Computers in Biology and Medicine ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Functional grouping of similar genes using eigenanalysis on minimum spanning tree based neighborhood graph

Q1 R. Jothi *, Sraban Kumar Mohanty, Aparajita Ojha

Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, India

ARTICLE INFO

Article history:
Received 5 October 2015
Accepted 12 February 2016

Q3 *Keywords:*
Gene expression analysis
Minimum Spanning Tree
Spectral clustering
Similarity graph,
Microarray analysis

ABSTRACT

Gene expression data clustering is an important biological process in DNA microarray analysis. Although there have been many clustering algorithms for gene expression analysis, finding a suitable and effective clustering algorithm is always a challenging problem due to the heterogeneous nature of gene profiles. Minimum Spanning Tree (MST) based clustering algorithms have been successfully employed to detect clusters of varying shapes and sizes. This paper proposes a novel clustering algorithm using Eigenanalysis on Minimum Spanning Tree based neighborhood graph (E-MST). As MST of a set of points reflects the similarity of the points with their neighborhood, the proposed algorithm employs a similarity graph obtained from k' rounds of MST (k'-MST neighborhood graph). By studying the spectral properties of the similarity matrix obtained from k'-MST graph, the proposed algorithm achieves improved clustering results. We demonstrate the efficacy of the proposed algorithm on 12 gene expression datasets. Experimental results show that the proposed algorithm performs better than the standard clustering algorithms.

1. Introduction

Microarray technology is a boon for biological researchers to monitor the expression of thousands of genes simultaneously. The results of microarray experiments are often organized as gene expression matrices whose rows represent genes and columns represent various environmental conditions or samples such as tissues. Set of genes with similar expression patterns are called as co-expression genes and extracting such genes from microarray experiments is an important task as it helps in identifying the group of genes involved in the same cellular processes [1,2].

Clustering is an unsupervised learning technique that is used extensively in various exploratory analyses to discover natural grouping in a given set of objects. The objective of gene-based clustering is to partition the microarray genes into k distinct groups, where k is the number of clusters which may or may not be known in advance. Genes with similar expression profiles should be put into a single cluster which corresponds to a particular macroscopic phenotype, such as clinical syndromes or cancer types [1]. Many clustering algorithms have been proposed in the literature for gene expression analysis [5,9,10,14,18,26,28].

Various approaches to clustering can be broadly categorized into hierarchical and partitional methods [3,4]. While hierarchical clustering generates a nested sequence of partitions in the form of dendrogram tree, partitional clustering directly divides the dataset into k clusters [1]. The traditional approaches like hierarchical and partitional methods lack in ability to detect the intrinsic clusters because of the heterogeneous nature of gene profiles [6].

Recently more new techniques using graph theory have been proposed for gene expression analysis. Graph models are widely studied in various problem domains due to their ability to represent complex data. In graph-based clustering, the gene profiles are represented as a weighted undirected graph $G=(V,E)$ called as gene network, where each node $v\epsilon V$ corresponds to a gene and the edgeset E represents the dissimilarity (distance) between the genes [8]. Gene expression data are often highly connected and clustering algorithms based on graph connectivity are becoming popular in identifying the co-expression genes with highly intersecting profiles. Refs. [9,10,14,18] witness the effectiveness of graph-based clustering algorithms for the identification of similar genes based on their connectivity.

Given a connected, undirected graph, a spanning tree of the graph is an acyclic subgraph that spans all the vertices in the graph. Minimum Spanning tree (MST) is the spanning tree with minimum weight. MST is a well-known combinatorial optimization problem that is successfully applied for various tasks such as image segmentation and cluster analysis [14–18]. The key idea of MST based clustering is to identify and remove the inconsistent

* Corresponding author. Tel.: +91 761 2632273.
E-mail addresses: r.jothi@iiitdmj.ac.in (R. Jothi), sraban@iiitdmj.ac.in (S.K. Mohanty), aojha@iiitdmj.ac.in (A. Ojha).

http://dx.doi.org/10.1016/j.compbiomed.2016.02.007
0010-4825/

edges in order to obtain a set of clusters. Besides having a number of clustering algorithms using MST, an unified approach to detect and remove the inconsistent edges of MST that results in desired cluster structure is always being a challenging one.

Another popular clustering method that is predominantly used in several fields of data analysis is spectral clustering. Spectral methods find a wide range of practical applications including bioinformatics [26,28]. The core of spectral clustering is the Laplacian matrix which is the difference between degree and adjacency (weight) matrices of the similarity graph. The spectrum (eigenvalues) of the Laplacian matrix is used to partition the given dataset [24].

Although spectral methods are widely discussed, there has been a little attention on how the similarity graph is constructed from the given set of points. It is worthwhile to note that the results of spectral clustering depend on the choice of similarity graph used to encode the given dataset [20]. A similarity graph should depict the underlying structure of the dataset in order to achieve better clustering results. The most commonly used similarity graphs in the literature are ε-neighborhood graph, K-nearest neighbor graph and fully connected graph [19]. To the best of our knowledge, no theoretical study has been done about which similarity graph suits a particular dataset and how to choose the parameters e.g., ε, K, σ in case of ε-neighborhood graph, K-nearest neighbor graph and fully connected graph respectively.

