Yeast Gene Expression Analysis and MRI Lesion Detection with MST

Latha Parthiban

Assistant Professor, Department of Computer Science
Pondicherry University Community College
lathap.ucc@pondiuni.edu.in

***Abstract*— Clustering is an investigative data analysis tool which gained enormous concentration mainly in gene expression field. *K*-means is widely used but it does not find the optimal cluster. The data from microarray experiments is usually in the form of large matrices. The method to process such a huge volume of data must be capable of detecting patterns efficiently. In this paper, Minimum Spanning Tree (MST) approach for gene database clustering is proposed to group similar patterns. MST has several advantageous over other methods in terms of speed, high accuracy and automation. In MST representation, the inter-data relationship is greatly simplified so that no essential information for clustering is lost and it does not depend on detailed geometric shape of the clusters. The entire approach of this paper is divided into five stage process such as Representation of gene matrix, preprocessing, rewriting the matrix into sparse matrix format, applying MST algorithm on the sparse matrix and calculating validity measure to enhance the clustering process. Experimental results of this paper also show that proposed approach can also be applied to medical images to find lesions.**

***Keywords*—Clustering, Minimum Spanning Tree, DNA Microarray, Validity Measure, Sparse Matrix.**

I. INTRODUCTION

DNA microarrays has been carried out using statistical analysis, machine learning and data mining approaches to analyze the expression of many genes simultaneously and it provide the means to analyze the patterns of gene expression at different time points in a living cell that is experimented in glass slides that have thousands of short DNA strands attached to their surfaces.

Gene expression data show the level of activity of several genes under experimental conditions over time series[2]. Time series expression data is classified into: short and long time series with 80% of microarray data belonging to short time series [12]. A good clustering is one where the Intra-cluster distance(Intra) is the sum of distances between objects in the same cluster are minimized and the Inter-cluster distance(Inter) is the distances between different clusters are maximized[4].

The rationale of gene clustering is that genes with high degree of expression similarity are functionally related, forms complex structures, participate in common pathways and is co-regulated by common upstream regulatory elements[3]. SOM (Self Organizing Map), Hierarchical clustering, Principal Component Analysis (PCA), and Multi-Dimensional Scaling & ESOM(Enhanced Self Organizing Map) are visualization methods for high dimensional data[5][6] .Clusters from Short Time series Expression Miner (STEM) and Fuzzy clustering by Local Approximation of Membership (FLAME) are consistent[10].

Fuzzy C-Means[11] and Fuzzy clustering by Local Approximations of Memberships (FLAME) [10] in which each gene is assigned a cluster membership which indicates the degree of belonging in each cluster. But, Fuzzy clustering approaches are sensitive to initialization and possibility for coincident clusters. The performance of different clustering algorithms is strongly dependent on both data distribution and application requirements.

The K-means algorithm is one of the standard tools for clustering but the major limitation is inability to determine the number of clusters and high computational complexity. The major drawback of basic K-means algorithm is that it produces different clusters for different set of values of the initial centroids. Several alternatives of k-means have been proposed but they also have drawback of automation.

II. MINIMUM SPANNING TREE

Graph based clustering algorithms are very powerful in clustering process and MST is a concept from graph theory used for representing multidimensional gene expression data. MST is well suited for finding arbitrary shaped clusters[6] and they are highly efficient and much flexible than Fuzzy C Means in detecting discriminatory patterns from medical images[7]. Spanning tree clustering produces high classification accuracy compared to other standard approaches [8]. Zahn's MST-based algorithm [1] first

constructs a MST of the data set and then removes those edges whose weight is significantly larger than the average of nearby edge weights. But this method fails when one vertex is connected with more than one edge.

Definition: MST is a sub graph of an weighed undirected graph G, such that i) it is a tree (acyclic), ii)it covers all vertices V and it contains |V|-1 edges And iv) the total cost connected with tree edges is the minimum among all possible spanning trees.

A gene expression data set from a micro array experiment can be represented by a real-valued expression matrix M={ w_{ij}| $1 \le i \le n, 1 \le j \le m$} where the rows (R={ $g_1, g_2 \ldots g_n$}) form the expression patterns of genes, the columns(S={$S_1, \ldots S_m$}) represent the expression profiles of samples and each cell is w_{ij} is the measured expression level of gene i in sample j. We define a weighted (undirected)graph G(M)=(V,E) as the vertex set V={$w_k|w_k \in M$} and the edge set E={$(w_k,w_l)|$ for $w_l,w_k \in M$ and $k \neq l$}. Hence G(M) is a complete graph. Each edge $(u,v) \in E$ has a weight which represents the distance $\rho(u,v)$ and distance between u and v is defined as Euclidean distance.

