Classification of Microarray Gene Expression Data by Gene Combinations using Fuzzy Logic (MGC-FL)

V.Bhuvaneswari[1] and .Vanitha[2]

[1]Assistant Professor, Department of Computer Applications, Bharathiar University, Coimbatore, India
bhuvanes_v@yahoo.com
[2]M.Phil Research Scholar, Department of Computer Applications, Bharathiar University, Coimbatore, India
kvanithapraveen@gmail.com

Abstrct

Feature selection has attracted a huge amount of interest in both research and application communities of data mining. Among the large amount of genes presented in gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. Hence, one of the major tasks with the gene expression data is to find groups of co regulated genes whose collective expression is strongly associated with the sample categories or response variables. A framework is proposed in this paper to find informative gene combinations and to classify gene combinations belonging to its relevant subtype by using fuzzy logic. The genes are ranked based on their statistical scores and highly informative genes are filtered. Such genes are fuzzified to identify 2-gene and 3-gene combinations and the intermediate value for each gene is calculated to select top gene combinations to further classify gene lymphoma subtypes by using fuzzy rules. Finally the accuracy of top gene combinations is compared with clustering results. The classification is done using the gene combinations and it is analyzed to predict the accuracy of the results. The work is implemented using java language.

Keywords:

Feature selection, T-Test, Fuzzy, Classification, Clustering

1. INTRODUCTION

Data mining or knowledge discovery is the process of discovering meaningful, new correlation patterns and trends by shifting through large amount of data store in repositories, using pattern recognition techniques as well as statistical and mathematical techniques. Data mining is considered as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [13].

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.

DOI : 10.5121/ijcsea.2012.2409

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 importance ranking of each gene is done using a feature ranking measure called T-Test which ranks the genes based on their statistical score.

The method T-Test includes the classes with different samples. The mean value of each gene expression in a class is calculated. In fact, the TS (T-Scores) used here is a t-statistic between the centroid of a specific class and the overall centroid of all the classes. The T-scores of the genes are sorted and the genes with the highest T-scores are ranked from 1 to 100. The genes with the highest scores are retained as informative genes which are used for gene combinations.

Fuzzy logic is a superset of conventional Boolean logic. Fuzzy logic, unlike other logical systems, deals with imprecise or uncertain knowledge. The set of informative genes with gene expression data are converted into fuzzy values using Type 1 fuzzy. The different gene combinations are identified and intermediate value is calculated for each gene combination. Further, the lymphoma subtypes are classified based on the fuzzy rules on a test dataset.

The fuzzified informative genes are used to find out gene combinations which are used for classifying the dataset to find its lymphoma subtypes. Specifically Single gene, Two-gene and Three-gene combinations are done with the selected informative genes. The purpose of generating gene combinations is to find out whether it will classify lymphoma subtypes.

A fuzzy rule involves a fuzzy condition and a fuzzy conclusion. The intermediate values calculated for single gene, two gene and three gene combinations are used to frame fuzzy rules to classify the lymphoma subtypes such as DLBCL, FL and CLL of the test dataset. The test dataset consists of hundred random genes and it is selected from the whole dataset of 4026 genes with its samples.

Clustering is the process of organizing objects into groups whose members are similar in some aspects. Here the gene combinations such as two gene and three gene combinations are grouped into a set of disjoint classes, called clusters so that genes within a class have high similarity to each other, while genes in separate classes are more dissimilar. Finally gene combinations are verified and its correlation is compared with hierarchical clustering approach by grouping the entire informative genes. Then the classification accuracy of the gene combination is analyzed based on its efficiency of subtype’s classification such as DLBCL, FL and CLL of the test dataset.

This paper is organized as follows. Section 2 provides the literature study of the various Feature selection methods, Gene classification and Fuzzy logic for Bio-logical database. Section 3 explores the methodology for Microarray Gene classification using Fuzzy Logic (MGC-FL). In Section 4 the implemented results are verified and validated. The final section draws the conclusion of the paper.

2. REVIEW OF LITERATURE

In [3] Qinghua Huang et al., (2011) have discussed the importance of feature selection. The objective of feature selection is to find optimal or suboptimal subsets from the original feature sets for irrelevant features removal, intrinsic class information preservation.

In [15] Patharawut Saengsiri et al., (2011) have provided the benefits of Feature Selection. They proposed three feature selection methods. They are Correlation based Feature Selection, Gain ratio and Information gain. The concept of Correlation based Feature Selection is relevance of feature and target class that is based on heuristic operation. Gain Ratio technique improves the problem of Information gain. Gain Ratio is based on evaluation of information theory.

