Spatial Expectation Mechanism Framework for Micro Array Analysis

D. Rajkumar
Computer Science and Engineering,
Dhanalakshmi Srinivasan College of Engineering,
Perambalur, India.

P. Ayesha Barvin M.E.,M.B.A.
Assitant.Professor.
Computer Science and Engineering,
Dhanalakshmi Srinivasan College of Engineering,
Perambalur, India.

***Abstract*:- Microarray technology is one of the important biotechnological that allows recording the expression levels of thousands of genes simultaneously within a number of different samples. A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample, and each entry of the matrix is the measured expression level of a particular gene in a sample, respectively. Hence, one of the major tasks with the gene expression data is to find groups of co-regulated genes whose collective expression is strongly associated with the sample categories or response variables. So implement feature subset selection approach to reduce dimensionality, removing irrelevant data and increase diagnosis accuracy and presents learning method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data using Spatial EM algorithm. An important finding is that the proposed semi supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. The experimental results prove that Spatial EM based classification approach provides improved accuracy rate in disease diagnosis.**

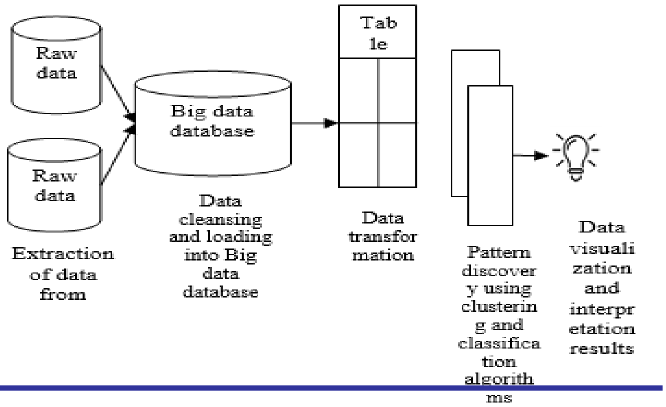
Keywords— Spatial EM, Clustering, Classification

I. INTRODUCTION

Classification and clustering are two major tasks in gene expression data analysis. Classification is concerned with assigning memberships to samples based on expression patterns, and clustering aims at finding new biological classes and refining existing ones. To cluster and/or recognize patterns in gene expression datasets, dimension problems are encountered. Typically, gene expression datasets consist of a large number of genes (attributes) but a small number of samples (tuples). Many big data algorithms (e.g., classification association rule mining, pattern discovery, linguistic summaries and context-sensitive fuzzy clustering are developed and/or optimized to be scalable with respect to the number of tuples, so as not to handle a large number of attributes. To apply existing clustering algorithms to genes, various algorithms have been used. Well-known examples are: k-means algorithms, self-organizing maps (SOM) and various hierarchical clustering algorithms. As for distance measures, Euclidean distance and Pearson's correlation coefficient are widely used for clustering genes. The genes regarded as similar by Euclidean distance may be very dissimilar in terms of their shapes or vice versa. It considers each gene as a random variable with n observations and measures the similarity between the two genes by calculating the linear relationship between the distributions of the two corresponding random variables. Recently, Spatial EM algorithms have been proposed to cluster both genes and samples simultaneously. Spatial EM algorithms aim at identifying subsets of genes and subsets of samples by performing simultaneous clustering of both rows and columns of a gene expression table instead of clustering columns and rows (genes and samples) separately. Specifically, these algorithms group a subset of genes and a subset of samples into a matrix such that the genes and samples exhibit similar behavior. Based on similar behavior, diseases can be classified using KNN classification techniques and evaluate the performance of the system.

II. BIG DATA WITH CLUSTERING

Big data (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems The overall goal of the Big data process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Generally, big data (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Technically, Big data is the process of finding correlations or patterns among dozens of fields in large relational databases.


Raw data
Raw data
Extraction of data from
Big data database
Data cleansing and loading into Big data database
Tab le
Data transfor mation
Pattern discover y using clusterin g and classifica tion algorith ms
Data visuali zation and interpr etation results

III. GENE EXPRESSION DATA.

Gene expression data is obtained by extraction of quantitative information from the images/patterns resulting from the readout of fluorescent or radioactive hybridizations in a microarray chip. Usually, gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column. Each element of this matrix represents the expression level of a gene under a specific condition, and is represented by a real number, which is usually the logarithm of the relative abundance of the mRNA of the gene under the specific condition.

Several obvious aims of these data analysis are the following:

1. Identify genes whose expression levels reflect biological processes of interest (such as development of cancer).
2. Group the tumors into classes that can be differentiated on the basis of their expression profiles, possibly in a way that can be interpreted in terms of clinical classification. For example, one hopes to use the expression profile of a tumor to select the most effective therapy.
3. Finally, the analysis can provide clues and guesses for the function of genes (proteins) of yet unknown role.

A microarray experiment typically assesses a large number of DNA sequences (genes, CDNA clones, or expressed sequence tags under multiple conditions. These conditions may be a time series during a biological process (e.g., the yeast cell cycle) or a collection of different tissue samples (e.g., normal versus cancerous tissues) and focus on the cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called “genes”. Similarly, will uniformly refer to all kinds of experimental conditions as “samples” if no confusion will be caused. A gene expression data set from a microarray experiment can be represented by a real-valued expression matrix $M = \{w_{i,j} | 1 \le i \le n, 1 \le j \le m\}$ where the rows $(G = \{\vec{g_1}, \ldots, \vec{g_m}\})$ form the expression patterns of genes, the columns $(S = \{\vec{s_1}, \ldots, \vec{s_m}\})$ represent the expression profiles of samples, and each cell w_{ij} is the measured expression level of gene i in sample j. Figure includes some notation that will be used in the following sections.

