Neurocomputing 143 (2014) 44–50

Contents lists available at ScienceDirect

Neurocomputing

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

Gene expression data clustering based on graph regularized subspace segmentation

Xiaoyun Chen *, Cairen Jian

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350116, China

ARTICLE INFO

Article history:
Received 16 November 2013
Received in revised form
7 June 2014
Accepted 9 June 2014
Communicated by L. Kurgan
Available online 17 June 2014

Keywords:
Gene expression data
Clustering
Graph regularization
Subspace segmentation

ABSTRACT

Gene expression data clustering offers a powerful approach to detect cancers. Specifically, gene expression data clustering based on nonnegative matrix factorization (NMF) has been widely applied to identify tumors. However, traditional NMF methods cannot deal with negative data and easily lead to local optimum because the iterative methods are adopted to solve the optimal problem. To avoid these problems of NMF methods, we propose graph regularized subspace segmentation method (GRSS) for clustering gene expression data. The global optimal solution of GRSS can be achieved by solving a Sylvester equation. Experimental results on eight gene expression data sets show that GRSS has significant performance improvement compared with other subspace segmentation methods, traditional clustering methods and various extensions of NMF.

1. Introduction

Cancer has already been one of the most serious diseases threatening public health for several decades. Precise and reliable identification of cancers is important for cancer diagnosis and treatment [1]. The different types of cancer can be distinguished based on the patterns of gene activity in the tumor cells by the microarray technology. With the rapid development of DNA microarray technology, it is now possible to measure expression levels of thousands of genes simultaneously. DNA microarray technology has changed many aspects of biological research [2]. Increasingly, the challenge is to gain insight into the mechanisms of biological processes and human diseases from such gene expression data [10]. Over the past few decades, a number of methods have been proposed and applied to gene expression data [1,3–6]. The biggest challenge of studying gene expression data lies in small sample sizes and high data dimensions, i.e. the 'large p, small n' problem. The sample size is usually from tens to hundreds while the genes often amount to more than one thousand, even to tens of thousands. In despite of these difficulties, many classification and clustering methods have been used to analyze gene expression data. In this paper, we focus on cluster analysis for gene expression data.

Clustering methods have been proved to be helpful to understand gene function and tumor typology. Gene-based clustering algorithms [7,8] group thousands of genes into several smaller clusters to find out different levels of gene expression, which is useful for understanding the functions of many genes. Sample-based clustering methods [1,5,6] cluster samples with the same or similar expression pattern to facilitate the discovery of new tumor types.

Sample-based clustering is a challenging issue due to the curse of dimensionality. Many clustering methods, such as hierarchical clustering (HC), self-organizing map (SOM), nonnegative matrix factorization (NMF) and its extension models, have been successfully used in gene expression data [6,9–13]. Nonnegative matrix factorization aims to find two nonnegative matrices whose product provides a good approximation for the original matrix [14]. Some extensions of NMF methods based on graph regularization have been proposed and applied to such fields as image recognition [15,16] and disease diagnosis [17]. Recently, NMF has been introduced to analyze the gene expression data, the study of Brunet et al. showed that NMF is more accurate than HC and more stable than SOM [10], and the accuracy of the gene expression data clustering is improved via sparse NMF [11]. More other NMF-based clustering algorithms for gene expression data can also be found in [6,12,13]. However, the standard NMF only works for nonnegative data, which undoubtedly limits its applications in gene expression data clustering. Fortunately, Ding et al. proposed convex nonnegative matrix factorization algorithm (CNMF) and semi-nonnegative matrix factorization algorithm (SNMF), which can be used on both positive and negative data sets [18]. What's worse, the NMF-based methods usually lead to local optimum because the iterative optimization procedures are adopted.

The main purpose of this paper is to solve those problems of NMF-based clustering methods mentioned above. Subspace segmentation has been shown to be a powerful tool for clustering image data set [19–21]. If gene expression data set in the same

* Corresponding author.
E-mail address: c_xiaoyun@21cn.com (X. Chen).

http://dx.doi.org/10.1016/j.neucom.2014.06.023
0925-2312/

cluster is treated as a subspace, we can use subspace segmentation to cluster gene expression data. However, the state-of-the-art subspace segmentation methods fail to discover the intrinsic geometrical structure of the data space, which is important to the real applications [15]. In this paper, we propose a novel graph regularized subspace segmentation method which makes the mapping function of subspace segmentation as smooth as possible. We also develop a global optimization scheme to solve the objective function by using a Sylvester equation. To the best of our knowledge, the subspace segmentation method has not been used to cluster the gene expression data yet.

The organization of the rest of this paper is as follows. In Section 2, we review some related work, such as subspace segmentation, normalized cuts and some extensions of nonnegative matrix factorization. Section 3 presents our graph regularized subspace segmentation method. In Section 4, experiments on gene expression data clustering are conducted. Conclusions are made in Section 5.

