IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 3 (Mar. - Apr. 2013), PP 16-19
www.iosrjournals.org

Gene Selection for Sample Classification in Microarray: Clustering Based Method

K. Mangala Prabin Libi
Department of information technology/Rajalakshmi engineering college/Anna university, Thandalam, Chennai.

Abstract: *Micro array technology is one of the important biotechnological means that allows recording the expression levels of thousands of genes simultaneously within a number of different samples. An important application of micro array gene expression data is to classify samples according to their gene expression profiles. The gene expression dataset can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample. The relevance of each attribute (attribute represents the gene expression conversion into numerical values) with respect to the class label and the redundancy between two attributes in terms of mutual information are calculated using supervised similarity measure. The proposed system uses supervised attribute clustering algorithm which determines the relevance of each attribute and growing the cluster around each relevant attribute by adding one attribute after the other. Min-hash algorithm is used to reduce the redundancy between the genes and also reduce the cluster size. The performance of the system can be improved by reducing the redundancy of genes.*

Key words: *Micro array, Gene expression, Mutual Information, Attribute Clustering, Supervised methods.*

I . Introduction

The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. Besides the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structures, visualization, and online updating. Kaijun Wang et al[1], proposed an Evolution for Cluster Analysis of Gene Expression Data , the system evolution method is to estimate the nature of clusters based on partitioning around mediods (PAM) clustering algorithms. It will be stable when natural clusters are found. It is easier to use. It provides intuitive information about separable degrees between clusters. It will judge separability of twin clusters and slightly overlapping between clusters. But it is difficult to estimate NC when overlapping clusters exist. Not applicable to the case of heavily overlapping clusters. PAM not suitable for irregular clusters shape such as circles or embedded within each other. Patrick C. H et al[2], they studied on clustering algorithms to inherent clusters in gene expression. Typically, the gene expression data is characterized by lot of noise, so they proposed an evolutionary clustering algorithm. It encodes an entire cluster grouping in a chromosome so that each gene in the chromosome encodes one cluster. It does not require the number of clusters to be decided in advance. Pattern hidden in each cluster can be explicitly revealed. Evo cluster is very robust in presence of noise. Evo clustering algorithm is a time consuming because number of clusters to be formed are not needed.

L.Wang et al[3], proposed an "Accurate Cancer Classification Using Expressions of Very Few Genes", they addressed to find small set of genes. The effective method is used to select small set of genes which involves two steps. First, choose some important genes using ranking scheme. Second, test the classification capability of all combinations of genes by using classifier (fuzzy neural network, SVM). Reduce computational burden and noise arising from irrelevant genes. Simplifies gene expression tests to include only a very small number of genes.

Shuanhu et al[4], studied an "Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning" they addresses to find the natural clusters in the data and estimate the correct number of clusters by using OPTOC(One-Prototype-Take-One-Cluster) algorithm and cluster splitting and merging strategy. It can be viewed as top-down process. Loose clusters are split into two clusters until pre-specified number of clusters is obtained. Merging attempts to merge similar clusters together. It can find natural clusters i.e.) number of clusters is less than the natural clusters in the data. Final partition of the dataset is not sensitive. It is very difficult to estimate reliably the correct number of cluster. It find only for minimum number of gene combinations.

C. Tang et al[5], "Expression Data: A Survey", they divide the cluster analyze for gene expression data in three categories, then relate each clustering category (gene based clustering, sample based clustering, subspace clustering). In which partitions the gene set into the subgroups each of which should be as homogeneous as possible. Genes in a cluster are more correlated with each other, whereas genes in different

cluster are less correlated. By using classification algorithm it will reduce a set of genes can be selected for analysis.

