescalings can be optimized with a modified LDA [36]. In above methods, the class information is globally involved because LDA is introduced to extract features. However, manifold learning is a nonlinear approach by locality learning. Thus it will contribute to explore the local structure discriminatively using the local label information associated to the corresponding points contained in local patch.

In this paper, a discriminant manifold learning method is applied to extract relevant biological correlations or "molecular logic" from tumor gene expression data. In the method, we have taken advantage of genre information of tumor gene expressive data, by which an intra-class graph and an inter-class graph can be constructed, respectively. In the intra-class graph, any point and its k nearest neighbors should be sampled from the same class points. On the contrary, for any points in the inter-class graph, it must select those with different classes to it as its k nearest neighbors. At last a Fisher criterion can be reasoned to find the optimal projection, which cam maximize the reconstruction errors in the inter-class graph and minimize the reconstructions errors in the intra-class graph in the low dimensional space, simultaneously.

The rest of paper is organized as follows. Section 2 describes classical LLE algorithm. Section 3 presents the proposed algorithm. Some experimental results and simulations are offered in Section 4. Then the whole paper is finished with conclusions in Section 5.

2. Review of LLE

Let $X=[X_1, X_2, ..., X_n] \in R^{D\times n}$ be n points in high dimensional space. The data are well sampled from a nonlinear manifold. The goal of LLE is to map the high dimensional data into a low dimensional manifold space with dimensionality $d(d \ll D)$. Let us denote the corresponding set of n points in the embedding space as $Y=[Y_1, Y_2, ..., Y_n] \in R^{d\times n}$. The outline of LLE can be summarized as follows:

Step 1: For each data point X_i, identify its k nearest neighbors by KNN.
Step 2: Compute the optimal reconstruction weights which can minimize the error of linearly reconstructing X_i by its k nearest neighbors.
Step 3: Calculate the low-dimensional embedding Y for X that best preserves the local geometry represented by the reconstruction weights and the corresponding k nearest neighbors.

In Step 1 Euclidean distance is always used to define neighborhood, which is composed of k points with the sorted bottom