Scientific Journal Impact Factor (SJIF): 3.518

IJMTER International Journal of Modern Trends in Engineering and Research

ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

ASSAY OF LYMPHOMA USING GENE EXPRESSION DATA-GENETIC FUZZY SYSTEM AND MODIFIED K-MEANS ALGORITHM

Ms. B.Nagarathna

Assistant Professor, AVS Arts and Science College, Salem, TN, India

Abstract: Microarray gene expression data are playing an essential role in cancer classifications. However, due to the availability of small number of effective samples compared to the large number of genes in microarray data, many computational methods have failed to identify a small subset of important genes. Therefore, it is a challenging task to identify small number of disease-specific significant genes related for precise diagnosis of cancer sub classes. In this research, Genetic fuzzy system along with Modified K-means algorithm based gene selection technique is proposed to distinguish a small subset of useful genes that are sufficient for the desired classification purpose.

Keywords: Genetic Algorithm, Fuzzy system, Modified k-means

I. INTRODUCTION

Microarrays are capable of profiling the gene expression patterns of tens of thousands of genes in a single experiment. Gene expression data can be a valuable source for understanding the genes and the biological associations between them. It has high dimension, small samples and the gene selection i.e. Feature selection is very important to determine the classification accuracy. The dataset utilized for this work is called Lymphoma Dataset which includes 4026 gene expression values with its subtypes[1].

The task of feature selection is generally divided into two aspects eliminating *irrelevant* features and *redundant* ones. Irrelevant features usually disturb the learner and degrade the accuracy, while redundant features add to computational cost without bringing in new information. All the genes used in the expression profile are not informative; also many of them are redundant. Finding informative genes greatly reduces the computational burden and noise arising from irrelevant genes. Reducing the number of genes by feature selection and still retaining best class prediction accuracy for the classifier is vital in case of classification [2].

Gene ranking simplifies gene expression tests to include only a very small number of genes rather than thousands of genes. The goal is to identify a small subset of genes which together give accurate predictions.

The basic aim of the research work is to reduce the number of genes, particularly suitable for identification of particular type of carcinoma. Genetic fuzzy system and Modified K means algorithm based gene selection technique is used to discriminate a small subset of useful genes that are adequate for the desired arrangement purpose. Gene expression data contains lymphoma dataset. The complete data set includes the expression data of 4,026 genes each measured using a focused cDNA microarray.The lymphoma data set is used in this research and the results are more efficient than existing system. The lymphoma data set has tested with 4026 genes with 62 rows and 4026 columns.

The data set contains noisy data and missing values. Pre processing is to be done for removing of the noisy data. After removing the noisy data the ranking should be done for the Genes who have high values for the informative genes.

Modified K means would help to form the appropriate numbers of clusters to be explored and hence to classify the dataset accurately and processor maintains the cluster structures in its own local

DOI:10.21884/IJMTER.2016.3107.YPL5U

memory. First, the number of gene clusters in a gene expression data set is usually unknown in advance. To detect the optimal number of clusters, users usually run the algorithms repeatedly with different values of *k* and compare the clustering results.

While a genetic algorithm is a very powerful tools to identify the fuzzy membership functions of a pre-defined rule base, they have their limitation especially when it also comes to identify the input and output variables of a fuzzy system from a given set of data. Genetic programming has been used to identify the input variables, the rule base as well as the involved membership functions of a fuzzy model [6].

Genetic based fuzzy algorithm for clustering and classified the genes[7]. A fuzzy set generalizes the concept of an ordinary set whose membership function only takes two values, zero and unity [8]. Thus, an element must either belong to an ordinary set or not belong to it. Fuzzy logic value determines 0 and 1 it predicts whether the corresponding gene has cancer or not then given result will be analyzing by genetic algorithm. The genes were processed under selection, mutation and crossover.

A normalizing input of threshold value is set as 5.If the given input values is less than the given threshold value then the given genes is not suffered from cancer or else if the inputs value is above the threshold value then the given gene is suffered from cancer and it will be indicate as high. If the given value is between 3 to 4.9 it indicates that the gene is suffered from cancer or some other diseases and it will indicate as average.

These algorithms are more efficient for accuracy and time period, clustering range, classification time is high and most efficient. The system is implemented and tested in matlab 2013 and executed successfully.

