ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue 1 Ver 2 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Interactive Supervised Attribute Clustering for Microarray Sample Classification

[I]N.Sakthi Devi, [II]Ms. S.Banupriya, [III]G.Parvathi
[I,III]M.E. student, [II]Assistant Professor
[I,II,III]Computer Science And Engineering, Srinivasan Engineering College
E-mail : [I]nirmaladevi254@gmail.com, [II]banupriyasubbiah@gmail.com, [III]Parvathimit39@gmail.com

Abstract
Effective identification of co expressed gene and coherent patterns in gene expression data is an important task in biomedical applications. Many clustering method have been proposed to identify co expressed genes that share similar coherent index pattern. However there is no objective standard for groups of co expressed gene. Further groups of co expressed genes in gene expression data are often highly connected through a large number of intermediate genes. So that there may be no clear boundaries to separate the clusters. Clustering the number of gene expression data will also face some challenges of satisfying biological domain requirements and addressing the high connectivity of the data sets. Here we propose an interactive framework for exploring coherent pattern in gene expression data. An novel coherent index pattern is proposed to give users highly confident indications of the existence of coherent patterns. To drive a coherent pattern index and facilitate clustering we devise an attraction tree structure that summarizes the coherence information among genes in the data sets. We present efficient and scalable algorithms for constracting attraction trees and coherent pattern indices from gene expression data sets. Our experimental result shows that our approach is effective in mining gene expression data and is scalable for mining large data sets.

Key Words
Bioinformatics, Gene expression data, Clustering, Interactive Data Mining.

I. Introduction

Microarray technology can simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. An important task of analyzing gene expression data is the detection of coexpressed genes and coherent pattern index. Clustering is used to identify the coexpressed gene and coherent pattern index. Coexpressed genes may belong to the same or similar functional categories and indicate co-regulated families or similar families, and coherent patterns may characterize important cellular processes and suggest the regulating mechanism in the cells. Clustering is an important topic in data mining research is given in an matrix table, a conventional clustering algorithm groups tuples, into a set of attributes, and into clusters based on similarity. Intuitively, tuples in a cluster are more similar to each other than those belonging to different clusters. Clustering is very useful in many data mining applications. When applied to gene expression data analysis, conventional clustering algorithms encounter the problem related to the nature of gene expression data which is normally "wide" and "shallow." In another words, data sets usually contain a huge number of genes (attributes) and a small number of gene expression profiles (tuples). This characteristic of gene expression data often compromises the performance of conventional clustering algorithms. This paper, present a methodology to group attributes that are interdependent or correlated with each other. It refer to such a process as attribute clustering. In this sense, attributes in a cluster are more correlated with each other whereas attributes in different clusters are less correlated. Attribute clustering are used to reduce the search dimension of a data mining algorithm to effectuate the search of interesting relationships or for construction of models in a tightly correlated subset of attributes rather than in the entire attribute space. After attributes are clustered, user can able to select a small number for further analysis.

A. 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. Usually, gene expression data is arranged in an data matrix, where each row corresponds to one gene and each column corresponds to cell of the gene. Each element of this matrix represents the expression level of a gene.

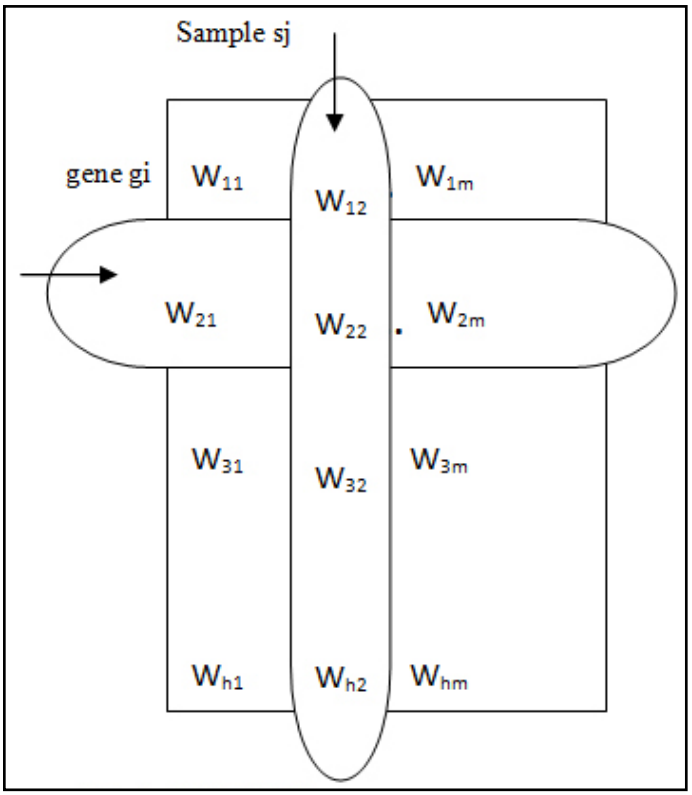


Fig 1: Matrix Table

Gene expression matrices have been extensively analyzed in two dimensions the gene dimension and the condition dimension. These analysis correspond to analyze the expression patterns of genes by comparing the rows in the matrix, and to analyze

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

the expression patterns of samples by comparing the columns in the matrix. Several obvious aims of these data analysis are the following.

