ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

SURVEY ON SUPERVISED ATTRIBUTE CLUSTERING AND GENE CLUSTERING TECHNIQUES IN DATA MINING

L.Boopathi[1], D.Vijaybabu[2]

PG Scholar, CSE Dept, Erode Sengunthar Engineering College, Thudupathi, India [1]

Assistant Professor, CSE Dept, Erode Sengunthar Engineering College, Thudupathi, India [2]

Abstract: In this paper focuses on supervised attribute clustering and gene clustering technique in data mining. Clustering is an important task of explorative data mining. Clustering is the process of grouping a set of objects into groups of similar objects. A good clustering method will produce high quality clusters in which intra class similarity is high and interclass similarity is low. Introduce the concepts of microarray technology and discuss the basic elements of clustering on gene expression data. This paper gives a brief introduction to clustering and various techniques have been discussed and a detailed study has been performed. The comparison table shows it clearly that a technique satisfies the clustering requirement.

Keywords: Clustering, Attribute Clustering Algorithm, Comparison of techniques, microarray technology, gene expression data

1. INTRODUCTION

Data Mining is the process of extracting information or patterns from large databases. In medical field, it is used to handle high dimensional data's. [1]Clustering is an important task of explorative data mining. Clustering is the process of grouping a set of objects into groups of similar objects. A good clustering method will produce high quality clusters in which intra class similarity is high and interclass similarity is low.

II.GENE EXPRESSION

Every organism is formed of cells and each cell contains a nucleus. The nucleus is formed of double stranded DNA molecule called Chromosomes. Gene is a DNA sequence located in a particular chromosome which encodes the information for the synthesis of proteins, which are the main structural and functional units of an organism's cell. Gene expression is the process by which the gene's coded information is converted into proteins as shown in Fig. 1 The formation of proteins from gene's coded information involves the formation of mRNA as an intermediary product. The level of mRNA or proteins (in case of protein synthesizing genes) produced for a gene at a particular time or an experimental condition is referred to as gene expression level. The expression level differs for different genes under different condition.

A.GENE EXPRESSION DATA

Gene expression data shows the expression levels of thousands of genes simultaneously under a condition. The microarray data or Gene Expression data is represented in the form of matrix X where the rows represents genes, columns represent conditions and each value X[i,j] shows the expression levels of gene under the jth condition. Monitoring the expression levels of thousands of genes simultaneously under a condition is called gene expression analysis or microarray data analysis.

B.GENE EXPRESSION PROFILING

Gene expression profiling is an emerging technology for identifying genes whose activity may be helpful in assessing disease prognosis and guiding therapy. Gene expression profiling examines the composition of cellular mRNA populations. The identity of the RNA transcripts that make up these populations and the number of these transcripts in the cell provide information about the global activity of genes that give rise to them.The number of mRNA transcripts derived from a given gene is a measure of the "expression" of that gene. Given that messenger RNA (mRNA) molecules are translated into proteins, changes in mRNA levels are ultimately related to changes in the protein composition of the cells, and consequently to changes in the properties and functions of tissues and cells in the body. Gene Expression Profiling has been applied to numerous mammalian tissues, as well as plants, yeast, and bacteria. These studies have examined the effects of treating cells with chemicals and the consequences of over expression of regulatory factors in transected cells. Studies also have compared mutant constraints with parental strains to delineate functional pathways. In the cancer research, such investigation has been used to find gene expression changes in transformed cells and metastases, to identify diagnostics markers, and to classify tumors based on their gene expression profiles.

C.GENE CLUSTER

A gene cluster is a set of two or more genes that serve to encode for the same or similar products. An example of a gene cluster is the Human β-globin gene cluster, which contains five functional genes and one non-functional gene which code for similar proteins. Haemoglobin molecules contain any two identical proteins from this gene cluster, depending on their specific role.Gene clusters are created by the process of genes duplication and divergence. A gene is accidentally duplicated during cell division, so that its descendants have two copies of the gene, which initially code for the same protein or otherwise have the same function. In the course of subsequent evolution, they diverge, so that the products they code for have different but related functions, with the genes still being adjacent on the chromosome. This may happen repeatedly.

Challenges of gene clustering

Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene-based clustering presents several new challenges and is still an open problem.