II. FUZZY RULE OPTIMIZATION BY GENETIC ALGORITHM

Individuals are characterized by chromosomes (or genomes) Sk, k = {1, . . .,n}. The chromosome is a string of symbols, which are called genes, Sk = (Sk1,. . .,SkM), and M is a string length. Individuals are evaluated via calculation of a fitness function. To evolve through successive generations, GA performs three basic genetic operators: selection, crossover and mutation.

A roulette wheel selection method is used to select the individuals that go on to produce an intermediate population. Parents are selected based on their fitness. Chromosomes have more chances to be selected if they are better (have higher fitness) than the others. Imagine all chromosomes in the population are placed on a roulette wheel, and each has its place big according to its fitness function.

The wheel is rotated and the selection point indicates which chromosome is selected when the wheel is stopped. It is obvious that the chromosome with bigger fitness will be selected more times (competing rule in the evolutionary theory).

The crossover operator selects random pairs from the intermediate population and performs 1-point crossover. Genes from parent chromosomes are selected to create new offspring. Finally, individuals are mutated and they form the new population. The mutation prevents falling all solutions in the population into a local optimum of the problem being solved. A few randomly chosen bits are switched from 1 to 0 or from 0 to 1.

Through chromosomes' evolution, GA searches for the best solution(s) in the sense of the given fitness function. We employ GA to train the complicated FSAM comprising many parameters. The fitness function is designed with the aim to reduce the number of fuzzy rules and also to decrease the learning error at the same time. The following formula is proposed:

$$fit(m) = ln(\bar{\sigma}^2) + \frac{log_n(m)}{n}$$

Where m is the number of fuzzy rules, n is the number of data samples, and s_2 is the error term defined by the following equation:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - F(x_i))^2$$

where yi is the real value and F(xi) is the output of the FSAM[10]. Parameters of FSAM are coded into genes of the GA chromosomes/individuals. With a population of individuals, GA can simultaneously explore different parts of the training model's parameter space and thus it is able to find the global solution to simultaneously minimize the error term and reduce the number of fuzzy rules [9].

Pseudo code for Genetic Fuzzy System

```
Input: real number set $X$
Output: optimal fuzzy set $Y$ for decision support
Procedure: FGA (m, $X(t)$ , $X(t)$ , $X'(t)$ ,$Y$ )
t := 0; //start with an initial time
//initialize a fuzzy random population of individuals $X(t)$ by fuzzifying the real number sets $X(t)$
with proper membership functions $X$ m ,
evaluate $X(t)$ ;//evaluate the fitness of all initial individuals of population based on fuzzy evaluation
While (not done) do  //test for termination criterion t := t + 1; //increase the time counter
production //select a fuzzy sub-population set $X'(t)$ for offspring
$X'(t)$ := select $X(t)$ ; //crossover the "genes" of the selected parents $X'(t)$
crossover $X'(t)$ ; //perturb the mated population stochastically
mutate $X'(t)$ ; //fuzzily evaluate its new fitness
evaluate $X'(t)$ ; //select the survivors $Y$ from actual fitness
$Y$ := survive $X(t)$ , $X'(t)$ ; End
rank $Y$ ;   //fuzzily rank the survivors
export $Y$ ;  //defuzzify and export the final survivors
return population // population of optimal clusters ;end
```