1) Identify genes whose expression levels reflects the 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. In another way that can be interpreted in terms of clinical classification. For example user can use the expression profile of a tumor to select the most effective one.
3) Finally, the analysis can able to provide clues and guesses for the function of genes (proteins) of yet unknown role.

A microarray experiment typically assess a large number of DNA sequences (genes, cDNA clones, or expressed sequence tags [ESTs]) under multiple conditions. These conditions may be a timeseries during a biological process (e.g., the yeast cell cycle) or a collection of different tissue samples (e.g., normal versus cancerous tissues). In this paper, we will focus on the cluster analysis of gene expression data without making a distinction among DNA sequences, which will uniformly be called "genes". Similarly, we will uniformly refer to all kinds of experimental conditions as "samples" if no confusion will be caused.

B. Techniques Used

Supervised Clustering

We focus on microarray data where experiments monitor gene expression in different tissues and where each experiment is equipped with an additional response variable such as a cancer type and brain tumor etc. So that the number of measured genes is in thousands, and also it is assumed that only a few marker components of gene subsets determine the type of a tissue in the cell. In this paper we present a new method for finding such groups of genes by directly incorporating the response variables into the grouping process, by yielding a supervised clustering algorithm for genes.

II. Related Work

A. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns

Large gene expression studies, such as using DNA arrays, It provide millions of different pieces of gene expression data. To address the problem of analyzing such data, then describe a statistical method, which they have called 'gene shaving'.This method identifies subsets of genes with coherent expression patterns and large variation across conditions. Gene shaving method which differs from hierarchical clustering [4] and other widely used methods for analyzing gene expression studies in that genes may belong to one or more cluster,The technique can be 'unsupervised', [14] that is, the genes and samples or cells are treated as unlabeled, or fully supervised by using known properties of the genes or samples to assist in finding meaningful groupings.

B. Minimum Redundancy Feature Selection from Microarray Gene Expression Data

Discriminant analysis is now widely used in bioinformatics, such as distinguishing cancer tissues or tumour tissues from normal tissues or one cancer subtype vs another, predicting protein fold or super-family from its sequence, etc. A critical issue in discriminant analysis is feature selection instead of using all available variables (features or attributes) in the data.This article propose a minimum redundancy – maximum relevance (MRMR) feature selection framework.[5] Genes selected via MRMR provide a more balanced coverage of the space and capture broader characteristics of phenotypes.

C. Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data

One important application of gene expression microarray data is classification of samples or genes into categories, such as the type of cancer and tumor. The use of microarrays allows simultaneous monitoring of thousands of genes expressions per sample. [6]This method can also used to measure gene expression data or profile has resulted in data with the number of variables P (genes) far exceeding the number of samples *N*. Standard statistical methodologies in classification and prediction do not work well or even at all when N<P.Modification of existing statistical methodologies or development of new methodologies is needed for the analysis of microarray data and propose a novel[8]analysis procedure for classifying (predicting) human tumor samples based on microarray gene expressions. This procedure involves dimension reduction using Partial Least Squares (PLS) and classification using Logistic Discrimination (LD)

III. System Design

First upload the dataset.Dataset may be microarray data. That is converted into values. Data pre-processing contains cleaning, normalization, transformation, feature extraction and selection, etc. Identify the two types of gene selection such as occurrence based selection and sequence based selection. In occurrence based selection, we provide the separate gene and show all gene which is provided by users. Then in sequence based selection, we provide sequence and identify all sequences with position information.Three types of periodic patterns are present in time series they are symbol,sequence and segment periodic.

Symbol Periodicity: A Time-Series is supposed to be a symbol periodicity, if collected gene contain no less than one symbol occurs then it is said to be as symbol periodicity respectively.

Sequence Periodicity: A Time-Series is supposed to be a Sequence Periodicity, if more than one symbol might be occur in the gene data setthen it is termed as limited periodic patterns.To apply existing clustering algorithms to genes, some of the algorithm used are(Well known examples)k-means algorithms, self-organizing maps (SOM) and various hierarchical clustering algorithms.