- First, cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis. For example, a clustering algorithm which can accurately estimate the "true" number of clusters in the data set would be more favored than one requiring the pre-determined number of clusters.
- Second, due to the complex procedures of microarray experiments, gene expression data often contain a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise.
- Third, our empirical study has demonstrated that gene expression data are often "highly con- nected" , and clusters may be highly intersected with each other or even embedded one in another. Therefore, algorithms for gene-based clustering should be able to effectively handle this situation.
- Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters (e.g., which clusters are more close to each other, and which clusters are remote from each other), and the relationship between the genes within the same cluster (e.g., which gene can be considered as the representative of the cluster and which genes are at the boundary area of the cluster). A clustering algorithm, whichcan not only partition the data set but also provide some graphical representation of the cluster structure would be more favored by the biologists.

D.MICROARRAY

The goal of microarray experiments is to identify genes that are differentially transcribed with respect to different biological conditions of cell cultures or tissue samples. With the development of the microarray technology, the necessary processing and analysis methods grow increasingly critical. It becomes gradually urgent and challenging to explore the appropriate approaches because of the large scale of microarray data comprised of the large number of genes compared to the small number of samples in a specific experiment.For the data obtained in a typical experiment, only some of genes are useful to differentiate samples among different classes, but many other genes are irrelevant to the classification. Those irrelevant genes not only introduce some unnecessary noise to gene expression data analysis, but also increase the dimensionality of the gene expression matrix, which results in the increase of the computational complexity in various consequent researches such as classification and clustering. As a consequence, it is significant to eliminate those irrelevant genes and identify the informative genes, which is a feature selection problem crucial in gene expression data analysis. With the development of microarray technology, DNA microarrays with millions of genes have been obtained. Finding the genes which are related to cancer is significant to medical treatment. There are various kinds of cancers. Each type of cancer may connect to different genes. Distinguishing classes of cancer based on gene expression levels has great importance on cancer diagnosis.

Measuring mRNA levels:

Compared with the traditional approach to genomic research, which has focused on the local examination and collection of data on single genes, microarray technologies have now made it possible to monitor the expression levels for tens of thousands of genes in parallel. The two major types of microarray experiments are the cDNA microarray and oligonucleotide arrays (abbreviated oligochip) . Despite differences in the details of their experiment protocols, both types of experiments involve three common basic procedures Chip manufacture: A microarray is a small chip (made of chemically coated glass, nylon membrane or silicon), onto which tens of thousands of DNA molecules (probes) are attached in fixed grids. Each grid cell relates to a DNA sequence. Target preparation, labeling and hybridization: Typically, two mRNA samples (a test sample and a control sample) are reverse-transcribed into cDNA (targets), labeled using either fluorescent dyes or radioactive isotopics, and then hybridized with the probes on the surface of the chip. The scanning process: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets. Generally, both cDNA microarray and oligo chip experiments measure the expression level for each DNA sequence by the ratio of signal intensity between the test sample and the control sample, therefore, data sets resulting from both methods share the same biological semantics. In this paper,unless explicitly stated, we will refer to both the cDNA microarray and the oligo chip as microarray technology and term the measurements collected via both methods as gene expression data.

E. CLUSTERING BASED ON SUPERVISED INFORMATIVE GENE SELECTION

The supervised approach assumes that phenotype information is attached to the samples, for example, the samples are labeled as diseased vs. normal. Using this information, a “classifier” which only contains the informative genes can be constructed. Based on this “classifier”, samples can be clustered to match their phenotypes and labels can be predicted for the future coming samples from the expression profiles. Supervised methods are widely used by biologists to pick up informative genes. The major steps to build the classifier include:

- Training sample selection. In this step, a subset of samples is selected to form the training set. Since the number of samples is limited (less than), the size of the training set is usually at the same order of magnitude with the original size of samples.

- Informative gene selection. The goal of informative gene selection step is to pick out those genes whose expression patterns can distinguish different phenotypes of samples. Forexample, a gene is uniformly high in one sample class and uniformly low in the other .

- A series of approaches to select informative genes include: the neighborhood analysis approach ; the supervised learning methods such as the support vector machine (SVM) [10], and a variety of ranking based methods [6, 43, 47, 49, 68, 70]. Sample clustering and classification. After about [24, 42] informative genes which manifest the phenotype partition within the training samples are selected, the whole set of samples are clustered using only the informative genes as features.

- Since the feature volume is relatively small, conventional clustering algorithms, such as K-means or SOM, are usually applied to cluster samples. The future coming samples can also be classified based on the informative genes, thus the supervised methods can be used to solve sample classification problem.