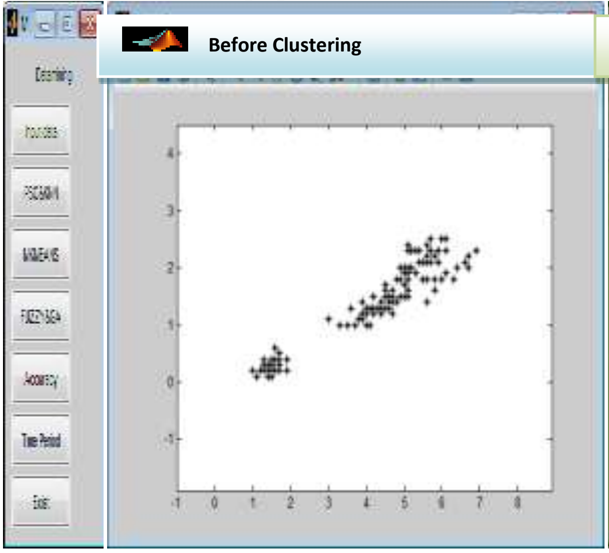


Figure 1: Before Clustering

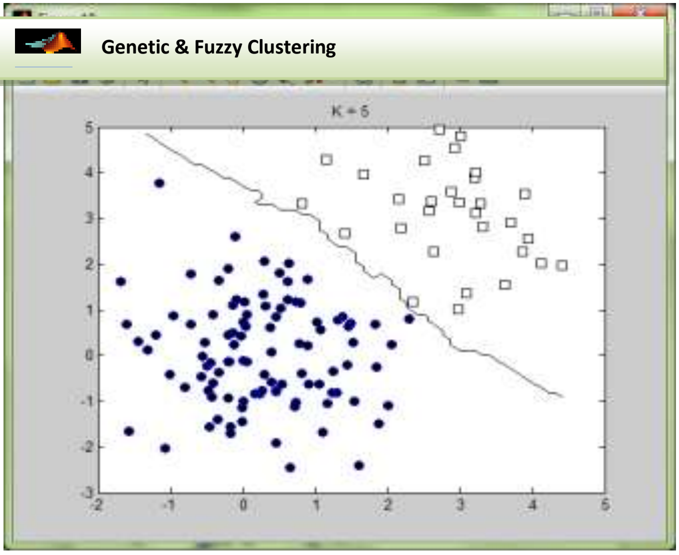


Figure 2: Genetic fuzzy system clustering

III. MODIFIED K-MEANS CLUSTERING

Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster error criterion[12].

The *k*-means clustering algorithm faces two major problems. One is the problem of obtaining non-optimal solutions. As the algorithm is greedy in nature, it is expected to converge to a locally optimal solution only and not to the global optimal solution, in general. This problem is partially solved by applying the *k*-means in a stochastic framework like simulated Fuzzy and genetic algorithm (GA) etc. The second problem is that of empty cluster generation. This problem is also referred to as the singularity problem in literature. Singularity in clustering is obtained when one or more clusters become empty. Both the problems are caused by bad initialization. Algorithms that use the *k*-means at their core suffer from the empty cluster problem too.

First phase is to determine initial centroids, for this compute the distance between each data point and all other data points in the set D. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to A1 and delete it from D.

Repeat this procedure until the number of elements in the set A1 reaches a threshold. Then again form another data-point set A2. Repeat this till 'k' such sets of data points are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set. The Euclidean distance is used for determining the closeness of each data point to the cluster centroids.

Next phase is to assign points to the clusters. Here the main idea is to set two simple data structures to retain the labels of cluster and the distance of all the data objects to the nearest cluster during the each iteration, that can be used in next iteration, we calculate the distance between the current data object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in its cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k- 1clustering center, saving the calculative time to the k-1 cluster centers.

Pseudo code for modified k -Means algorithm

```
Algorithm: Modified approach (S, k), S={x1,x2,…,xn }
Input: The number of clusters k1( k1> k ) and a dataset containing n objects(Xij+).
Output: A set of k clusters (Cij) that minimize the Cluster - error criterion
Compute the distance between each data point and all other data- points in the set D.
Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k+1) which contains these two data- points, Delete these two data points from the set D.
Find the data point in D that is closest to the data point Ap, Add it to Ap and delete it from D.
Repeat step 4 until the number of data points in Am reaches (n/k).
If p<k+1, then p = p+1, find another pair of data points from D between which the distance is  the shortest, from another data point set Ap and delete them from  D,Go to step 4.
Phase I
•   For each data-point set Am (1<=p<=k) find the arithmetic mean of the vectors of data
•   points Cp(1<=p<=k) in Ap.
•   Select nearest object of each Cp(1<=p<=k) as initial centroid.
•   Compute the distance of each data-point di (1<=i<=n) to all the centroids cj (1<=j<=k+1) as d(di, cj)
•   For each data-point di, find the closest centroid cj and assign di to cluster j
•   Set ClusterId[i]=j; // j:Id of the closest cluster
•   Set Nearest_Dist[i++]= d(di, cj)
•   For each cluster j (1<=j<=k), recalculate the centroids  ;Repeat
Phase II
For each data-point di
•   Compute its distance from the centroid of the present nearest cluster
•   If this distance is less than or equal to the present nearest distance, the data-point stays in the cluster   Else ;
•   For every centroid cj (1<=j<=k) Compute the distance (di, cj); Endfor
•   Assign the data-point di to the cluster with the nearest centroid Cj
•   Set ClusterId[i] =j
•   Set Nearest_Dist[i] = d (di, cj);   Endfor
```

