Hybrid System For Microarray Data

Mrs.Mangayarkkarasi
M.Tech
Computer Science & Engineering,
Kalasalingam University, Krishnan Kovil,
Tamilnadu,
India.

Ms.J.Jeyaranjani
Assistant Professor,
Computer Science & Engineering,
Kalasalingam University, Krishnan Kovil,
Tamilnadu,
India.

***Abstract*: A hybrid system is proposed by combining the unsupervised and supervised learning techniques in data mining methodology. Microarray data involves large number of genes and a small number of samples .Genes which is useful is selected from the dataset and a Fuzzy system is designed to perform classification. The optimization and the accuracy are achieved through Genetic Algorithm. The dataset used is Lung cancer.**

***Index Terms*—Hybrid system, Unsupervised, Supervised, Genetic Algorithm.**

1. Introduction

Lung cancer is a type of cancer that begins in the lungs. Your lungs are two spongy organs in your chest that take in oxygen when you inhale and release carbon dioxide when you exhale Lung cancer is the leading cause of cancer deaths in the United States, among both men and women. Lung cancer claims more lives each year than do colon, prostate, ovarian and breast cancers combined. People who smoke have the greatest risk of lung cancer. The risk of lung cancer increases with the length of time and number of cigarettes you've smoked. If you quit smoking, even after smoking for many years, you can significantly reduce your chances of developing lung cancer.

Microarray data is a collection of microscopic DNA spots attached to a solid surface. The dataset is comprised of large number of genes and it is difficult to classify the samples one by one. The genes which is valid and useful for the classification is collected and analysed.Genes are termed as features and they are selected using t-test filtering hypothesis method. The Fuzzy system is designed by applying Fcm clustering algorithm to this selected genes and the lung cancer classes are predicted as clusters. The if-then rules are coded for the dataset and the classification accuracy is accomplished through G.A.

1.1BACKGROUND OF THE PROJECT

[a] Data Mining

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining Microarray data is the one which provides accurate medical diagnosis and helps to find the right treatment and cure for many diseases.

- Descriptive
- Classification and Prediction

[b]Microarray

Microarrays are capable of determining the expression levels of thousands of genes simultaneously. One important application of gene expression data is classification of samples into classes. Microarray data is the one which provides accurate medical diagnosis and helps to find the right treatment and cure for many diseases. Nowadays analyzing the microarray data is very difficult in medical field.

[c]Challenges for Microarray

- Number of genes are in thousands
- Number of samples are small
- False positives
- Strong methods for validation

A hybrid method proposed here to overcome these challenges and to give the best classification accuracy for the lung cancer dataset.

2. Proposed Work

In this proposed system a novel hybrid system is developed to classify the large microarray geneset.

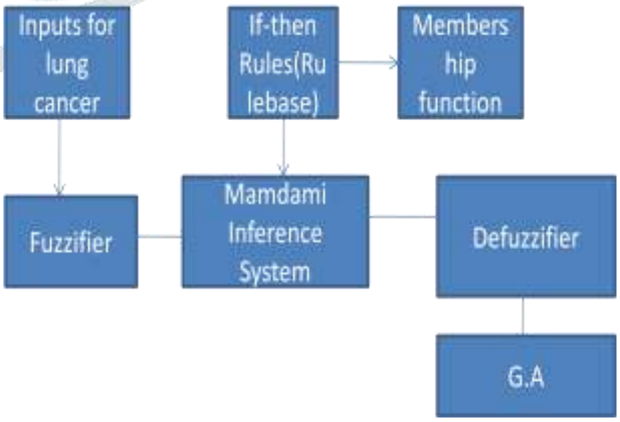


Figure1.Block Diagram of Hybrid System

[a]Tasks done in microarray

- **Gene Selection**
- **Fuzzy system Design**
- **Classification**

- **Optimization**

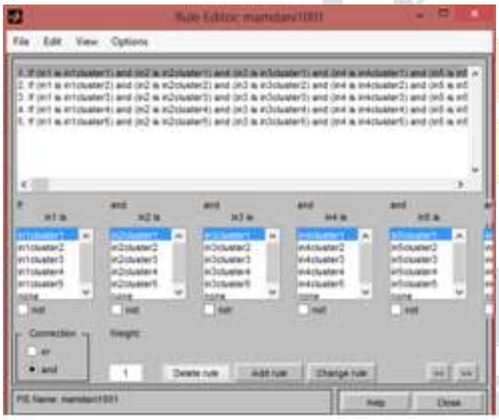
3. Gene Selection

The dataset is identified and visualized .It consists of thousands of genes with few samples. The whole dataset cant be classified and it is difficult so we select some informative genes by applying certain gene selection methods. The t-test hypothesis filter is used to select required features from the dataset. The feature selection is the process of selecting relevant features to use in the Fuzzy design. On using this techniques