In [17] the author Alok Sharma et al., (2011) have proposed a feature selection algorithm for classification problem using transcriptome data. The proposed algorithm explores and provides a way to investigate important genes. It is observed that the algorithm finds a small gene subset that provides high classification accuracy on several DNA microarray gene expression datasets.
In [6] Yan-Fei Wang et al., (2011) proposed a type-2 fuzzy membership test (Type-2 FM test) for disease-associated gene identification on microarrays to improve traditional fuzzy methods. The results showed that type-2 FM test performs better than traditional fuzzy methods when analyzing microarray data with similar expression values and noise.

In [7] Pablo Martin-Munoz et al., (2010) presented a new algorithm, FuzzyCN2, for extracting conjunctive fuzzy classification rules. This algorithm produced an ordered list of fuzzy rules. In [20] Yan-Fei Wang et al., (2010) proposed to combine the FCM method with the empirical mode decomposition (EMD) for clustering microarray data in order to reduce the effect of the noise. It was called as fuzzy C-means method with empirical mode decomposition (FCM-EMD).

In [4] Lipo Wang et al., (2010) discussed ranking of genes using two methods called T-Score (TS) and Class Separability CS). All genes in the training data set are ranked using a certain ranking criterion and small numbers of highly ranked genes are retained. In T-Test statistical method the T-Scores are calculated for each gene and gene with highest T-score is selected.

In [16] Wutao Chen,Huijuan Lu et al., (2009) compared various feature selection methods in selecting informative genes. It is choosing genes which have expression levels of high diversity in different types of samples. Among the various feature selection methods, such as SNR, *t*-test, Fisher and information gain, *t*-test has been proved to be an effective method in the binary-classification problem.

In [11] Zarita Zainuddin et al., (2009) have discussed about Microarray Data Preprocessing. Microarray data consists of an overwhelming number of genes relative to the number of samples. However, the majority of such genes are probably irrelevant in discriminating between the subclasses of the heterogeneous cancers. Hence, genes selection is a crucial aspect in microarray data analysis.

In [12] Wutao Chen et al., (2009) has introduced classification of gene expression data using artificial neural network based on samples filtering. Simulation tests were carried out to verify the proposed strategy using Leukemia data sets, and the test results were compared with those of single artificial neural network.

In [9] Jahangheer Shaik et al., (2009) presented Fuzzy-Adaptive-Subspace-Iteration-based Two-way Clustering (FASIC) of microarray data to find differentially expressed genes from two-sample microarray experiments. In [10] Keon Myung Lee et al., (2009) introduced three fuzzy set-based microarray data analysis techniques used to find local cluster, to locate contrasting group, and to filter group with specific pattern.

In [19] Mingrui Zhang et al., (2009) evaluated several validity measures in fuzzy clustering and developed a new measure for a fuzzy c-means algorithm which uses a Pearson correlation in its distance metrics. In [18] this paper Ming Chen et al., (2008) focused on a method of optimizing classifiers of neural network by Genetic Algorithm based on the principle of gene reconfiguration, and implemented classification by training the weight.

In [5] Qingzhong Liu et al., (2006) have presented a scheme of recursive feature addition for gene selection and combined classifiers for the purpose of classifying tumor tissues using DNA microarray data. In [8] Nilesh N. Karnik et al., (1999) introduced a type-2 fuzzy logic system (FLS), which handled rule uncertainties. It involved the operations of fuzzification, inference, and output processing.

3. PROBLEM FORMULATION AND METHODOLOGY

The proposed framework Microarray Gene Classification using Fuzzy Logic (MGC-FL) given in Figure 1 is used to find informative gene combinations and to classify gene combinations belonging to its relevant subtype by using fuzzy logic. In the initial phase the noisy data is removed and genes are ranked based on their statistical scores. The highly informative genes are filtered based on ranking of genes. In the classification phase informative genes are fuzzified and identified for 2-gene and 3-gene combinations. The intermediate value for gene combination is calculated to classify gene lymphoma subtypes by using fuzzy rules. In the final phase top gene combinations are compared with clustering and the classification accuracy of gene combinations is analyzed.