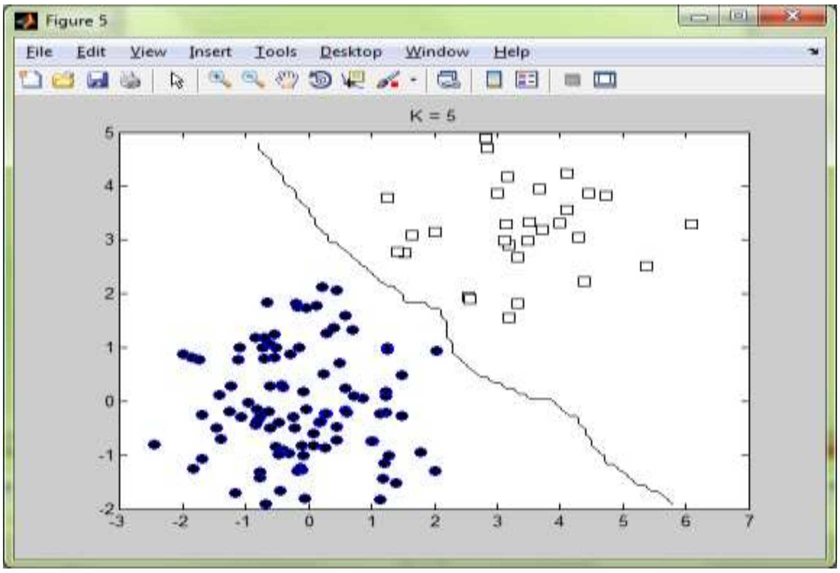


Figure 3: Modified k-Means clustering

IV.EXPERIMENTAL STUDY

4.1 Lymphoma Cancer Dataset

Gene Expression data is one of the latest breakthroughs in experimental molecular biology, which allow monitoring of gene expression for tens of thousands of genes in parallel and are producing huge amounts of valuable data. The Lymphoma dataset is downloaded from Lymphoma / Leukemia Molecular Profiling Project (LLMPP) webpage [http: //llmpp .nih. gov/lymphoma/ data /figure1 /figure1.cdt]. Human BCell contains about 4026 genes expressed in lymphoid cells or which are known as immunological or ontological importance with 96 conditions. **Format**

A list with two components:

x Gene expression data. A matrix with 62 rows and 4026 columns.

y Class index. A vector with 62 elements.

Details

The lymphoma dataset consists of

42 samples of Diffuse Large B-cell Lymphoma (DLBCL),

9 samples of Follicular Lymphoma (FL), and

GENEID	NAME	VALUES	VALUES	VALUES
GENE3129X	Autocrine motility factor receptor Clone=1072783	-0.3000	0.30000	0.5900
GENE3126X	2B catalytic submit Clone=1087686	-0.02200	-1.2000	1.4100
GENE3178Y	Probable ATP Clone=13560987	-0.0400	0.1500	0.6800
GENE3067X	SRC- like adapter protein Clone =7017678	0.4100	-0.3400	-0.1800
GENE4006X	AP Clone=1356783	1.7600	1.2100	0.9990

Table 1: A Sample data from Lymphoma Dataset

4.1.1 Pre-processing

Data pre-processing is an often neglected but important step in the data mining process. Pre-processing is the process of removal of noisy data and filtering necessary information. The lymphoma dataset downloaded consist of noisy and inconsistent data. The multiple empty spots as shown in Table 2 are filled with values in the pre-processing phase.