Benefits of Supervised Clustering:

This section briefly discusses the benefits of supervised clustering. In general, supervised clustering can be used for many different tasks that include:
1. Create background knowledge for a dataset
2. Dataset compression and editing
3. To learn subclasses and to use these subclasses to enhance classification algorithms
4. To evaluate distance functions in distance function learning
Due to space limitations we center on discussing the first two applications in this section.

III. RESEARCH ISSUES IN CLUSTERING AND GENE CLUSTERING TECHNIQUES

A. CLUSTER ANALYSIS OF GENE EXPRESSION DATA

Stephen Conrad Dinger (2011) examined the cluster analysis of gene expression data. The DNA microarray is a recently developed technology which enables biologists to measure gene expression profiles. Microarray experiments are developed for cancer patients for faster and more accurate diagnosis. A problem with cluster analysis is to determine the correct number of clusters in high dimensional data such as those obtained from microarrays. [4]The purpose here is to develop a clustering algorithm that can automatically determine the correct number of clusters whilst successfully clustering the data. The high dimensional space also remains a challenge since the distance metrics such as the Euclidean distance are not effective when the dimension increases. The common solution to the dimensionality problem involves reducing the dimensions of the data using Principal Component Analysis (PCA) and Isometric Mapping. The large amount of information embedded in the genome is hard to analyze statistically and is often compounded with noise from external factors like laboratory equipments. The sources of noises due to the laboratory procedure include: mRNA preparation, transcription, labeling, hybridization parameters and contaminants that affect the image analysis. So the classical approach to analyse microarrays involves defining the type of problem based on two different criteria: number of samples, a priori knowledge of the distribution. The statistics commonly used on the microarray data includes the F-statistics, the chi-square statistics and the Students t-statistics.The t-statistics is used on two different samples inorder to evaluate whether the samples have a distribution with the same mean. The t-statistics in the microarray analysis is used to determine whether the gene expression is stochastics or regulated between cancer and healthy patients. The Analysis of Variance (ANOVA) test is able to avoid the increase in error introduced from multiple group comparison. The ANOVA procedure determines whether a gene is statistically significant and therefore differentially expressed in any of multiple conditions tested.The two common procedures for achieving dimensions are called feature selection and feature extraction. In feature selection, a test is performed whereby the features (genes) that contribute the most to the class separability are chosen. The feature extraction deals with the linear or non-linear mapping of the dataset from the high dimensional feature space to lower-dimensional feature space.In supervised learning the class labels are supplied as training data set which is used to build up a model. The most common supervised learning techniques of T.Hastie are artificial neural networks, support vector machine and naive bayes. The various distance metrics inorder to measure the similarity between groups of genes or samples such as Euclidean distance, Manhattan distance, maximum distance, Minkowski distance are discussed here. The most common clustering algorithms are hierarchical clustering, k-means, Self-Organising map (SOM). To validate the cluster, the internal criteria, relative criteria and external criteria are used. The cluster validity indices such as Davies-Bouldin index, Dunn's Index, Calinski Harabasz Index and I index are also discussed.

B. ATTRIBUTE CLUSTERING

Wai-Ho Au presents an attribute clustering method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data. It can be used for gene grouping, selection, and classification. The partitioning of a relational table into attribute subgroups allows a small number of attributes within or across the groups to be selected for analysis. By clustering attributes, the search dimension of a data mining algorithm is reduced. The reduction of search dimension is especially important to data mining in gene expression data because such data typically consist of a huge number of genes (attributes) and a small number of gene expression profiles (tuples). Most data mining algorithms are typically developed and optimized to scale to the number of tuples instead of the number of attributes. The situation becomes even worse when the number of attributes overwhelms the number of tuples, in which case, the likelihood of reporting patterns that are actually irrelevant due to chances becomes rather high. It is for the aforementioned reasons that gene grouping and selection are important preprocessing steps for many data mining algorithms to be effective when applied to gene expression data. It defines the problem of attribute clustering and introduces a methodology to solve it.A new method is introduced so as to group interdependent attributes into clusters by optimizing a criterion function derived from an information measure that reflects the interdependence between attributes. By applying the Attribute Clustering Algorithm (ACA) to gene expression data, meaningful clusters of genes are discovered. The grouping of genes based on attribute interdependence within group helps to capture different aspects of gene association patterns in each group. Significant genes selected from each group then contain useful information for gene expression classification and identification. To evaluate the performance of this, we applied it to two well-known gene expression data sets and compared our results with those obtained by other methods. Finally the experiments show that this method is able to find the meaningful clusters of genes. By selecting a subset of genes which have high multiple-interdependence with others within clusters, significant classification information can be obtained. Thus, a small pool of selected genes can be used to build classifiers with very high classification rate. From the pool, gene expressions of different categories can be identified.

