ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

Improving Class Separability for Microarray datasets using Genetic Algorithm with KLD Measure

P.Aarthi, PG Scholar, Kongu Engineering College, Perundurai, Tamilnadu, India
E.Gothai, Associate Professor, Department of CSE, Kongu Engineering College, Perundurai, Tamilnadu, India

Abstract*—Among the great amount of attributes/genes presented in real life data sets, only a small fraction of them are effective for performing certain diagnostic test. The selection of relevant and non-redundant features of real valued datasets is a highly challenging issue. To solve the problem, one of the major tasks with gene expression data is to find the co-expressed gene groups whose joint expression is strongly related with the sample categories. For this, the gene clustering algorithm is used to cluster genes from microarray data. The Mutual information incorporates the information of sample categories to measure the similarity between attributes by sharing the information between each attributes. Thus the redundant and irrelevant attributes are eliminated. After forming the clusters, the Kullback-Leibler divergence measure is used to find the distance between testing and training datasets and the Genetic algorithm is used to find the significant feature from the divergence measure so as to increase the class separability. Using these techniques, the diagnosis can be made easier and effective. The predictive accuracy is estimated using three classifiers such as Linear Discriminant Analysis, Naive Bayes and Support Vector Machine. Thus the overall approach provides excellent predictive capability for accurate medical diagnosis.

Index Terms— **Classification, Genetic Algorithm, Gene clustering, Kullback-Leibler Divergence, Microarray, Mutual information.**

I. INTRODUCTION

In human each organ is formed of cells and each cell contains a nucleus. The nucleus is formed of double stranded Deoxy-ribo Nucleic Acid (DNA) molecule called Chromosomes. Gene is a DNA sequence located in a particular chromosome which encodes the information for the synthesis of proteins. Gene expression is the process by which the gene's coded information is converted into mRNA and then into proteins. The gene expression levels (level of mRNA or proteins produced for a gene) of thousands of genes are monitored and recorded using the microarray technology.

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 Ribo Nucleic Acid (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.

Microarray is a high throughput technology that allows uncovering of thousands of genes concurrently. The microarray data is represented in the form of matrix, where the row represents genes, columns represent the

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

samples and each value in the matrix shows the expression levels of genes. 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 unrelated 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. Accordingly, it is significant to eliminate those irrelevant genes and identify the informative genes for performing certain diagnostic test.

Thus in gene expression data analysis, the clustering methods such as Bayesian clustering, hierarchical clustering, k-means algorithm, self-organising map and principal component analysis groups a subset of genes that are interdependent or correlated with each other. The attribute clustering helps to reduce the search dimension of the classification algorithm and constructs the model using a tightly correlated subsets of genes rather than using the entire genes.

The main purpose with the gene expression data is to find groups of genes whose combined expression is strongly related with the sample categories. The unsupervised and some of the supervised clustering algorithms do not incorporate the information of sample categories. Thus the Supervised Attribute Clustering algorithm is used to find such groups of genes by incorporating the information of sample categories. A quantitative measure, based on Mutual Information is used to calculate the similarity between attributes. Subsequently, the cluster is formed for each relevant attribute that successively adds the attribute one after the other. The growth of the clusters is repeated until the cluster gets stabilized. The GA is applied to get the optimal feature to increase the class separability index. Thus as the CS gets increased, the new sample that we give gets into its respective class appropriately by obtaining a prediction of diagnosing diseases.

II. RELATED WORK

Au et al (2005) presents an attribute clustering method which groups genes based on their interdependence to mine the meaningful patterns from the gene expression data. This method is used for gene grouping, selection and classification. A new method ACA [1] was introduced so as to assemble the co-dependent attributes into clusters by optimizing a criterion function derived from an informative measure that reflects the interdependence between attributes. The meaningful clusters of genes are discovered on applying the Attribute Clustering Algorithm (ACA) to gene expression data. The clustering of genes based on the attribute interdependence within groups helps to capture different aspects of gene association patterns in each collection. Significant genes selected from each collection then contain useful information for gene expression classification and identification.