IV SPATIAL EM ALGORITHM:

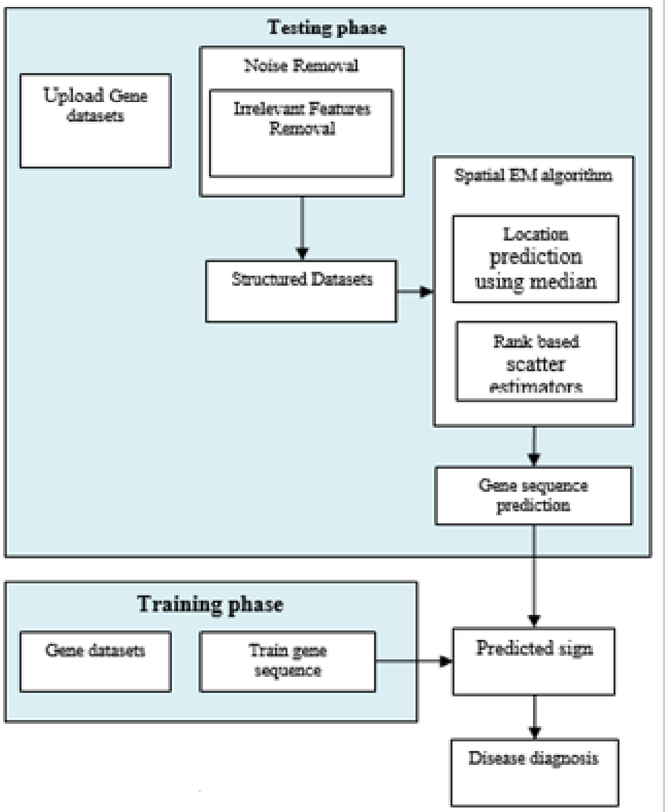
Spatial-EM modifies the component estimates on each M-step by spatial median and rank covariance matrix to gain robustness at the cost of increasing computational burden and losing theoretical tractability. Pseudo code of the algorithm is described as:

```
Initialization t = 0, $\mu_j, \Sigma_j = I, \tau_j = \frac{1}{K}\ for\ \forall_j$
Do until $\tau_j^t\ coverage\ for\ all\ j$
For j=1 to K
E-Step: Calculate $T_{ji}^t$
M-Step: Update $\tau_j^{t+1}$
Define$W_{ji}^t$,
Find $\mu_j^{t+1}$,
Find$(\Sigma_j^{t+1})^{-1}\ and\ (\Sigma_j^{t+1})^{-1/2}$
End
t=t+1
End
```

In spatial algorithm can first calculate the maximum coverage of data and then initialize all variables and perform Expectation and Maximization steps as in EM algorithm. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

EXISTING METHODS

With the use of EM framework to analysis the microarray gene to identify the disease and performing the clustering technique to group the gene to put in one term and ranking the predicate sign with the help of scatter based matrix, the accuracy of the finding diseases is 75% only. A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample, and each entry of the matrix is the measured expression level of a particular gene in a sample, respectively.



PROPOSED SYSTEM.

The proposed system is spatial expectation mechanism for micro array analysis. In this, I include the classification algorithm to identify the stage of the diseases. Analysis the Gene pattern with the help of predefined diseases patterns. I provide the accuracy of 80-90% of in my project. The experimental results prove that Spatial EM based classification approach provides improved accuracy rate in disease diagnosis.

V. CONCLUSION

Recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel. Gene expression data generated by microarray experiments offer tremendous potential for advances in molecular biology and functional genomics. This paper reviewed both classical and recently developed clustering algorithms, which have been applied to gene expression data, with promising results. The proposed semi-supervised spatial EM clustering algorithm is based on measuring mean values and scatter matrix using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised EM gene selection algorithm with accuracy rate. An important finding is that the proposed semi-supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

REFERENCES

[1] S. Bashir and E. M. Carter, “High breakdown mixture discriminant analysis,” J. Multivariate Anal., vol. 93, no. 1, pp. 102–111, 2005.

[2] C. Biernacki, G. Celeux, and G. Govaert, “An improvement of the NEC criterion for assessing the number of clusters in a mixture model,” Pattern Recognit. Lett., vol. 20, pp. 267–272, 1999.

[3] B. Brown, “Statistical uses of the spatial median,” J. Roy. Stat. Soc., B, vol. 45, pp. 25–30, 1983.

[4] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, “Knowledge-based analysis of microarray gene expression fidata by using support vector machines,” Proc. Nat. Acad. Sci., vol. 97, no. 1, pp. 262–267, 2000.

[5] N. A. Campbell, “Mixture models and atypical values,” Math. Geol., vol. 16, pp. 465–477, 1984.

[6] G. Celeux and G. Soromenho, “An entropy criterion for assessing the number of clusters in a mixture model,” Classification J., vol. 13, pp. 195–212, 1996.

[7] Y. Chen, Bart H. Jr, X. Dang, and H. Peng, “Depth-based novelty detection and its application to taxonomic research,” in Proc. 7th IEEE Int. Conf. Data Mining, Omaha, Nebraska, 2007, pp. 113–122.

[8] Y. Chen, X. Dang, H. Peng, and H. Bart Jr., “Outlier detection with the kernelized spatial depth function,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 288–305, Feb. 2009.