C. MOLECULAR CLASSIFICATION OF CANCER

Golub (1999) proposed the Classification of cancer for class discovery and class prediction by gene expression monitoring. Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

D. FUZZY ROUGH SUPERVISED ATTRIBUTE CLUSTERING ALGORITHM AND CLASSIFICATION

One of the main problems in gene expression data analysis is uncertainity.[9] Some of the sources of this uncertainity include incompleteness and vagueness in class definition. So, for this the possibility concept was introduced by fuzzy sets and rough sets which has gained popularity in modelling and propagating uncertainity. The generalized theories of rough-fuzzy sets and fuzzy-rough sets have been applied successfully to feature selection of real valued data, rough-fuzzy clustering. To cluster coexpresssed genes from microarray data, different fuzzy and rough-fuzzy clustering algorithms can be used. However these algorithms are unsupervised in nature, as the genes are clustered without using any information of class labels.Hence in diagnosis of cancer, we need the information of class labels or sample categories to be incorporated with it. Thus, an Supervised gene clustering Algorithm is required for better diagnosis. A new supervised gene clustering algorithm, termed as Fuzzy-Rough Supervised Attribute clustering (FR-SAC) is introduced by Pradipta Maji (2011) based on the fuzzy-rough sets. It finds the coregulated clusters of genes whose collective expression is strongly associated with the sample categories. A new quantitative measure based on fuzzy-rough sets is used to compute the similarity between genes. This measure incorporates the information of sample categories or class samples while measuring the similarity between genes. The FRSAC algorithm uses this measure to reduce the redundancy among genes. It involves the partitioning of the original gene set into some distinct subsets or clusters so that the genes within the clusters are highly coregulated with the strong association to sample categories, while those in different clusters are as dissimilar as possible.A single gene from each cluster having highest gene-class relevance value if first selected as the initial representative of that cluster. The representative of each cluster is then modified by averaging the initial representative with other genes of that cluster whose collective expression is strongly associated with the sample categories. Finally the modified representative of each cluster is selected to constitute the resulting reduced feature set. Using this, the diagnosis for cancer patients are made easily.

E. SUPERVISED CLUSTERING OF GENES

Microarray technology monitors the gene expression of different tissue samples, and where each experiment is equipped with an additional categorical outcome variable, describing e.g. a cancer type. An important problem in this setting is to study the relation between gene expression and tissue type. While microarrays monitor thousands of genes, it is assumed that only a few underlying marker components of gene subsets account for nearly all of the outcome variation, i.e. determine the type of a tissue. [2]The identification of these functional groups is crucial for tissue classification in medical diagnostics, as well as for understanding how the genome as a whole works. As a first approach, unsupervised clustering techniques have been widely applied to find groups of co-regulated genes on microarray data. Hierarchical Clustering identifies sets of correlated genes with similar behaviour across the experiments, but yields thousands of clusters in a tree-like structure. This makes the identification of functional groups very difficult. The unsupervised techniques usually fail to reveal functional groups of genes that are of special interest in tissue classification. This is because genes are clustered by similarity only, without using any information about the experiment's response variables.So **Marcel Dettling (2002)** focused on Supervised Clustering, defined as grouping of variables (genes), controlled by information about the Y variables, i.e. the tumor types of the tissues. One of the supervised techniques is Tree Harvesting. It is of a 2-step method which consists of generating numerous candidate groups by unsupervised hierarchical clustering. Then the average expression profile of each cluster is considered as a potential input variable for a response model and the few gene groups that contain the most useful information for tissue discrimination are identified. Only this second step makes the clustering supervised, since the selection process relies on external information about the tissue types. An interesting supervised clustering approach that directly incorporates the response variables Y in the grouping process is the Partial Least Squares (PLS) procedure an often applied tool in the chemometrics literature, which in a supervised manner constructs weighted linear combinations of genes that have maximal covariance with the outcome. PLS has the drawback that the fitted components involve all (usually thousands of) genes, which makes them very difficult to interpret.A promising new method for searching functional groups, each made up of only a few genes whose consensus expression profiles provides useful information for tissue discrimination. This supervised algorithm can be started with or without initial groups of genes, and then clusters genes in a stagewise forward and backward search, as long as their differential expression in terms of our objective function can be improved. This yields clusters typically made up of 3–9 genes, whose coherent average expression levels allow perfect discrimination of tissue types. Thus the clusters show excellent out-of-sample predictive potential, and permutation and randomization techniques show that they are reasonably stable and clearly more than just a noise artifact. The output of our algorithm is thus potentially beneficial for cancer type diagnosis. At the same time it is very accessible for interpretation, since the output consists of a very limited number of clusters, each summarizing the information of a few genes. Thus, it may also reveal insights into biological processes and give hints on explaining how the genome works.