- Models can be created simple
- The training time will be short
- Overfitting can be avoided
- Redundant and irrevalant features are removed

4. Fuzzy System Design

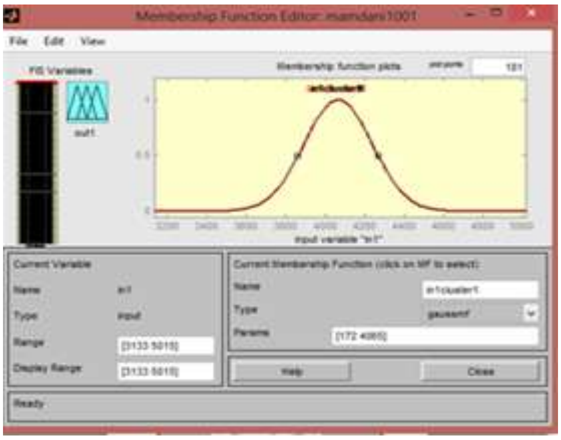
The features selected can’t be classified directly and it should be learned first through some unsupervised learning technique. The aim of the clustering is to group the data into clusters with its relations. Fuzzy Logic Toolbox tools allow you to find clusters in input-output training data Cluster information is used to generate a Sugeno-type fuzzy inference system that best models the data behavior using a minimum number of rules.


1. If (in1 is in1cluster1) and (in2 is in2cluster1) and (in3 is in3cluster1) and (in4 is in4cluster1) and (in5 is in5
2. If (in1 is in1cluster2) and (in2 is in2cluster2) and (in3 is in3cluster2) and (in4 is in4cluster2) and (in5 is in5
3. If (in1 is in1cluster3) and (in2 is in2cluster3) and (in3 is in3cluster3) and (in4 is in4cluster3) and (in5 is in5
4. If (in1 is in1cluster4) and (in2 is in2cluster4) and (in3 is in3cluster4) and (in4 is in4cluster4) and (in5 is in5
5. If (in1 is in1cluster5) and (in2 is in2cluster5) and (in3 is in3cluster5) and (in4 is in4cluster5) and (in5 is in5

FIS Name: mamdani1001

Figure 2.Rule Viewer

The rules are added in the rule viewer. For every input a membership function is generated within the clusters. The membership function of the input value is between 0 and 1.The membership function defines the point of view of simplicity, convenience, speed, and efficiency. A fuzzy set admits the possibility of partial membership in it. A membership function associated with a given fuzzy set maps an input value to its appropriate membership value.


Current Variable
Name: in1
Type: input
Range: [3133 5015]
Display Range: [3133 5015]

Current Membership Function (click on MF to select)
Name: in1cluster1
Type: gaussmf
Params: [172 4065]

Ready

Figure 3.Membership Function

4.1FCM Algorithm

The Clustering algorithm applied to design a fuzzy system is Fuzzy C-Means Algorithm. The greatest advantage of using fuzzy logic lies in the fact that scientists can model complex systems by implementing human experience, knowledge, non-linear, and imprecise and practice as a set of inference rules or if-then rules that use fuzzy variables. The gene expression data is predicted as the classes. The information returned from by fcm is used to build a fuzzy inference system. Fuzzy Logic Toolbox tools allow you to find clusters in input-output training data Cluster information is used to generate a Sugeno-type fuzzy inference system that best models the data behavior using a minimum number of rules.