GENE	NAME	VALUES	VALUES	VALUES	VALUES
GENE1853X	(Clone=1357915)	-0.1300			0.4000
GENE1836X	(Clone=1358277)	-0.3100	0.1600		0.2500

GENE1865X	(clone=1358604)	-0.1200	0.5200		0.8300
GENE1933X	(Clone=1358190)	0.0500			0.2800
GENE1932X	(Clone=1336836)	-0.2600		-.00990	0.1500
GENE1931X	(Clone=1336983)	-0.5500			

Table 2: Lymphoma Dataset with empty spots

4.1.2 Removal of Noisy Data

The lymphoma dataset contains 4026 genes out of which certain gene expression values are missing. The missing data is imputed by knn impute method. It replaces the data with the corresponding value from the nearest-neighbor column. The missing data in lymphoma dataset is replaced with nearest neighbor values as it is shown in Table 3.

GENE	NAME	VALUES	VALUES	VALUES	VALUES
GENE1853X	(Clone=1357915)	-0.1300	-0.2800	-0.2800	0.4000
GENE1836X	(Clone=1358277)	-0.3100	0.1600	0.1600	0.2500
GENE1865X	(clone=1358604)	-0.1200	0.5200	0.5200	0.8300
GENE1933X	(Clone=1358190)	0.0500	0.1750	0.8500	0.2800
GENE1932X	(Clone=1336836)	-0.2600	-0.0990	-.00990	0.1500
GENE1931X	(Clone=1336983)	-0.5500	-0.5500	-0.5500	-0.5500

Table 3: Pre-processed Lymphoma Dataset

The empty spots are filled with nearest values as data and the pre-processed values are given as input to the next process, called the ranking of genes.

4.1.3 Ranking of Genes

Gene ranking simplifies gene expression tests to include only a very small number of genes rather than thousands of genes. The importance ranking of each gene is done using a feature ranking measure which ranks the genes based on their statistical score. The value compares the actual difference between two means in relation to the variation in the data which is expressed as the standard deviation of the difference between the means. Genes includes the classes with different samples. The mean value of each gene expression in a class is calculated. In fact, the TS used here is a t-statistic between the centroid of a specific class and the overall centroid of all the classes. The rank of genes 'i' is defined as

$$\boldsymbol{Tsi = max\left\{\left|\frac{\overline{xik}-xi}{mksi}\right|\ k = 1, 2 \ldots .k\right\}} \qquad (4)$$

Where there are K classes. Max (yk, k=1,2…k) is the maximum of all yk.

$$\boldsymbol{\overline{xik} = \sum\nolimits_{j\in ck}\frac{\overline{xij}}{nk}} \qquad (5)$$

Ck refers to class k that includes *nk* samples, *xij* is the expression value of gene i in sample j and *xik* is the mean expression value in class k for gene. *N* is total number of samples. *xi* is the general mean expression value for gene i. *si* is the pooled within-class standard deviation for gene i. The gene rank is calculated for the entire set of 4026 genes in Lymphoma dataset as shown in Table 4.

GENE ID	GENE ORDER
GENE1943X	0.2047
GENE880X	0.1842
GENE324X	0.1785
GENE1557X	0.1641
GENE2231X	0.1598
GENE289X	0.1569
GENE1792X	0.1559
GENE910X	0.1548

GENE272X	0.1547
GENE692X	0.1541

Table 4: List of genes with order

4.1.4 Finding informative genes

Finding informative genes greatly reduces the computational burden and noise arising from irrelevant genes. The genes are sorted and the genes with the highest gene are ranked from 1 to 100. Hundred out of 4026 genes with the highest gene rank are selected. Every gene is labelled after its importance rank. For example, Gene 1 means the gene ranked first as shown in Table 5. The genes with the highest scores are retained as informative genes.