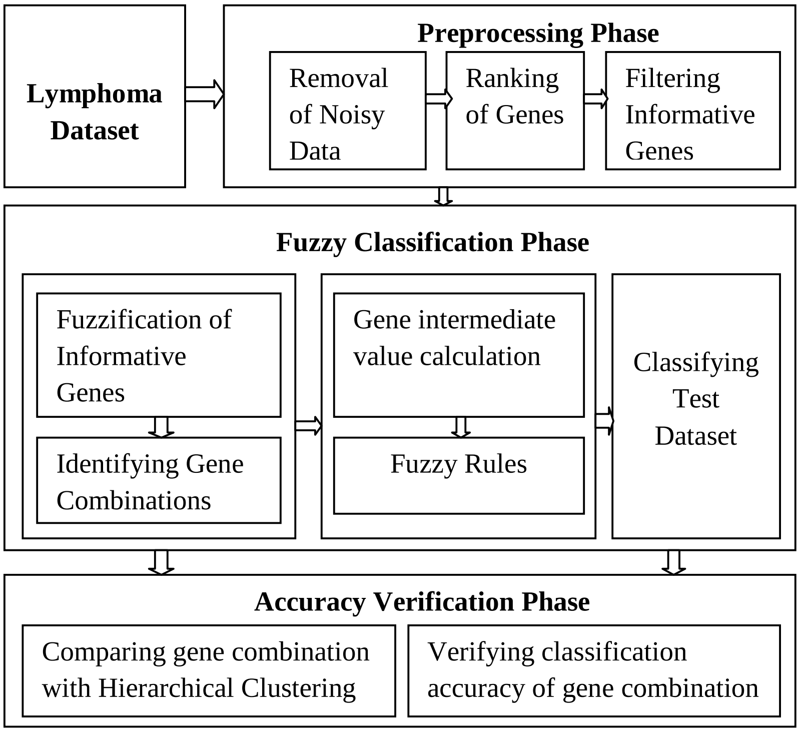

Lymphoma Dataset
Preprocessing Phase
Removal of Noisy Data
Ranking of Genes
Filtering Informative Genes
Fuzzy Classification Phase
Fuzzification of Informative Genes
Identifying Gene Combinations
Gene intermediate value calculation
Fuzzy Rules
Classifying Test Dataset
Accuracy Verification Phase
Comparing gene combination with Hierarchical Clustering
Verifying classification accuracy of gene combination

Figure 1. Framework for Microarray Gene Classification using Fuzzy Logic (MGC-FL)

3.1 Dataset

Microarrays 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] as shown in Table 1. Human B-Cell contains about 4026 genes expressed in lymphoid cells or which are known as immunological or oncological importance with 96 conditions. There are three types of lymphomas such as diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), and chronic lymphocytic leukaemia (CLL) [1]. The entire data set includes the expression data of 4,026 genes each measured using a specialized cDNA microarray with its relevant Genbank accession number, Name and Clone IDs. A part of the dataset is chosen for the proposed work to classify lymphoma subtypes consists of hundred genes with gene expression values of 62 samples, with a total of 6200 samples and it is called as the Test dataset.

Table 1. A Sample data from Lymphoma Dataset

GENE ID	**NAME**	**VALUES**	**VALUES**	**VALUES**
GENE3129X	Autocrine motility factor receptor Clone=1072873	-0.3000	0.3000	0.5900
GENE3126X	2B catalytic subunit Clone=637173	-0.2200	-1.2100	1.4100
GENE3072X	APC Clone=125294	-0.0400	0.1500	0.6800
GENE3067X	Probable ATP Clone=1350869	0.4100	-0.3400	-0.1800
GENE4006X	SRC-like adapter protein Clone=701768	1.7600	1.2100	0.9900

3.2 Preprocessing

Data pre-processing is an often neglected but important step in the data mining process. Preprocessing 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 preprocessing phase.

Table 2. Lymphoma Dataset with empty spots

GENE	**NAME**	**VALU**	**VALU**	**VALU**	**VALU**
GENE1835X	(Clone=1357915)	-0.1300		-0.2800	0.0400
GENE1836X	(Clone=1358277)	-0.3100	0.1600		0.2500
GENE1865X	(Clone=1358064)	-0.1200	0.5200		0.8300
GENE1933X	(Clone=1358190)	0.0500			0.2800
GENE1932X	(Clone=1336836)	-0.2600		-0.0900	0.1500
GENE1931X	(Clone=1336983)	-0.5500			

3.2.1 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 knnimpute method. It replaces NaNs in 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.

Table 3. Preprocessed Lymphoma Dataset

GENE	**NAME**	**VALU**	**VALU**	**VALU**	**VALU**
GENE1835X	(Clone=1357915)	-0.1300	-0.2800	-0.2800	0.0400
GENE1836X	(Clone=1358277)	-0.3100	0.1600	0.1600	0.2500
GENE1865X	(Clone=1358064)	-0.1200	0.5200	0.5200	0.8300
GENE1933X	(Clone=1358190)	0.0500	0.0500	0.0500	0.2800
GENE1932X	(Clone=1336836)	-0.2600	-0.0900	-0.0900	0.1500
GENE1931X	(Clone=1336983)	-0.5500	-0.5500	-0.5500	-0.5500