A. Data representation

Representation of data influences the clustering process. The following is the one of the common approaches to data normalization for each gene vector:

$$w'_{i,j} = \frac{w_{i,j} - \overline{w_i}}{\sigma_i} \quad (1)$$

$$\text{Where } w_i = \sum_{j=1}^{n_s} w_{i,j} \quad \sigma_i = \frac{\sum (w_{i,j} - \overline{w_i})^2}{n_s - 1} \quad (2)$$

After vector normalization , the data set G(M) is filtered to remove the values that do not show any interesting changes during the experiment. During this process the genes with small variance over time are removed by threshold and then sparse matrix is generated on the preprocessed image .

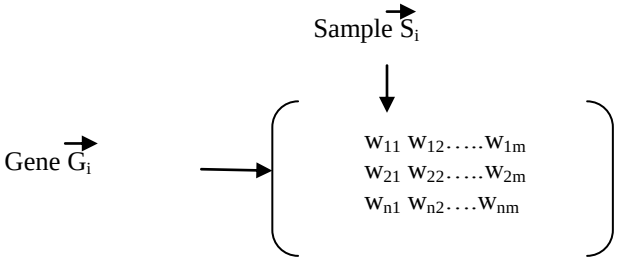

Fig.1. Gene Matrix

The concept of sparse matrix save a significant amount of memory by saving only non zero elements and speed up the processing of that data. A MST of a weighted graph can be found by a greedy method as illustrated by the following strategy used in the classical Prims algorithm. The algorithm grows MST one edge at a time by adding a minimal edge that connects a node in the growing MST with any other node. Time complexity is O(E*log(N)), where N and E are the number of nodes and edges respectively.

III. ITERATIVE ALGORITHM

A ny clustering algorithm that partition the MST T into K subtrees, to optimize a more general objective function(3) .

$$\sum_{i=1}^{K} \sum_{d \in T_i} \rho(d, \text{center}(T_i)) \quad (3)$$

so that distance between the center of each cluster and its data points is minimized, here the center is the average weight of a cluster. Here Correlation distance is used for the center

$$\text{Center}(C) = \sum c_i / \boldsymbol{\sigma_i} \quad (4)$$

where σ_i is the standard deviation of $c_i \in C$.

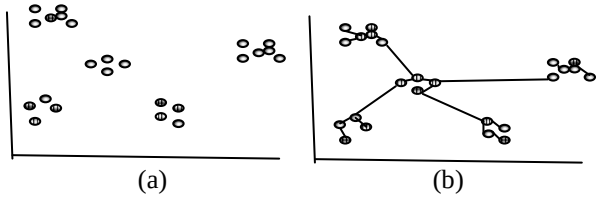

Fig. 2. (a) MST Clustering Results (b) Data points connected using distance measure.

The algorithm begins with K-partitioning and iterative repetition of the following steps till convergence: For every pair of neighboring clusters, visit all edges to locate the edge to cut and optimization achieved with objective function (1). By examining the Fig 2. (b) we observe that data points of the same cluster are connected with each other by sort tree edges whereas long tree edges link clusters together. We provide a necessary condition for a subset of a set to be a cluster. Let M be a dataset and ρ represent the distance between two data points of M. Each of the four clusters in Figure 2 satisfies this necessary criterion.

IV. VALIDITY MEASURE

The validity measure is calculated as intra/inter. The intra and inter distances are given in (5) and (6) respectively. The clustering which gives a minimum value for the validity measure will tell us what the ideal value of *K*.

$$\text{intra} = 1/N \sum_{i=1}^{K} \sum_{x \in K_i} \|x - C_i\|^2 \quad (5)$$

The intra cluster distance is the average of distance between a point and its cluster center . Here N is the number of data items, *K* is the number of clusters, and C_i is the cluster center of cluster K. and the requirement is to minimize this measure. The inter-cluster distance, or the distance between clusters is also measured which must be as large as possible. This is calculated as the distance between cluster centers, and the minimum of this value is defined as

$$\text{inter} = \min(\|C_i - C_j\|^2) \quad i=1,2,\ldots\ldots K-1, \; j=i+1,i+2,\ldots\ldots K \quad (6)$$

V. EXPERIMENTAL RESULTS

The experimental results on yeast gene expression profile is tested using MATLAB with bioinformatics toolbox. Figure 3a provides the Yeast gene expression profiles and figure 3b provides Yeast Gene Expressions after MST. Figure 4a provides the K-means clustering of Cancer data used in [9] and figure 4b provides a part of MST of cancer data profiles and figure 4c the clustering results. Figure. 5a provides the test MRI image with lesions, 5b its MST and 5c its clustering results.