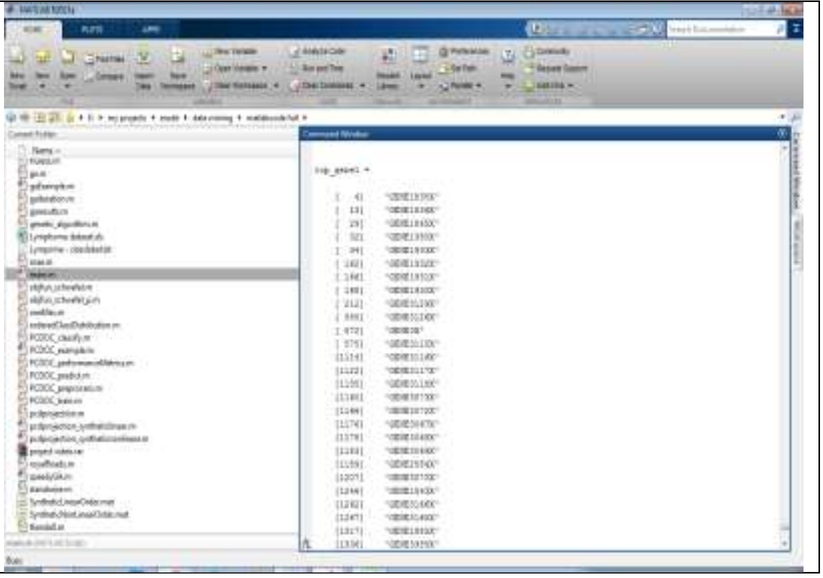
GENE ID	GENE ORDER	GENE RANK	
GENE1943X	0.2047	1	
GENE880X	0.1842	2	
GENE324X	0.1785	3	
GENE1557X	0.1641	4	
GENE2231X	0.1598	5	
GENE289X	0.1569	6	
GENE1792X	0.1559	7	
GENE910X	0.1548	8	
GENE272X	0.1547	9	
GENE692X	0.1541	10	

Table 5: Informative genes based on their rank

V.FINDINGS

Here the gene based lymphoma cancer data set using the machine learning algorithms Genetic Algorithm with Fuzzy based and Modified K -Means algorithm is used to classified and cluster the data .These algorithm have high efficient for clustering and classification of the data. When the lymphoma data set is loaded pre-processing is done for removal of noisy data and filtering necessary information and then genes are ranked according to their range of values.

The system was tested with 4026 genes of real dataset for different age groups are selected for verification and validation.

Genetic fuzzy and Modified K - means gives the high accuracy compared to other machine learning algorithms

The Modified K-Means shows accuracy of 92.4 and it takes the time period of 2.5 ms.

The Genetic with Fuzzy shows accuracy of 96.5 and it takes the time period of 1.6 ms.