Segment Periodicity : A Time-Series is supposed to be a Segment Periodicity, if the entire gene data set is typically symbolized as a replication of a model.The proposed supervised attribute clustering method uses this measure to reduce the redundancy among genes.

The issues of selecting a 'good' clustering method and determining the 'correct' number of clusters are reduced to model selection problems in the probability framework.

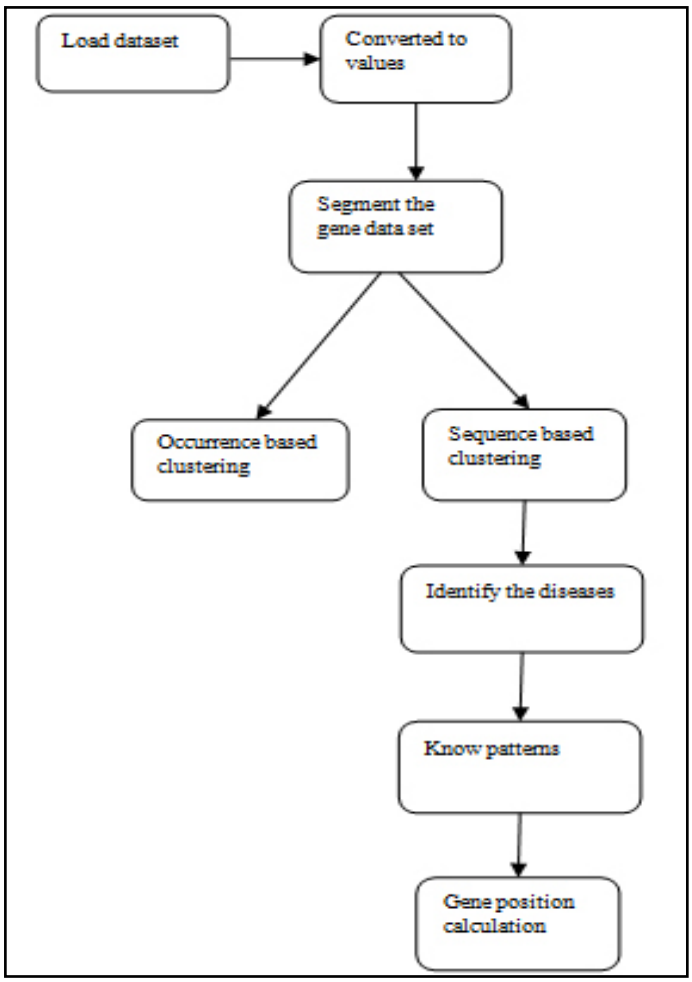
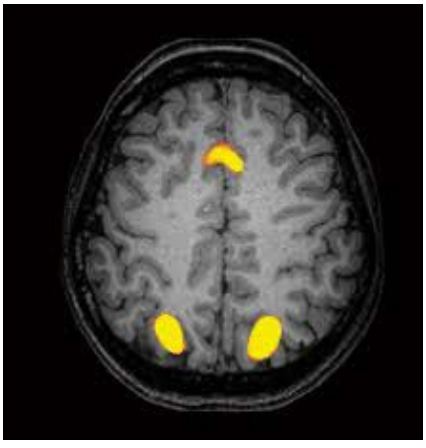


Fig. 2: Architecture Diagram

IV. Datasets

Data sets are the sample gene collected from the human body. The datasets are splited into several small number of clusers it is consider as coexpressed gene and Intermediate genes.so there is no absolute standard.The data sets is also splited into several large clusters It also contains both coexpressed gene and intermediate genes.By using the collected data sets we are identifying the diseases occurred in the human body. Sample dataset collected is given below.We can able to reduce the redundancy among the data sets and also noise datasets are removed.In enhancement phase we are using the image as dataset.

Sample Datasets



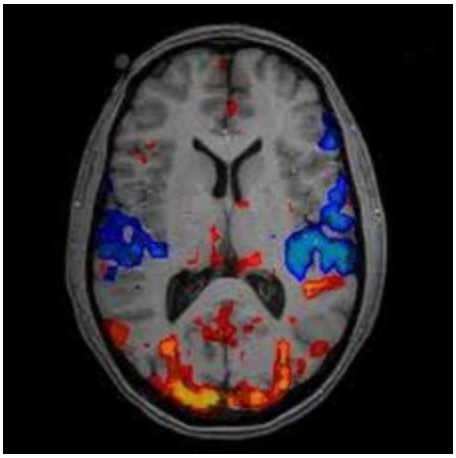


Fig 3: Image dataset

V. Proposed Work

In real data analysis, one of the important issues is computing both relevance and redundancy of attributes by discovering dependencies among them.The proposed supervised attribute clustering algorithm relies on mainly two factors, namely, determining the relevance of each attribute and growing the cluster around each relevant attribute incrementally by adding one attribute after the other. One of the important properties of the proposed clustering approach is that the cluster is augmented by the attributes those satisfy following two conditions:

1. Suit best into the current cluster in terms of a supervised similarity measure defined above.
2. Improve the differential expression of the current cluster most, according to the relevance of the cluster representative or prototype

Upload Dataset

We upload the datasets. The dataset may be microarray dataset. A microarray database is containing microarray gene expression data. The user uses the microarray database to store the measurement of data, and also to manage the searchable index, and make the data available to other applications for analysis and interpretation.