F. GENE SELECTION BASED ON MUTUAL INFORMATION FOR THE CLASSIFICATION OF MULTI-CLASS CANCER

Sheng-Bo Guu, Michael R.Lyu and Tat-Ming Lok (2006) says that with the development of microarray technology, microarray data are widely used in the diagnosis of cancer subtypes. However, people are still facing the complicated problem of accurate diagnosis of cancer subtypes.[10] Building classifiers based on the selected key genes from microarray data is a promising approach for the development of microarray technology; yet the selection of non-redundant but relevant genes is complicated. The selected genes should be small enough to allow diagnosis even in regular laboratories and ideally identify genes involved in cancer-specific regulatory pathways. Instead of traditional gene selection methods used for the classification of two categories of cancers, a novel gene selection algorithm based on mutual information is proposed for the classification of multi-class cancer using microarray data, and the selected key genes are fed into the classifier to classify the cancer subtypes. In our algorithm, mutual information is employed to select key genes related with class distinction. The application on the breast cancer data suggests that the algorithm can identify the key genes to the BRCA1 mutations/BRCA2 mutations/the sporadic mutation class distinction since the result of our proposed algorithm is promising, because our method can perform the classification of three types of breast cancer effectively and efficiently. And two more microarray datasets, leukemia and ovarian cancer data, are employed to validate the performance of our method. The performance of these applications demonstrate the high quality of our method. Based on this, our method can widely used to discriminate different cancer subtypes, which all contribute to the development of technology for the recovery of the cancer.

G. OPTIMAL SEARCH-BASED GENE SUBSET SELECTION FOR GENE ARRAY CANCER CLASSIFICATION

Jiexun Li (2007) proposed the optimal search based gene subset selection methods because they evaluate the group performance of genes and helps to pinpoint the global optimal set of marker genes. The overall methodology are: use an optimal search method to generate candidate gene subsets, assess these subsets based on evaluation criteria, then the gene subset with the goodness score is regarded as optimal. [8]The two optimal search methods to generate candidate gene subsets are Genetic Algorithm (GA) and Tabu Search (TS).A GA is an optimal search method which is used in many applications such as Internet Search Engines, feature selection and Intelligent Information retrieval. In a GA, each solution to a problem is represented in as a chromosomes, which is the string representing a gene subset. A pool of strings forms a population. A fitness function is defined to measure the goodness of a solution. Based on the principle of “Survival of the Fittest,” strings with higher fitness are more likely to be selected and assigned a number of copies into mating pool. Next, crossovers randomly choose pairs of strings from the pool and produce two offspring strings by exchanging genetic information between the two parents. Mutations are performed on each string by changing each element. Each string in new population is evaluated based on the fitness function. By repeating this process for a number of generations, the string with the best fitness of all generations is regarded as the optimum.TS algorithm is a metaheuristic method that guides the search for the optimal solution making use of flexible memory, which exploits the search history. TS is based on the assumption that solutions with higher objective value have a higher probability of either leading to a near-optimal solution, or to a good solution in a fewer number of steps. In each iteration, a TS moves to the best admissible neighboring solution, either with the greatest improvements or the least deterioration. A tabu list records the reverse of the most recent T moves to avoid cycling. A move in the tabu list is forbidden until it exist the tabu list in a first-in, first-out procedure.The evaluation criteria for gene subset selection inorder to assess the candidate gene subsets are Filter MRMR and wrappers (SVM). By combining the two optimal search algorithms with two evaluation criteria, we have four gene subset selection methods: GA/MRMR, TS/MRMR, GA/SVM and TS/SVM.