2. Related work

2.1. A brief review of subspace segmentation

Subspace segmentation is an important clustering method for machine learning, which has been successfully applied in machine vision and other fields, i.e. clustering and image representation [19,20,22].Given a data set drawn from a union of subspaces, the target of subspace segmentation is to group the data set into clusters with each cluster corresponding to a subspace [19]. The mathematical definition of this description is

Definition 1. (Subspace segmentation) [20]

Given a data set $X=[X_1,\ldots,X_k]=[x_1,\ldots,x_n]\in\mathbb{R}^{d\times n}$, which drawn from a union of k subspaces $\{S_i\}_{i=1}^{k}$, where d is the feature dimension and n is the sample size. Let X_i be a collection of n_i data points drawn from the subspace S_i, $n=\sum_{i=1}^{k}n_i$. The task is to segment the data set according to the underlying subspaces they are drawn from.

In the past two decades, many subspace segmentation methods have been proposed. Existing works on the subspace segmentation can be roughly divided into four categories: algebraic methods, statistical methods, iterative methods and spectral clustering-based methods [20]. A review of subspace segmentation can be found in [21].

The important task of spectral clustering-based subspace segmentation methods is to find an affinity matrix Z, where Z_{ij} measures the similarity between data points x_i and x_j. A common similarity measure between the two data points is $Z_{ij}=\exp(-\|x_i-x_j\|/\sigma)$, $\sigma>0$. However, this method cannot characterize the structure of data from subspaces [20]. Many recent works on the subspace segmentation proposed some new methods to construct the affinity matrix, such as Low Rank Representation (LRR) [19] and Least Square Regression (LSR) [20]. These methods express each data point x_i as a linear combination of all other data points $x_i=\sum_{j\neq i}Z_{ij}x_j$, and use the representational coefficient $(|Z_{ij}|+|Z_{ji}|)/2$ to measure the similarity between data points x_i and x_j. LRR and LSR use different regularizations on Z, which lead to different affinity matrices.

LRR is a low rank representation method, the goal of LRR is

$$\min_{Z} rank(Z) \quad s.t. \quad X=XZ \tag{1}$$

where $rank(Z)$ is the rank of Z, the solution of this problem is NP-hard. In [19], the nuclear norm is used to instead of $rank(Z)$. Thus LRR is used instead to solve the following problem:

$$\min_{Z} \|Z\|_* \; s.t. \quad X=XZ \tag{2}$$

where $\|Z\|_*$ is the nuclear norm of Z, defined as the sum of all the singular values of Z. It can be further extended to a noisy case:

$$\min_{Z} \|X-XZ\|_{2,1}+\lambda\|Z\|_* \tag{3}$$

where $\lambda>0$, $\|Z\|_{2,1}$ is $l_{2,1}$ norm of Z, defined as the sum of the l_2 norm of each column. More details about LRR can be found in [19].

LSR minimized the Frobenius-norm of Z:

$$\min_{Z} \|Z\|_F \quad s.t. \quad X=XZ \tag{4}$$

Its extended model for noisy case is as follows:

$$\min_{Z} \|X-XZ\|_F^2+\lambda\|Z\|_F^2 \tag{5}$$

where $\lambda>0$ and $\|Z\|_F$ is the Frobenius-norm of Z. More details about LSR can be found in Ref. [20].

2.2. Normalized cuts method [23]

Given a set of data points, a nice way of representing the data is in form of weighted graph $G=(V,E)$. The nodes of the graph are the data points, and an edge is formed between every pair of nodes. The weight on each edge $w(i,j)$ is the similarity between nodes i and j. Assume A and B are the two disjoint subsets of G, $A\cup B=V$ and $A\cap B=\varnothing$. In graph theory, the dissimilarity between A and B is called *cut*, defined as

$$cut(A,B)=\sum_{i\in A,j\in B} w(i,j) \tag{6}$$

The optimal bipartitioning of a graph is the one that minimizes this *cut* value. One of the most common segmentation approaches is Normalized Cuts method [23], the objective function is

$$Ncut(A,B)=\frac{cut(A,B)}{asso(A,V)}+\frac{cut(A,B)}{asso(B,V)} \tag{7}$$

where $asso(X,V)=\sum_{u\in X,t\in V} w(u,t)$. However, the problem in (7) is NP-hard. Normalized Cuts method relaxes y to take on arbitrary values, and then minimize the relaxed cost as follows:

$$Ncut(y)=\frac{y^T(D-W)y}{y^TDy} \tag{8}$$

where W is a symmetrical matrix with $W(i,j)=w(i,j)$ and D is a diagonal matrix with $D(i,i)=\sum_j w(i,j)$.

The solution of (8) can be solved by the generalized eigenvalue system

$$(D-W)y=\lambda Dy \tag{9}$$

Let $z=D^{1/2}y$ and rewrite (9) as

$$D^{-1/2}(D-W)D^{-1/2}=\lambda z \tag{10}$$

We can segment the graph by using z. More details about Normalized Cuts method can be found in [23].

2.3. Some extensions of NMF

Nonnegative matrix factorization has been successfully applied in image recognition, text analysis and other fields. Two representatives of NMF variants are convex nonnegative matrix factorization (Convex NMF) and semi-nonnegative matrix factorization (Semi-NMF) [18], because they removed the non-negative constraints of the input data set.

Convex NMF can be expressed as follows:

$$\min_{U,V} \|X-XUV\|_F^2 \quad s.t\ U,\ V\geq 0 \tag{11}$$

and Semi-NMF can be expressed as below:

$$\min_{U,V} \|X-UV\|_F^2 \quad s.t \quad V\geq 0 \tag{12}$$