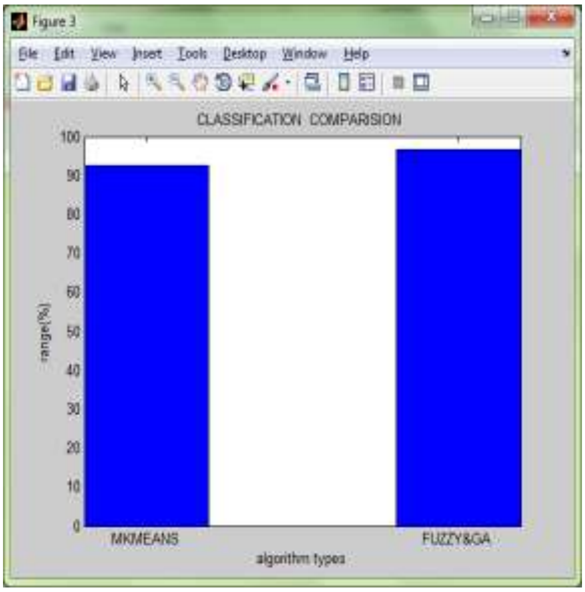


Figure 4: Accuracy comparison

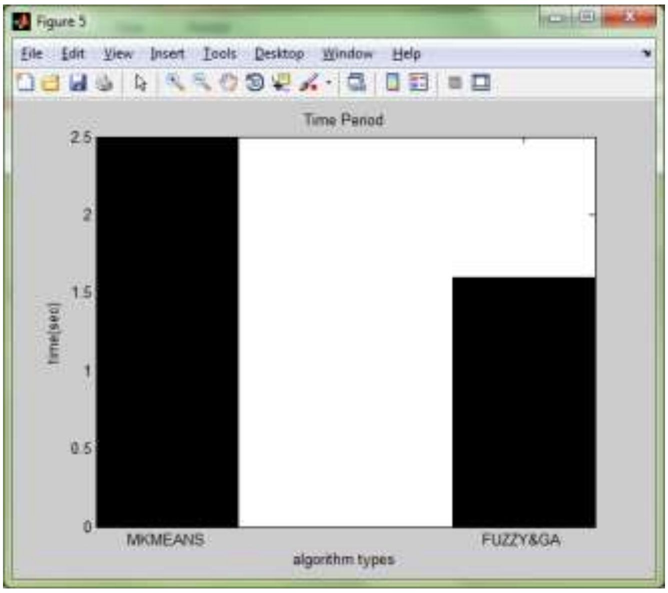


Figure 5: Time comparison

VI. CONCLUSION

Genetic Fuzzy and Modified K-Means algorithm gene selection method for proper classification of microarray data is proposed. A heuristic for selecting the optimal values of K efficiently, guided by the classification accuracy, is also proposed. In the earlier period, researchers used statistical methods to reduce the dimension of the dataset. After that they have applied some heuristic methods to select subset of informative genes. However, due to the reduced dimension, some informative genes would not take part in the heuristic method.

The proposed method is also compared to other gene selection methods applied on the lymphoma dataset and promising results are obtained. One of the most potential areas in applying microarray technology is clinical microbiology. It uses the low or middle density microarrays for the simultaneous assessment of large numbers of microbial genetic objects. The proposed method be speaks the possibility of developing simplified methods for an easy diagnosis of the cancer subgroups. Thereafter, histopathologists can identify the relevant genes efficiently and classify the blind test samples correctly.

VII.FUTURE ENHANCEMENT

Future work will be focused on using the other classification algorithms and prediction algorithm but some algorithms can’t find string and hexadecimal values of data mining. It is a known fact that the performance of an algorithm is dependent on the domain and the type of the data set. Hence, the usage of other classification algorithms like machine learning algorithms will be explored in future.

New microarray platforms should come up in tandem with the statistics and software for analysis and data mining technologies. Then only more precise and cheaper simplified technical and analytical procedures will benefit the patient in a large context.The main goal of this research consists of the exploration of new strategies and in the development of new clustering methods to improve the accuracy and robustness of clustering results, taking into account the uncertainty underlying the assignment of examples to clusters in the context of gene expression data analysis.

REFERENCES

[1] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieter’s, R., den Boer, M. L.,Minden,M.D., et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia/lymphoma. Nature Genetics, 30, 41–47.

[2] Bhattacharyya, C., Grate, L. R., Rizki, A., Radisky, D., Molina, F. J., Jordan, M. I., et al.(2003). Simultaneous classification and relevant feature identification in high dimensional spaces: Application to molecular profiling gene expression data. Signal Processing, 83, 729–743.

[3] Castillo, O., Melin, P., Ramirez, E., & Soria, J. (2012). Hybrid intelligent system for cardiac arrhythmia s with fuzzy K-nearest neighbors and neural Networks combined with a fuzzy system. Expert Systems with Applications, 39(3), 2947–2955.

[4] Cordon, O., Herrera, F., Villar, P (2001): Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base. IEEE Trans. Fuzzy System. 9(4), 667–674

[5] J. Li, X. Gao and L. Jiao, A new feature weighted fuzzy clustering algorithm, Acta Electronica Sinica, vol.34, no.1, pp.89-92, 2006.

[6] Lee, C. P., Lin, W. S., Chen, Y. M., & Kuo, B. J. (2011). Gene selection and sample

[7] Classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Systems with Applications, 38, 4661–4667.

[8] Li, S., Wu, X., & Tan, M. (2008). Gene selection using particle swarm optimization and genetic algorithm. Soft Computing, 12, 1039–1048.

[9] Melin, P., & Castillo, O. (2013). A review on the applications of type-2 fuzzy logic in Classification and pattern recognition. Expert Systems with Applications, 40(13), 5413–5423.

[10] Melin, P., Olivas, F., Castillo, O., Valdez, F., Soria, J., Mario, J., et al. (2013). Optimal

[11] Design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic. Expert Systems with Applications, 40(8).

[12] Shi .Y and Mizumoto .M, An improvement of neuro-fuzzy learning algorithm for tuning fuzzy rules,Fuzzy Sets and Systems, vol.118, no.2, pp.339-350, 2001.

[13] Wang, J. and X. Su, (2011). An improved K-means clustering algorithm, in 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an.