4. TABLE: SOME DATA SETS FOR GENE-BASED ANALYSIS

Data set	Description	Methods
Cho's data [11] K-means [15], SOM [16], CLICK, DHC [18]	6,220 ORFs in S. cerevisiae with 15 time points	K-means [15], SOM [16], CLICK, DHC
Iyer's data [17]	9,800 cDNAs with 12 time points	agglomerative hierarchi- cal [19], CLICK [5], DHC [10]
Wen's data [17]	112 rat genes during 9 time points	CAST [13]
Combined yeast data [14] agglomerative hierarchi- cal [19], model-based	6,178 ORFs in S. cerevisiae during 4 time courses	agglomerative hierarchi- cal [19], model-based
Colon cancer data [3]	6,500 human genes in 40 tumor and 22 normal colon tissue samples	divisive hierarchical [12], model-based [45]
human hematopoi- etic data	(1) 6000 genes in HL-60 cell lines with 4 time points (2) 6000 genes in four cell lines (HL-60, U937 and Jurkat with 4 time points and NB4 with 5 time points)	SOM [16]

IV. 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. In this paper, we reviewed both supervised clustering for gene expression and micro array technology. which have been applied to gene expression data, with promising results.

V.REFERENCE

[1] Au W.H., Chan K.C.C., Wong A.K.C. and Wang Y. (2005), '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.

[2] Dettling M. and Buhlmann P. (2002), 'Supervised Clustering of Genes', Genome Biology, Vol.3, No. 12, pp.0069.1-0069.15.

[3] Devijver P.A. and Kittler J. (1982), 'Pattern Recognition: A Statistical Approach', Prentice Hall.

[4] Domany E. (2003), 'Cluster Analysis of Gene Expression Data', J.Statistical Physics, Vol.110, Nos. 3-6, pp. 1117-1139.

[5] Golub T.R., Slonim D.K., Tamayo.P. and Huard.C. (1999), 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring', Science, Vol. 286, No. 5439, pp. 531-537.

[6] Huang D. and Chow T.W.S. (2004), 'Effective Feature Selection Scheme Using Mutual Information', Neurocomputing, Vol.63, pp.325-343.

[7] Lei Wang. (2008), 'Feature Selection with Kernel Class Separability', IEEE Trans.Pattern Analysis and Machine Intelligence., Vol. 30, No., 9.

[8] Li J., Su H., Chen.H. and Futscher B.W. (2009), 'Optimal Search-based Gene Subset Selection for Gene Array Cancer Classification', IEEE Trans. Biomedical Eng., Vol. 56, No .4, pp. 1063-1069.

[9] Pradipta Maji. (2011), 'Fuzzy-Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data', IEEE Trans. Cybernetics., Vol. 41, No.1.

[10] Sheng-Bo Guu., Michael Lyu R. and Tat-Ming Lok. (2004), 'Gene Selection Based on Mutual Information for the Classification of Multi-class Calcer', Science, Vol 134.

[11] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Molecular Cell, Vol. 2(1):65–73, July 1998.

[12] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonu- cleotide array. Proc. Natl. Acad. Sci. USA, Vol. 96(12):6745–6750, June 1999.

[13] Ben-Dor A., Shamir R. and Yakhini Z. Clustering gene expression patterns. Journal of Computational Biology, 6(3/4):281–297, 1999.

[14] Chu S., DeRisi J., Eisen M., Mulholland J., Botstein D., Brown PO., et al. The transcriptional program of sporulation in budding yeast. Science, 282(5389):699–705, 1998.

[15] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture. Nature Genet, pages 281–285, 1999.

[16] Tamayo P., Solni D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. and Golub T.R. In- terpretingpatternsof gene expressionwith self-organizingmaps: Methods and applicationto hematopoi- etic differentiation. Proc. Natl. Acad. Sci. USA, Vol. 96(6):2907–2912, March 1999.

[17] Iyer V.R., Eisen M.B., Ross D.T., Schuler G., Moore T., Lee J.C.F., Trent J.M., Staudt L.M., Hudson Jr. J., Boguski M.S., Lashkari D., Shalon D., Botstein D. and Brown P.O. The transcriptional program in the response of human fibroblasts to serum. Science, 283:83–87, 1999.

[18] Jiang, D., Pei, J. and Zhang, A. . DHC: A Density-based Hierarchical Clustering Method for Time- series Gene Expression Data. In Proceeding of BIBE2003: 3rd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, March 10-12 2003.

[19] Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David . Cluster analysis and dis- play of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95(25):14863–14868, December 1998.