Dettling et al (2002) focused on Supervised Clustering [2], defining as grouping of genes controlled by information about the tumour types of tissues. This clustering can be started with or without initial groups of genes, and then the genes are clustered in a stage wise forward and towards the back search, as long as their differential expression can be enhanced. This yields clusters typically made up of three to nine genes, whose logical average expression levels allow a perfect discrimination of tissue types. Although it is sensitive to noise or outlier of the dataset, the output of this algorithm is beneficial for cancer type diagnosis.

Sheng-Bo et al(2006) says that microarray data are widely used in the diagnosis of cancer subtypes. However, everyone is still facing the difficult problem of accurate diagnosis of cancer subtypes. Based on the selected key genes building classifiers from microarray data is a capable approach for the development of microarray technology; yet the selection of non-redundant but relevant genes is complicated. The genes that are selected should be small enough to allow diagnosis even in regular laboratories and ideally recognize genes involved in cancer-specific regulatory pathways.

A novel gene selection algorithm based on mutual information [6] is proposed for the classification of multi-class cancer using microarray data, instead of traditional gene selection methods used for the classification of two categories of cancers and the selected key genes are fed into the classifier to organize 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 the proposed algorithm is promising, because

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

this 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 the method. The high quality of their method is demonstrated by the performance of these applications. Based on this, their method can widely used to distinguish different cancer subtypes, which all contribute to the development of technology for the recovery of the cancer.

Pradipta Maji (2011) proposed a new supervised gene clustering algorithm, named as Fuzzy-Rough Supervised Attribute Clustering (FRSAC) [10] based on fuzzy-rough sets. To deal with the uncertainty in gene expression, the FRSAC algorithm is used. It finds the coregualted clusters of genes. To compute the similarity between genes, Fuzzy-Rough sets are used. This measure incorporates the information of sample categories or class labels while measuring the similarity between genes. The FRSAC algorithm uses this measure to reduce the redundancy among genes. It involves the portioning of original gene set into some distinct subsets or clusters so that genes within the clusters are highly correlated, while those in different are dissimilar as possible.

Li et al (2007) proposed the optimal search based gene-selection methods because they calculate the performance of genes and helps to find out the optimal set of marker genes. The Tabu Search and Genetic Algorithm [9] are the important optimal search methods. In GA, the chromosomes (strings) form the initial population. Each gene in the population is measured using the fitness function. The genes with higher fitness value are selected based on the 'Survival of the Fittest 'principle and copied into the matting pool. The next step is, crossovers randomly choose a pair of strings from the pool and two offspring's are produced by exchanging the genetic information between the parent strings. Mutations are performed by changing the elements on each string. By repeating this procedure for number of generations, the strings with the best function of all generation is regarded as the optimum.

III. PROPOSED WORK

The proposed method deals with the similarity between attributes, clustering of relevant attributes, classification prediction using different classifiers and class separability with GA and PSO [9] is shown in Figure 1. It reduces the dimensionality, avoids the noise sensitivity problem and increases the classification accuracy of microarray data.

A. Preprocessing

The Preprocessing of data is necessary so as to avoid noisy data's containing errors and outliers, missing values and inconsistent data. The preprocessing procedure of genes illustrates how to filter the data by removing genes that are not uttered, having only small variability across samples or do not change. The data sets of genes are quite large. To find the interesting genes, the size of the dataset is reduced by removing genes with expression profiles that do not show anything of interest. There are number of techniques to reduce the number of gene expression profiles to some subset that contains the most significant genes. The different techniques are, the gene expression data with empty gene symbols are removed, the gene with missing data are identified using the function isnan in matlab and the genes are removed using the indexing comments, the genes with very low absolute expression values are filtered out using the genelowvalfilter function, the genes with a small variance across samples are filtered using the genevarfilter function, the genes whose profile having low entropy values is filtered using the geneentropyfilter function. Thus the preprocessing module filters out the missing genes, genes with low absolute value and small variances. The dimensionality reduction is achieved by eliminating the unnecessary genes.

B. Similarity Measure

In real data analysis, the most important issues are computing both significance and redundancy of attributes [11] by discovering dependencies among them. Thus the similarity between attributes needs to be computed after the data is preprocessed. In order to obtain the similarity measure, the relevance of the attributes is to be calculated.