A.K.C. Wong et al [6], "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," presents an attribute clustering method which is able to group gene based on their interdependence. It can be used for gene grouping, selection and classification. Grouping of genes based on attribute interdependence within group helps to capture different aspects of gene association patterns in each group. Selected genes from each group contain information for gene expression classification and identification. Small data sets are used for results. W. Haiying, et al [7], addresses two algorithms self-organizing map and hierarchical clustering for SAGE (Serial Analysis of Gene Expression) data analysis. SOM perform pattern discovery and visualization for SAGE data in a more effective way. Estimation of optimal number of clusters hidden in SAGE data. Subsets of genes are interdependent with each other. Identifies set of related genes with similar samples yields tree like structure, makes identification of functional groups very difficult. B.W. Futscher et al[8], focuses on optimal search based subset selection methods because they evaluate the group performance of genes. They introduces tabu search to gene selection from high dimensional gene array data. Generate candidate gene subsets, assess based on evaluation criteria. Gene subset with the highest score is regarded as the optimal. It randomly picks the genes and evaluates solutions.

P. Majl[9], addresses f-information measures that may be suitable for selection of genes from gene expression data. It will evaluate the gene selection problem. It is used to measure the relevance of genes with respect to class labels or sample categories. More effective to evaluate the gene class relevance as well as gene-gene redundancy. T. Hastie, et al[10], proposes a method for supervised learning from gene expression data called as "tree harvesting". This includes hierarchical clustering of genes. The component of the model will be convenient for interpretation. By using clusters as inputs, bias input towards correlated set of genes. Do not use any similarity measure to cluster genes; rather use different predictive scores.

II. Proposed Methodology

GENE DATA SETS

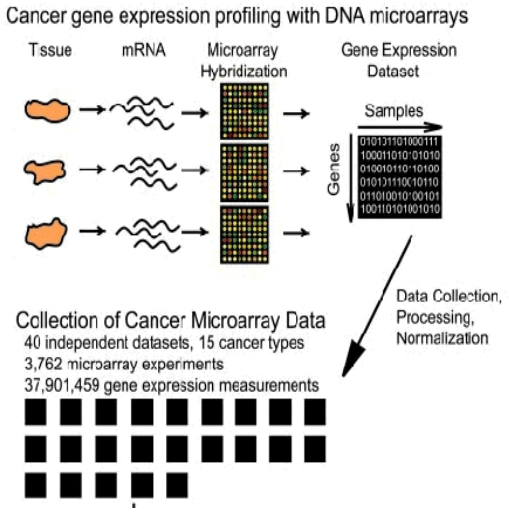

Fig: 1 gene data sets

Architecture

Figure 2 represents the architecture diagram of the system. In data client, the data acquisition takes place. Database is an repository which stores all the details of datasets. In data server, first data sets are pre-processed it includes data cleaning, normalization and testing. Here, the attributes are measured to calculate the relevance between the gene attributes. For this threshold has to be set manually. Based on the threshold values the attributes are grouped. Then the attributes are selected to form cluster. Representative attribute has to be chosen to form the finer cluster. It has to be chosen based on the maximum matches. That particular gene can be stated as representative attribute. After forming finer cluster sometimes the gene may be repeated. So, the repeated gene can be eliminated by using Min-Hash algorithm. Because of the repeated genes the cluster size is big. So redundant attributes are eliminated.

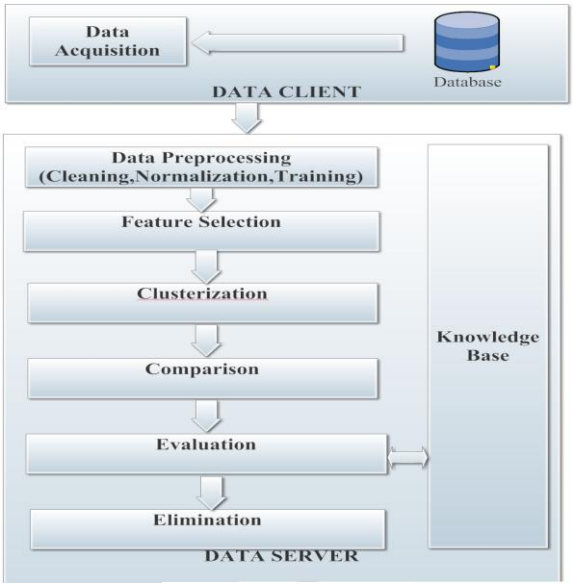

Fig:2 Proposed model

III. An Overview Of Proposed Solution

Module 1:

Gene expression profile data sets are collected. The expression profile consists of number of genes under different samples. Now Each and every sample is assigned with unique class labels known as sample categories.