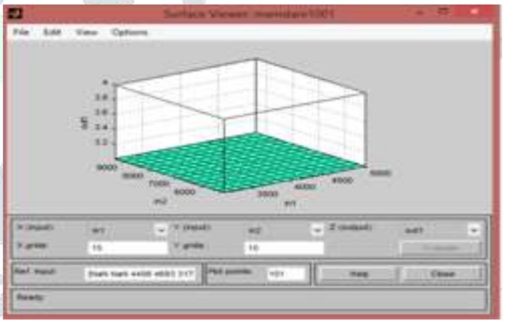

Ready

Figure4.Output of the Fuzzy System

4.2Algorithm

- Choose a number of clusters
- Assign randomly to each point coefficients for being in the clusters
- Repeat until the algorithm has converged
- Compute the centroid for each cluster, using the formula above
- For each point,compute its coefficients in the clusters,using the above formula

4.3Advantage of FCM

- Improved accuracy of clustering under noise.
- Minimization of the function
- Better results than K-Means Algorithm

5. Classification

The classification is done to accurately predict the target class for the data. The supervised learning technique is applied because the classes are known earlier for the data. Classification is done by reducing the dimension of the large dataset. The Dimensionality reduction methods are applied for the dataset. The features are compared with the samples and the number of gene values is reduced.

5.1Dimensionality Reduction

A classification procedure for the purpose may consist of two basic steps.

- Dimension reduction, in which the data are reduced from the high p-dimensional gene space to a lower K-dimensional
- Class prediction, in which response classes are predicted using a standard class prediction model on the gene components

There are three dimensional reduction methods such as PLS,SIR and PCA.Here PCA is applied for the reduction.After reducing the dataset classification becomes quite easy and the classification accuracy is obtained.

5.2PCA

PCA is termed as principal component analysis and it is a well known method of dimension reduction. The basic idea of PCA is to reduce the dimensionality of a data set, while retaining as much as possible the variation present in the original predictor variables. This is achieved by transforming the p original variables X = [x1, x2, …, xp] to a new set of K predictor variables, T = [t1, t2, …, tK], which are linear combinations of the original variables. The maximum number of components K is determined by the number of nonzero eigenvalues

6. Optimization

6.1Genetic Algorithm

A Genetic algorithm (or GA) is s search technique used in computing to find true or approximate solutions to optimization and search problems. Best features are selected from the classified dataset..Encode the solution on a chromosome should be right Parameters decided are initial population size and the procedure..Fitness function should be derived from the objective function and the operators such as selection, crossover and mutations are applied.

7. Conclusion

In this paper A Hybrid system is proposed for classifying the lung cancer gene samples. The proposed model consists of fuzzy system and genetic algorithm .it gives increase in accuracy of percentage of the genes selected for classification. The rules should be generated from the dataset to define a class from classifier. Thus the classified features are given to Genetic Algorithm to get the optimized output. The classification accuracy is improved by using this model.

8. Future Work

Analyzing the microarray data through mapreduce framework technique and deployed in hadoop.effiency of microarray will be high based on the performance and to do in a low cost value.

9.REFERENCES

1] P. Ganesh Kumar, 2S. Arul Antran Vijay, 3D.Devaraj, A Hybrid Colony Fuzzy System for Analyzing Diabetes Microarray Data,in: in:Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),2013,

[2]S.Ashwin, S.Aravind Kumar, S.Arun Kumar, Soft Computing Techniques Based Computer Aided System for Efficient Lung Nodule Detection – A Survey: in: International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-2, December 2012.

[3]Zakaria Suliman Zubi1, Marim Aboajela Emsaed,Identifying Cancer Patients using DNA Micro-Array Data in Data Mining Environment, Journal of Science and Engineering Vol. 3(2), 63-75. Vol. 3 (2), 2013

[4] Z. S. Zubi, M. A. Emsaed, (2013), Identifying Cancer Patients using DNA Micro-Array Data in Data Mining Environment, Journal of Science and Engineering, Vol. 3(2), 63-75.

[5] Thanh Nguyen, Abbas Khosravi,Douglas Creighton,Saeid Nahavandi Hierachial Gene Selection and Genetic Fuzzy System for Cancer Microarray Cancerous Cell Using Soft Computing Technique,in:Procedia Computer Science 49 (2015) 66 – 73.

[6]Selva Mary. G, Sachin Bojewar "Classification of Microarray Gene Expression: A Comparative Analysis using Dimensionality Reduction Techniques" "Pragyanam „14" the proceedings of International Conference on Recent Trends in Computer and Electronics Engineering(ICRTCEE), January, 2014.

[7] Mukesh Kumara,∗, Nitish Kumar Rath, Amitav Swain and Santanu Kumar Rath Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015).

[8] G.C.J. Alonso, I.Q. Moro-Sancho, A. Simon-Hurtado, and R. Varela- Arrabal, "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods," Expert Systems with Applications, vol. 39, pp. 7270 –7280, 2012.