The relevance of the attribute with respect to class label is defined using Mutual Information, which is used to quantify the information shared by two objects. If not similar information is exchanged between the independent objects, the value of mutual is small. Two highly correlated objects will display a elevated mutual information value. The objects can be the class label and the genes. If a gene has expression values randomly or uniformly distributed in unlike classes, its mutual information with these classes is zero. If a gene is strongly differentially expressed for different classes, it should have large mutual information. Thus the mutual information is measured based on the probability distribution of random variables using entropy and conditional entropy. The mutual

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

information may be used to measure the level of similarity or redundancy between two genes. The mutual information can be used as a measure of relevance of genes.

*

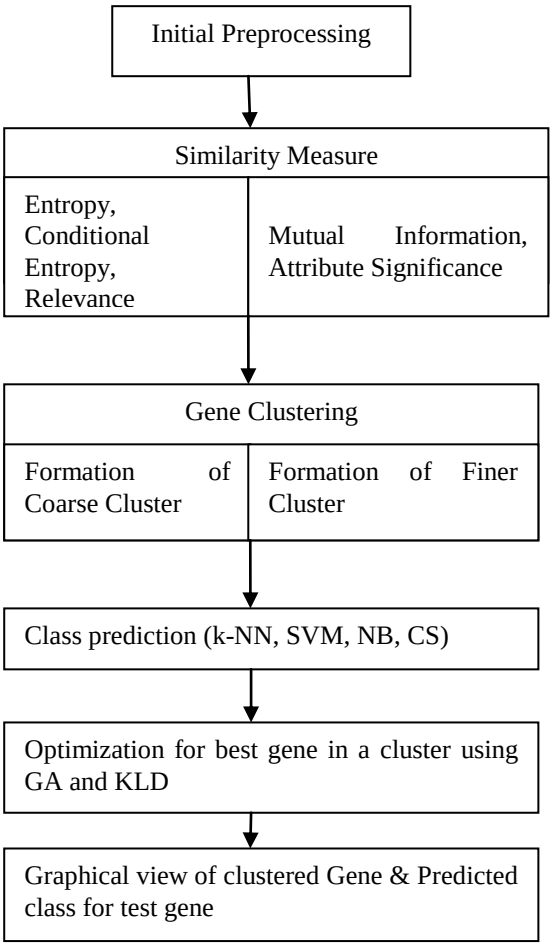

Fig 1 System Model

The next step is to find the significance of one attribute with respect to another attribute in the dataset. The significance of one attribute is the change in dependency when that attribute is removed from the set of attributes. The higher the change in dependency, the more significant that attribute is. If the significance is 0, then that attribute is dispensable. Based on the significance of an attribute, the supervised similarity measure between two attributes is calculated. Hence the supervised similarity measure directly takes into account the information of sample categories or class labels while computing the similarity between two attributes. If the two attributes are completely correlated with respect to the class labels then the supervised similarity between them will be 1. If the two attributes are uncorrelated, then their value is 0. Thus the output of the similarity measure lies between 0 and 1. Using this similarity measure the clusters can be formed.

C. Gene Clustering

The Gene clustering method [10] uses this similarity measure to reduce the redundancy among genes. It involves the separation of the original gene set into some diverse subsets or clusters so that the genes within the clusters are highly co expressed with the strong relation to sample categories, while those in different clusters are as dissimilar as possible. From each coarse cluster a gene having highest gene-class relevance value is first selected as the initial representative of that cluster. The representative of each cluster is then changed by averaging the initial representative with other genes of that cluster. Finally the modified representative of each cluster is selected to constitute the resulting gene set in the finer cluster. This algorithm yields biologically significant gene clusters, where Gene Ontology Term finder is used and avoids the noise sensitivity problems. Thus the output of this

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

supervised attribute clustering is shown using the clustergram where red shades represents higher expression level, green shades represents lower expression level and the black shades represents the absence of differential expression values.

