Expert Systems with Applications 36 (2009) 7644–7650

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis

Xuehua Li *, Lan Shu

School of Applied Mathematics, University of Electronic Science and Technology of China, Jianshe Road, Chengdu 610054, PR China

ARTICLE INFO

Keywords:
Genomic microarray
Dimensionality reduction
Locally linear embedding
Kernel methods
Support vector machine

ABSTRACT

Accurate recognition of cancers based on microarray gene expressions is very important for doctors to choose a proper treatment. Genomic microarrays are powerful research tools in bioinformatics and modern medicinal research. However, a simple microarray experiment often leads to very high-dimensional data and a huge amount of information, the vast amount of data challenges researchers into extracting the important features and reducing the high dimensionality. This paper proposed the kernel method based locally linear embedding to selecting the optimal number of nearest neighbors, constructing uniform distribution manifold. In this paper, a nonlinear dimensionality reduction kernel method based locally linear embedding is proposed to select the optimal number of nearest neighbors, constructing uniform distribution manifold. In addition, support vector machine which has given rise to the development of a new class of theoretically elegant learning machines will be used to classify and recognise genomic microarray. We demonstrate the application of the techniques to two published DNA microarray data sets. The experimental results and comparisons demonstrate that the proposed method is effective approach.

1. Introduction

Genomic microarray data have attracted a great deal of attention, as reflected by the ever increasing number of publications on this technology in the past decade. A common application in microarray data analysis is to identify genes that, based on their expression levels, discriminate between known classes of experiments. This identification is often achieved by using various statistical measures to, gene-by-gene, correlate the expression levels with the classes of interest. The application of microarrays technology encompasses many fields of study. From the search for differentially expressed genes, genomic microarrays data present enormous opportunities and challenges for machine learning, data mining, pattern recognition, and statistical analysis, among others. In particular, microarray technology is a rapidly maturing technology that provides the opportunity to assay the expression levels of thousands or tens of thousands of genes in a single experiment (Shalon, Smith, & Brown, 1996). Nevertheless, the analysis of microarray data remains a challenge as one wish to investigate the possibility of the expression of thousands of genes across multiple samples. Naturally the issue of multiplicity arises as one examines the significance of large numbers of genes. Manifold learning is a perfect tool for data mining that discovers the structure of high dimensional data sets and provides better understanding of the data. Several different manifold learning algorithms have been developed to perform dimensionality reduction of low-dimensional nonlinear manifolds embedded in a high dimensional space. locally linear embedding (Roweis & Saul, 2000), Isomap (Tenenbaum, Silva, & Langford, 2000), Laplacian eigenmaps (Belkin & Niyogi, 2003), and Stochastic neighbor embedding (Hinton & Roweis, 2003) were originally proposed as a generalization of multidimensional scaling.

The locally linear embedding (LLE) algorithm belongs to a group of manifold learning methods is an unsupervised learning algorithm that can compute low dimensional, neighborhood preserving embeddings of high dimensional data. LLE is considered one of effective algorithms for dimensionality reduction, and has been used to solve various problems in information processing, pattern recognition, and data mining (Zhang, Wang, Zhao, & Zhang, 2004; Elgammal & Lee, 2004; Mekuz, Bauckhage, & Tsotsos, 2005). LLE algorithm computes a different local quantity, and calculates the best coefficients to approximate each point by a weighted linear combination of its neighbors, and then tries to find a set of low-dimensional points, which can be linearly approximated by its neighbors with the same coefficients that have been determined from high-dimensional points. However, LLE impressionable uniform distribution manifold and the number of nearest neighbors. When LLE is applied to real world data sets and applications, it displays limitations, such as sensitivity to the noise, outliers, missing

* Corresponding author.
E-mail addresses: leesoftcom@gmail.com (X. Li), shul@uestc.edu.cn (L. Shu).

0957-4174/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2008.09.070

data, and poor linear correlation between variables due to poorly distributed variables. In LLE algorithms, the free parameter is the LLE's neighborhood size, which unfortunately, has no direct method of finding the optimal parameter. The optimal neighborhood size for each problem is determined by the experimenter's experience. On the other hand, if the density of training data is uneven, it will decrease the precision of classification if only the sequence of first k nearest neighbors is considered and not the differences of distances.

The purpose of this paper is to fill these gaps by presenting a kernel method based LLE algorithm (KLLE). The kernel method (Schölkopf, Smola, & Müller, 1998; Shawe & Cristianini, 2004) is demonstrated as having the ability to extract the complicated nonlinear information from application data sets. The kernel function of the kernel method is a nonlinear mapping from input space $\mathscr{X} \subseteq \mathscr{R}^n$ onto feature space $\mathscr{H} \subseteq \mathscr{R}^N$, $\phi : \mathscr{X} \subseteq \mathscr{R}^n \to \mathscr{H} \subseteq \mathscr{R}^N$. The kernel method provides a powerful and principled way of detecting nonlinear relations using well-understood linear algorithms in an appropriate feature space. This approach decouples the design of the algorithm from specification of the feature space. Most importantly, based on the kernel method, the kernel matrix is guaranteed to be positive semi-definite, convenient for the learning algorithm receiving information about the feature space and input data, and projects data onto an associated manifold, such as PCA. In addition, to solve LLE KNN's parameter problems, fuzzy KNN (Keller, Gray, & Givens, 1985) adopts the theory of fuzzy sets to KNN, and fuzzy KNN assigns fuzzy membership as a function of the object's distance from its K-nearest neighbors and the memberships in the possible classes. This combination has two advantages. Firstly, fuzzy KNN can denoise training data sets. And secondly, the number of nearest neighbors selection, though not the most important, can consider the neighbor's fuzzy membership value.