Module 2:

First of all, expression profile of a gene (attribute) is taken into account. It is compared with all the class labels in data set and the relevance value is measured. Similarly all the attributes are compared with all the class labels and their relevance value is measured. Likewise supervised similarities between different attributes are calculated. Here the information of sample categories is also considered while calculating similarity between attributes. Now select a default attribute (A_i) which has higher relevance value to all the class labels.

Module 3:

Now from the whole data set, only the attributes (relevance measured) that are similar to A_i are chosen and then clustered. This cluster is known as coarse cluster and A_i is known as representative of the cluster. To increase the relevance value of A_i, only subsets of attributes are chosen from coarse cluster. Their expression profiles are then merged with A_i to increase its relevance value. i.e. the relevance value is only increased if the attributes of same sample categories are merged. These subsets of attributes are grouped as finer cluster. Hence the coarse cluster size is reduced. Similarly numbers of clusters are formed.

Module 4:

Now Min hash clustering algorithm is applied to finer clusters to find duplicate genes. This algorithm eliminates duplicate genes from all the finer clusters thereby refined clusters are formed. This will reduce the size of clusters.

VI. Conclusion

The system has been designed for train the gene data sets. More over the system has been designed for reducing the cluster size by eliminating the redundant attributes**.** The analysis on the requirements and a design for the proposed system has been screened. The requirement analysis process includes learning and determining about the working environment, technical requirements and logical aspects or features of the system. The design of the system has been sketched out using this analyzed information. The design has been implemented and applicable to certain changes when it is required in order to develop a prototype for the proposed work.

Acknowlegements

I would like to thank Mrs. J. Jeya Lakshmi, assistant professor, information technology, Rajalakshmi engineering college, Chennai for guiding me in writing this paper. Also I am grateful to prof. S. Poonkuzhali, Head of the department, Rajalakshmi engineering college, Chennai for her encouragement and motivation.

References

[1] Kaijun Wang, Jie Zheng, Junying Zhang, Member, IEEE, and Jiyang Dong "Estimating the Number of Clusters via System Evolution for Cluster Analysis of Gene Expression Data" IEEE transactions on information technology in biomedicine, vol. 13, no. 5, September 2009.

[2] Patrick C. H. Ma, Keith C. C. Chan, Xin Yao, Fellow, IEEE, and David K. Y. Chiu "An Evolutionary Clustering Algorithm for Gene Expression Microarray Data Analysis" IEEE trans on Evolutionary Computation, vol. 10, no. 3, June 2006.

[3] L.Wang,F.Chu, and W.Xie,"Accurate Cancer Classification Using Expressions of Very Few Genes," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 1, pp. 40-53, Jan.-Mar. 2007.

[4] Shuanhu Wu, Alan Wee-Chung Liew, Member, IEEE, Hong Yan, Senior Member, IEEE, and Mengsu Yang, "Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning" IEEE Trans on information technology in bio-medical, vol. 8, no. 1, march 2004.

[5] D. JianCluster Analysis for Gene g, C. Tang, and A. Zhang, "Expression Data: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

[6] W.-H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 2, no. 2, pp. 83-101, Apr.-June 2005.

[7] W. Haiying, Z. Huiru, and A. Francisco, "Poisson-Based Self- Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 2,pp.163-175,Apr.-June 2007.

[8] J. Li, H. Su, H. Chen, and B.W. Futscher, "Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification," IEEE Trans. Information Technology in Biomedicine, vol. 11, no. 4, pp. 398-405, July 2007.

[9] P. Maji, "f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data," IEEE Trans. Biomedical Eng., vol. 56, no. 4

[10] T. Hastie, R. Tibshirani,D. Botstein and brown "Supervised Harvesting of Expression Trees," Genome Biology

SITES

[1] http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/ASPMgene/

[2] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[3] http://www.ebi.ac.uk/arrayexpress/browse.html

[4] http://www.ncbi.nlm.nih.gov/gds?term=brain%20cancer