D. Optimization using GA

After the formation of finer cluster, three classifiers such as LDA, NB and SVM are used to measure the accuracy of the generated clusters. The classification of different types of cancer is obtained based on their sample categories. The class separation between different classes is measured using the Class Separability Index. The Class Separability filter calculates the class separability of each feature using the Kullback-Leibler (KL) distance between histograms of feature values [13]. For each feature there is one histogram for each class. The histograms are normalized dividing each bin count by the total number of elements to estimate the probability that the j-th feature takes a value in the i-th bin of the histogram given a class n, $p_j(d = i \mid c = n)$. For each feature j, we calculate the class separability in Equation (1) as,

$$\Delta_j = \sum_{m=1}^{c}\sum_{n=1}^{c}\delta_j(m,n) \qquad (1)$$

Where c is the number of classes and $\delta_j(m,n)$ is the KL distance between histograms corresponding to classes m, n is represented in Equation (2) as,

$$\delta_j(m,n) = \sum_{i=1}^{b} p_j(d = i/c = m)\log\left(\frac{p_i(d = i/c = m)}{p_j\left|(d = i/c = n)\right|}\right) \qquad (2)$$

Where b is the number of bins in the histograms. The features are then sorted in descending order of the distances Δ_j.

In order to achieve class separation at higher rate, the class separability index should attain lower optimum value. Thus to get the optimum value, GA [9] and PSO is used. As much good the optimal solution obtained, that much higher the class separation will improve. Hence the test dataset is given and the entire step is proceeded to get the optimal gene.

IV. RESULTS & ANALYSIS

The datasets are obtained from the Kent-Ridge Bio-medical data center. This provides three cancer datasets such as Breast Cancer, Leukemia and Colon Cancer, and two arthritis datasets such as Rheumatoid Arthritis versus Osteoarthritis (RAOA) and Rheumatoid Arthritis versus Healthy Controls.

A. Breast Cancer

The breast cancer training data contains 78 patient samples, 34 of which are from patients who had developed distance metastases within 5 years (labelled as "relapse"), the rest 44 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years (labelled as "non-relapse"). Correspondingly, there are 12 relapse and 7 non-relapse samples in the testing data set. The number of genes is 24481. We replaced "NaN" symbol in original ratio data with 100.0.

B. Leukemia

Training dataset consists [6] of 38 bone marrow samples (27 ALL and 11 AML), over 7129 probes from 6817 human genes. Also 34 samples testing data is provided, with 20 ALL and 14 AML.

C. Colon Tumor

The colon cancer contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors (labelled as "negative") and 22 normal (labelled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

Using MATLAB, the Supervised Gene Clustering Algorithm is represented in figure 2 by the clustergram.

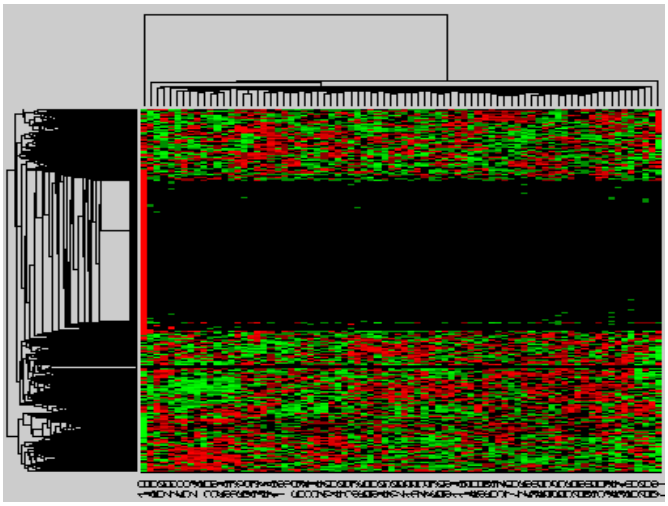

Fig 2 Cluster gram

Table1. Results of the Classification

Performance of the Classifier: SVM	Existing: Supervised Algorithm	Proposed: CS with KLD & Genetic Algorithm
Correct rate	0.7680	0.8900
Error rate	0.2320	0.1100
Sensitivity	0.8627	0.9435
Specificity	0.6085	0.7820

Table1 represents the result of the classifier using breast cancer datasets and compares the performance of the existing supervised algorithm and the proposed Genetic Algorithm along with KLD**.** The accuracy has been improved using the proposed KLD with GA method.