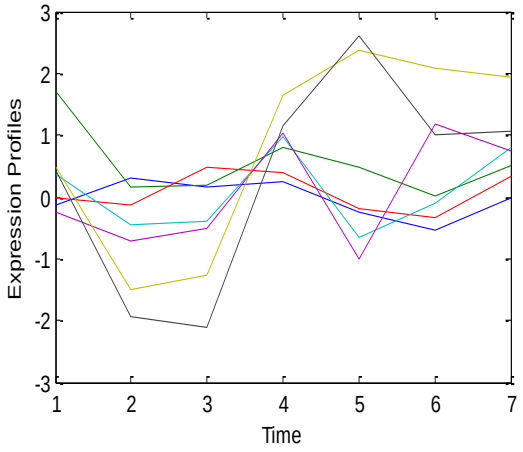

Fig. 3. (a) Yeast gene expression profiles

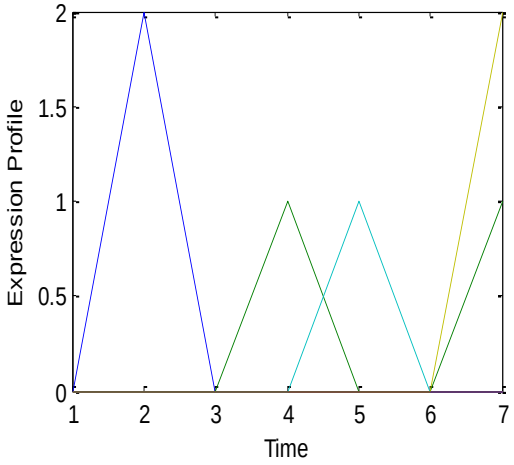

Fig. 3. (b) Yeast Gene Expressions after MST

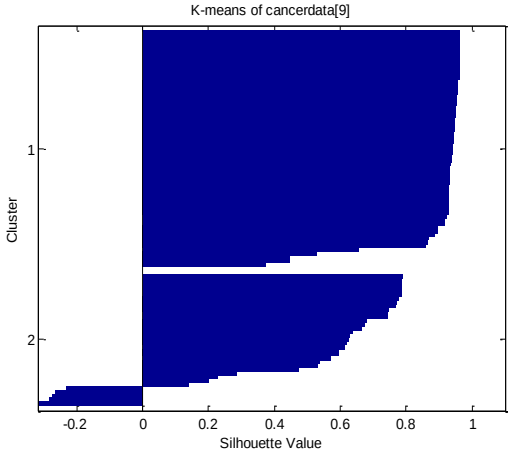

Fig 4. (a) K-means Clustering of Cancer data[9]

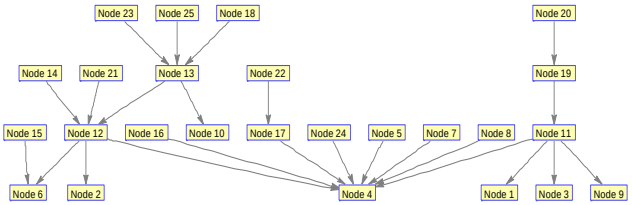

Fig. 4. (b) Part of MST of Cancer data profiles [9]

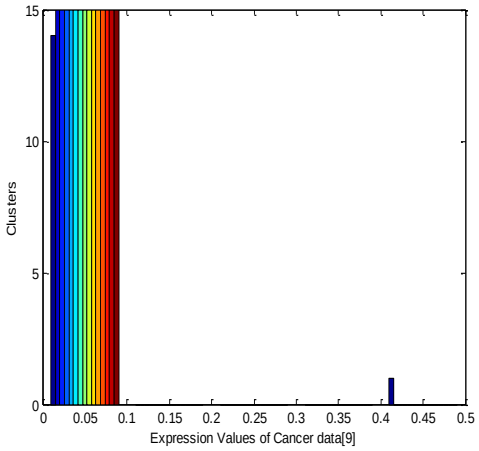

Fig. 4. (c) Clustering results of Cancer data[9]