Most of the algorithms become inefficient when applied on heterogeneous datasets which are diverse in shape, size and densities which is a common problem in microarray analysis. Moreover, microarray datasets often contain noise and outliers and thus finding a robust clustering algorithm is always a challenging problem [1]. As MST-based and spectral clustering methods have shown better clustering performance in identifying arbitrary shaped clusters, the proposed algorithm E-MST aims to integrate these two methods. Our proposed algorithm is free from any user-defined parameters.

The rest of the paper is organized as follows. In Section 2, related work in gene data clustering is presented with particular attention given to the graph based cluster analysis. Section 3 highlights our contribution. Spectral clustering method is introduced in Section 4. The proposed algorithm is described in Section 5. The results of experimental validation are reported and discussed in Section 6, and finally our research is concluded in Section 7.

2. Related work

Hierarchical clustering algorithms have been used extensively for microarray analysis due to their ability to describe the clustering results through multi-level visual structure. Eisen et al. applied hierarchical agglomerative clustering for gene expression analysis and also developed a graphical display called as Eisen plot to aid visual illustration of group of genes that share similar expression patterns [5]. The hierarchical algorithms work on distance matrix of $O(n^2)$ computations, which is expensive for large datasets.

As opposed to hierarchical approach, partitional algorithms directly divide the data points into number of clusters without imposing the hierarchical structure. K-means is one of the most popular partitional clustering algorithms used for gene expression analysis due to its ease of implementation and linear complexity [1]. However, K-means is highly sensitive to the choice of initial centers. If the initial centers are improperly chosen, then the algorithm may converge to a local optimum [1,7].

Graph based clustering algorithms are becoming popular in recent years as they seek partition based on connectivity rather than on centroid. CLuster Identification via Connectivity Kernels (CLICK) [9] models the gene expression clustering as a graph partitioning problem. It iteratively determines the min-cut to partition the graph into highly connected components according to certain homogeneity threshold. The results were further refined using two post-pruning steps such as adoption and merging. The results of CLICK demonstrated its improved cluster quality in terms of cluster separation and homogeneity. However it may not perform well while the gene expression data contains highly intersecting profiles, which is a commonly occurring scenario in microarray experiments [1].

Clustering gene network by exploiting neighborhood properties of each gene in the network is also widely studied in the microarray analysis [10–12]. Nearest Neighbor Network (NNN) algorithm makes use of mutual nearest neighborhood principle to extract more complex and biologically relevant clusters [10]. The NNN algorithm achieves higher precision in detecting functionally related genes. However the results of the algorithm depend on the size of the neighborhood [8].

Ruan et al. applied an efficient graph partitioning algorithm QCut to identify dense subgraphs from the gene co-expression network [11]. First a rank-based gene network is constructed and then QCut algorithm with an objective function called modularity is used to automatically determine the optimal partitioning and the number of partitions. The authors of the paper believe that the rank-based co-expression networks (those that utilize the rank-transformed similarities) can better capture the global topology of the network, identifying both strongly and weakly co-expressed modules, whereas the conventional value-based methods (those that utilize similarity values) can only exhibit the strongly co-expressed modules.

Baya et al. proposed penalized K-Nearest Neighbor graph based clustering for gene expression analysis (PKNNG) [12]. They first construct KNN graph of the given gene expression data for low value of K in the range 3–7. If the KNNG is disconnected, then they connect the subgraphs by adding edges with an exponentially penalized weight $W = d(i,j) * \exp^{d(i,j)}$. Different connecting schemes discussed in [12] are: MinSpan (uses edges of the minimum spanning set to connect the subgraphs), AllSubgraphs (connects each subgraph to all other subgraphs using minimum length edges), AllEdges (similar to fully connected graph, but the edges across different subgraphs have penalized weight) and Medoid (connects medoids of each subgraph to medoid of remaining subgraphs). Once a connected graph known as Penalized KNNG (PKNNG) is obtained, the underlying neighborhood relation is used as a metric for measuring the pairwise similarity of the points and any traditional clustering approaches such as K-means or hierarchical methods can utilize this metric for clustering. They reported that PKNNG with MinSpan connecting scheme and applying PAM as a final clustering method have been shown to provide the better clustering results.

MST of a set of points can be used to reflect the similarity of the points with their neighborhood and simply removing $k-1$ inconsistent edges (for eg. longest edges) from the MST results in k disjoint subsets such that each subset would represent a cluster [13]. The key advantage of the MST-based clustering algorithm is that the points of a cluster are neither grouped around cluster center nor separated by a geometric boundary. Thus the performance of these clustering algorithms do not depend on the shape of the cluster and they can detect clusters of irregular boundaries [14,15].

Identifying a suitable measure of inconsistent edges to partition the MST is a crucial step in MST-based clustering algorithms. In most of the cases, longest edges of the MST may not actually represent the inconsistent edges and this is illustrated in Fig. 1. Many researchers proposed different partitioning criteria [14–18]. Xu et al. proposed a MST based clustering algorithm for gene

Please cite this article as: R. Jothi, et al., Functional grouping of similar genes using eigenanalysis on minimum spanning tree based neighborhood graph, Comput. Biol. Med. (2016), http://dx.doi.org/10.1016/j.compbiomed.2016.02.007