This paper focused on genomic microarray analysis, which enables researchers to monitor the expression levels of thousands of genes simultaneously (Young, 2000). With the help of gene expressions, heterogeneous cancers can be classified into appropriate subtypes. To classify tissue samples or diagnose diseases based on gene expression profiles, both classic discriminant analysis and contemporary classification methods have been used and developed. Recently, different kinds of machine learning and statistical methods (Brown et al., 2000; Lee & Lee, 2003) have been used to classify cancers using genomic microarrays expression data. To evaluate the effectiveness of the proposed KLLE dimensionality reduction method for classification, two published data sets are used. The experiment shows that dimensionality reduction of genes can significantly increase classification accuracy.

2. Summary of kernel method and locally linear embedding

2.1. Kernel method

Kernel method has been demonstrated to be able to extract the complicated nonlinear information embedded on a data set (Schölkopf et al., 1998; Schölkopf & Smola, 2002), and has become one of the most popular approaches to learning from examples with many potential applications in science and engineering (Wang & Paliwal, 2003). Many algorithms for data analysis are based on the assumption that the data can be represented as vectors in a finite dimensional vector space, such as linear discrimination, PCA, or least squares regression, making extensive use of the linear structure. Kernel method solution comprises two parts: a module that performs the mapping input space $\mathscr{X}$ into the embedding feature space $\mathscr{H}$, the nonlinear mapping is $\phi : \mathscr{X} \subseteq \mathscr{R}^n \to \mathscr{H} \subseteq \mathscr{R}^N$, and a learn algorithm designed to discover linear patterns in that space. Firstly, creates a complicated linear feature space, and then work out what the inner product in that space would be, and finally find a direct method for computing that value in terms of the original inputs. In fact, the kernel function K is directly defined by the nonlinear mapping $\phi(\cdot)$, and the feature space $\mathscr{H}$ is simply derived from its definition. The main property of kernel function is that the fundamental concept of the kernel method is the deformation of the vector (lower) space itself to a higher dimensional space.

However, an explicit mapping $\phi(\cdot)$ does not always exist, and kernel method's conditions are not sufficient in guaranteeing the existence of a feature space. In practice, the mapping is performed implicitly by choosing the kernel trick (Schölkopf & Smola, 2002), which consists of expressing the inner product in $\mathscr{H}$ as an evaluation of a kernel function $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ for the data points x_i and x_j in the input space. Moreover, there is a problem when choosing the function $K(x_i, x_j)$, since not every function is guaranteed to give a valid feature space. One way of searching for a valid kernel function is to draw on Mercer's theorem (Haykin, 1999) which states that any continuous symmetric function $K(x_i, x_j)$ that satisfies the positive semi-definite condition

$$\int \int_{\mathscr{X} \times \mathscr{X}} K(x_i, x_j)\phi(x_i)\phi(x_j)dx_i dx_j \geqslant 0 \quad \text{and} \quad \int_{\mathscr{X}} \phi(x)^2 dx < \infty \qquad (1)$$

which is ensured to be a kernel for some valid feature space. This provides a flexible way of choosing the kernel mapping functions.

2.2. Locally linear embedding algorithm

The LLE (Roweis & Saul, 2000) is a manifold learning method that has aroused a great deal of interest in machine learning. It computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs and recovers the global nonlinear structure from locally linear fits. Essentially, the algorithm attempts to compute a low dimensional embedding with the property that nearby points in the high dimensional space remain nearby and similarly co-located with respect to one another in the low dimensional space. Put another way, the embedding is optimized to preserve the local configurations of nearest neighbors. The standard LLE algorithm (Roweis & Saul, 2000) consists of three steps: (1) finding the k nearest neighbours of each point, (2) measuring reconstruction error resulting from the approximation of each point by the neighbour points and compute the reconstruction weights that minimize the error, (3) computing the low-embedding by minimising an embedding cost function with the reconstruction weights. Steps 1 and 2 aim to characterize the geometry property and preserve it in a matrix. In Step 3, the low-dimensional embedding is calculated while preserving the geometry property in the high-dimensional space.

LLE computes dimensionality reduction that preserves the local neighborhood structure of the input data in the low-dimensional transformation. The transformation models the subspace manifold as a connected patchwork of locally linear surfaces. LLE is commonly justified using Taylor's theorem which states that any differentiable function is linear at the limit in a small area around a point. LLE works by identifying local neighborhood distance relationships, and by finding a mapping into a lower dimensionality that preserves them as much as possible. The selection of k value is the key to dimensionality reduction. There have been numerous papers (Lee & Verleysen, 2005; Marina & Shi, 2001; Kouropteva, Okun, & Pietikainen, 2002) suggesting that the selection of the neighborhood number k is important to the original LLE. If the number k is larger, the algorithm will ignore or even lose the local nonlinear features on the manifold, just as the traditional PCA performs. In contrast, if the number k is defined as smaller, LLE will split the continuous manifold into detached locality pieces, because the global characteristics are lost. On the other hand, it is well known that