Preprocessing

Data pre-processing is an important step in the data mining process. The representation and quality of data is first checked before running an analysis.If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery and maintaining the data during the training phase is more difficult.

Pattern Evaluation

In this module we identify the two types of gene selection such as occurrence based selection and sequence based selection. In occurrence based selection, we provide the separate gene and show all gene which is provided by users. In sequence based selection, we provide sequence and identify all sequences with position information.

Clustering approach

The proposed supervised attribute clustering algorithm relies on mainly two factors, namely, determining the relevance of each attribute and growing the cluster around each relevant

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

attribute incrementally by adding one attribute after the other. A new supervised attribute clustering algorithm is proposed to find co regulated clusters of genes whose collective expression is strongly associated with the sample categories or class labels. A new quantitative measure, based on mutual information, is introduced to compute the similarity between attributes.The proposed supervised attribute clustering method uses this measure to reduce the redundancy among genes.

Coherent index selection

We can identify the gene positions. Then finding good clustering configurations which contain interdependence information within clusters and discriminative information for classification.

Evaluation criteria

The performance of the proposed supervised attribute clustering algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on five microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separability index and classification accuracy of naive bayes classifier, K-nearest neighbor rule, and support vector machine. To compute the classification accuracy, the leave one-out cross validation is performed on each gene expression data set.

VI. Evaluation Result

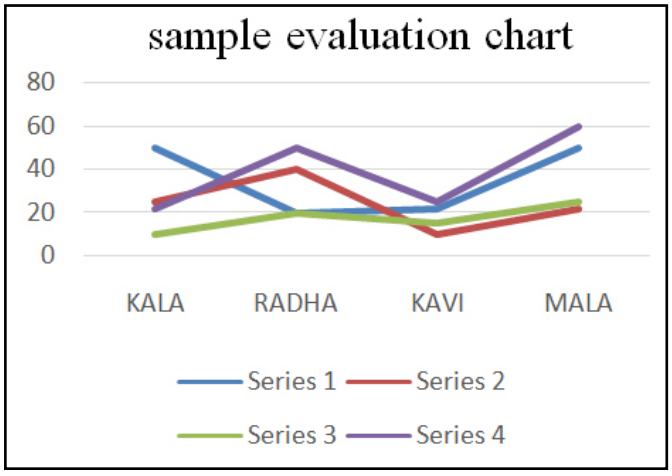


Fig 4: Evaluation chart

VII. Conclusion

Recent DNA microarray technologies monitor transcription levels of tens thousands of genes in parallel.This project reviewed both classical and recently developed clustering algorithms, which have been applied to gene expression data, with promising results. The proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed.An important finding is that the proposedsupervisedattribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

References

[1] *U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Array," Proc. Nat'l Academy of Sciences USA, vol. 96, no. 12, pp. 6745-6750.*

[2] *M. Ankerst, M.M. Breunig, H.P. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure," Proc. SIGMOD, pp. 49-60.*

[3] *Z. Bar-Joseph, E.D. Demaine, D.K. Gifford, N. Srebro, A.M. Hamel, and T.S. Jaakkola, "K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data," Bioinformatics, vol. 19, no. 9, pp. 1070- 1078, 2003.*

[4] *A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns," J. Computational Biology, vol. 6, nos. 3-4, pp. 281-297.*

[5] *M. Blatt, S. Wiseman, and E. Domany, "Super-Paramagnetic Clustering of Data," Physical Rev. Letters, vol. 76, 1996.*

[6] *Y. Cheng and G.M. Church, "Biclustering of Expression Data," Proc. Eighth Int'lConf. Intelligent Systems for Molecular Biology (ISMB), vol. 8, pp. 93-103, 2000.*

[7] *R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis, "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle," Molecular Cell, vol. 2, no. 1, pp. 65-73, July 2010*

[8] *M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc. Nat'l Academy of Sciences USA, vol. 95, no. 25, pp. 14863-14868.*

[9] *M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proc. Second Int'l Conf. Knowledge Discovery and Data Mining, pp. 226-231.*

[10] *C. Fraley and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," The Computer J., vol. 41, no. 8, pp. 578-588.*