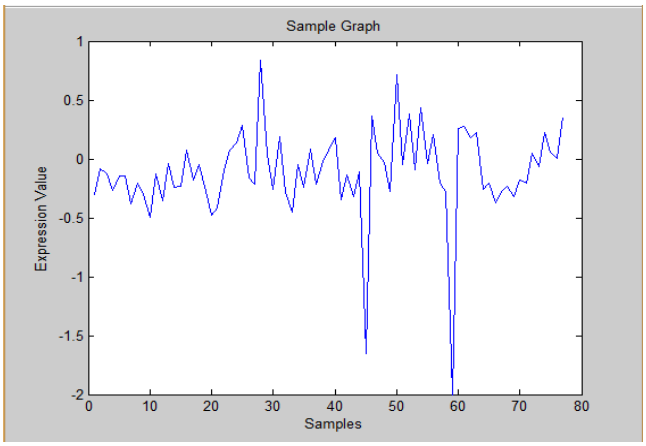

Fig 3 Graph of proposed algorithm

Figure 3 represents the graph of the samples vs expression value for the breast cancer datasets. The results obtained using the proposed algorithm KLD with GA expresses the cancer affected in each sample. Similarly the other datasets leukemia and colon tumor are used in this algorithm and tested using other different classifiers such as Linear Discriminant Analysis and Naive Bayes.Thus from the results totally obtained , it is analyzed that the

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

proposed algorithm finds the genes affected by cancer efficiently by increasing the class separability and the clustering gives 100% accuracy.

V. CONCLUSION

This paper presents a Gene clustering algorithm for cancer classification using microarray experiments. The co-expressed clusters of genes are found out using the Gene Clustering Algorithm and the redundant and irrelevant genes are eliminated using the mutual information measure. The GA along with the Kullback-Leibler Divergence measure selects the optimal feature from the clusters and increases the class separability which improves the classification and predictive accuracy of each cell. Thus the proposed algorithm is potentially useful in the context of medical diagnosis as it identifies groups of interacting genes that have high explanatory power for given tissue types, and which in turn can accurately predict the class labels of new samples. Hence the performance is estimated using three different classifiers such as LDA, NB and SVM to predict the accuracy. The future work can be extended by improving the performance of classification using different kernel functions of SVM and the similarity can be calculated using rough sets and fuzzy sets.

ACKNOWLEDGMENT

I extend my gratitude to my supervisor, E. Gothai M.E., for her valuable ideas and suggestions, which have been very helpful in the project. I am grateful to my Head Of the Department, Dr.R.Thangarajan M.E,Ph.D and all the faculty members of the Computer Science and Engineering Department, for their support.

REFERENCES

[1] 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-Jun 2005.

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

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

[4] S. C. Dinger, "Cluster Analysis of Gene Expression Data on Cancerous Tissue Sample", J.Statistical Physics, Vol.110, Nos. 3-6, pp. 1117-1139, 2011.

[5] P. Ghosh, Arka, Ranjan Maitra and D. Anna Peterson, "A Separability Index for Distance-based Clustering and Classification Algorithms", 2012.

[6] T. R. Golub, D. K. Slonim, P. Tamayo and C. Huard, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, Vol. 286, No. 5439, pp. 531-537, 1999.

[7] S. B. Guu, R. Michael Lyu, and Tat-Ming Lok, "Gene Selection Based on Mutual Information for the Classification of Multi-class Cancer", Science, Vol 134, 2004.

[8] D. Huang and T. W. S. Chow, "Effective Feature Selection Scheme Using Mutual Information", Science, Vol 152, 2004

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

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

[11] Pradipta, Maji, "Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification", IEEE transaction on Knowledge and data engineering. Vol 24, No.1, Jan2012.

[12] 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

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 2, March 2014

[13] L.Wang, "Feature Selection with Kernel Class Separability", IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 30, no., 9, 2008.

[14] Wendy Stevens, Meir Perez and Jonathan Featherston "Differentially Expressed Gene Identification based on Separability Index", IEEE, Machine Learning and Application, 2009.