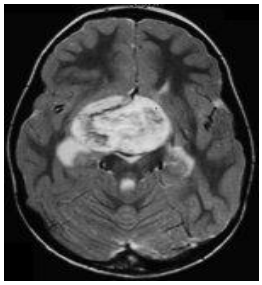

Fig. 5. (a) MRI image with lesions

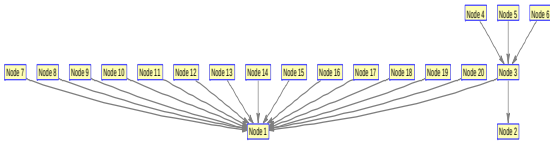

Fig.5. (b) MST of MRI image

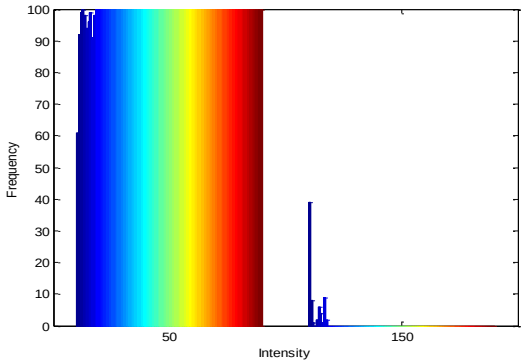

Fig. 5. (c) Clustering results of Fig.5. (a)

VI. CONCLUSION

MST is highly efficient, much flexible and is capable of detecting discriminatory patterns from medical images and produces high classification accuracy compared to other standard approaches. The clustering approach of this paper has been applied on microarray gene profiles as well as on the medical images to detect and group the similar patterns. The approach of this paper concentrates on outlier elimination first then clustering is obtained. To achieve automation in the clustering process the Iterative algorithm is applied on the MST and those clusters are validated by the measure. The overall performance of the approach has been improved in terms of automation, accuracy and efficiency compared to other methods. Experimental results show the proposed approach is efficient and competitive with existing clustering algorithms. Future work will focus on how to find the requirements of biomedical applications and to match those requirements with clustering results.

REFERENCES

[1] C.T. Zahn, Graph –theoretical methods for detecting and describing gestalt clusters,IEEE Trans.on Computers (1971), 68-86.

[2] Ying Xu, Victor Olman and Dong Xu, Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees., 2001.

[3] Aaron M Newman1, James B Cooper1,2, Newman and Cooper ,AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number, BMC Bioinformatics 2010.

[4] A. Brazma, J. Vilo, Minireview: gene expression data analysis, European Molecular Biology Laboratory, Outstation Hinxton—the European Bioinformatics institute, Cambridge CB10 ISD UK, 2000.

[5] K.Y. Yeung, W.L. Ruzzo, Principal component analysis for clustering gene expression data, Bioinformatics 17 (9) (2001) 763–774..

[6] Yu He and Lihui Chen, A threshold criterion, auto-detection and its use in MST-based clustering Department of Electrical and Electronic Engineering, Nanyang Technological University, Republic of Singapore, 639798,2002.
[7] Jana, P.K.; Naik, A.; An efficient minimum spanning tree based clustering Algorithm; Methods and Models in Computer Science, 2009. ICM2CS 2009.
[8] Jiang Qiang-rong; Gao Yuan; Spanning-Tree Kernels on Graphs ;Measuring Technology and Mechatronics Automation (ICMTMA), 2010
[9] Saravanan, R Mallika, **"**Statistical Based Classification Methods for Micro Array Gene Expression Data – A Survey**"**, International Journal of Computer Science and Knowledge Engineering, Volume 3, No 2, July-December 2009, pp 211-219. ISSN: 0973-6735.
[10] Limin Fu and Enzo Medico FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data, BMC Bioinformatics, January 2007.
[11] James C. Bezdek, Robert Ehrlich and William Full, FCM: The fuzzy c-means clustering algorithm Computers & Geosciences ,Volume 10, Issues 2-3, 1984.
[12] Jason Ernst1,Gerard J. Nau and Ziv Bar-Joseph, Clustering short time series gene expression data, Bioinformatics Journal(January,2005).

LATHA PARTHIBAN received her B.E in Electronics and Communication Engineering from Madras University, M.E in Computer Science and Engineering from Anna University, Chennai and PhD in Computer Science and Engineering from Pondicherry University. Currently she is with Computer Science department of Pondicherry University Community College. Her research interests are in Medical Image Processing, Data mining and Computer aided medical Diagnosis.