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

3.2.2 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 called T-Test which ranks the genes based on their statistical score. The t-test 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. T-Test 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 T-Score of gene 'i' is defined as

$$Tsi = \max\left\{ \left| \frac{\bar{x}ik - \bar{x}i}{mksi} \right| \; k = 1,2...k \; \right\} \longrightarrow \qquad \text{Eq.(1)}$$

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

$$\bar{x}ik = \sum\nolimits_{j\in ck} \bar{x}ij \Big/ nk \longrightarrow \qquad \text{Eq.(2)}$$

Ck refers to class k that includes *nk* samples, *xij* is the expression value of gene i in sample j and $\bar{x}ik$ 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 T-scores is calculated for the entire set of 4026 genes in Lymphoma dataset as shown in Table 4.

Table 4. List of genes with T-scores

GENEID	T-SCORE
GENE1943	0.2047
GENE880	0.1842
GENE324	0.1785
GENE1557	0.1641
GENE2231	0.1598
GENE289	0.1569
GENE1792	0.1559
GENE910	0.1548
GENE272	0.1547
GENE692	0.1541

3.2.3 Finding informative genes

Finding informative genes greatly reduces the computational burden and noise arising from irrelevant genes. The T-scores of the genes are sorted and the genes with the highest T-scores are ranked from 1 to 100. Hundred out of 4026 genes with the highest T-Scores are selected. Every gene is labeled 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.

Table 5. Informative genes based on their T-scores

GENEID	T-	GENE
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

The set of informative genes are passed as input to the next phase for fuzzy classification.

3.3 Fuzzy Classification

In this phase the set of informative genes with gene expression data are converted into fuzzy values using Type 1 fuzzy. The different gene combinations are identified and intermediate value is calculated for each gene combination. Further, the lymphoma subtypes are classified based on the fuzzy rules on a test dataset.

3.3.1 Fuzzification of Informative Genes

Gene expression data is quantitative and it contains numerical values. The numeric values are converted into fuzzy linguistic variables and terms using the concept of fuzzy set. The fuzzification process includes Type-1 fuzzy. The first step in fuzzification is to take the crisp inputs, i.e. gene expression data and covert to fuzzy values. The second step is to take the fuzzified inputs, and apply them to the antecedents of the fuzzy rules. In Type-1 fuzzy the constant value is calculated and the gene expression states are represented by the constant values. In type-2 fuzzy appropriate ranges are provided for the fuzzified values. The states used in Type-1 fuzzy are Low, Average and High. The maximum and minimum values in each parameter is calculated and sorted in ascending order. The value of p_{jk} calculated and sorted in ascending order. The value of p_{jk} is calculated by using the equation,

$$P_{jk} = low_i + \frac{R_k - Cf_{i-1} \quad * 8}{F_i}$$

Where low_i is the lower limit of the ith class interval, R_k is the rank of the k th partition value, Cf_{i-1} is the cumulative frequency.

if Aj1<eij<Aj2

$$\mu A = \frac{Aj3 - eij}{Aj3 - Aj2} \quad \text{---------} \quad \text{if } Aj2<eij<Aj3$$

$$\mu H = \frac{eij - Aj2}{Aj3 - Aj2} \quad \text{----------} \quad \text{if } eij<Aj2$$

By using the above equations the lymphoma gene expression data as shown in Table 6 is converted into fuzzy values and it is shown in Table 7.

Table 6. Lymphoma Gene Expression values

GENE ID	VALU	VALU	VALU	VALUE
GENE1943X	0.4600	0.2100	-0.0100	-0.3400
GENE880X	0.8900	0.7700	0.3000	0.6000
GENE324X	0.4600	0.0200	-0.0200	-0.5400

Table 7. Informative Gene Data with Type-1 fuzzy values

GENE ID	VALU	VALUE	VALUES	VALUE
GENE1943X	0.5210	0.9115	0.1386	0.4190
GENE880X	1	0.0414	0.7717	0.3056
GENE324X	0.5231	0.1131	0.1471	0.5890

The fuzzified informative genes are passed as input to the next process to identify various gene combinations.

3.3.2 Identifying gene combinations

The fuzzified informative genes are used to find out gene combinations which are used for classifying the dataset to find its lymphoma subtypes. Specifically Single gene, Two-gene and Three-gene combinations are done with the selected informative genes. The Single gene, two gene and three gene combinations are identified to classify the lymphoma subtypes such as

DBLCL, FL and CLL. The single gene is identified from the whole informative gene set which consists of 100 genes.

3.3.3 Gene Intermediate value calculation (IVC)

The intermediate value is the arithmetic mean commonly known as standard average. As shown in Table 8 let the set of informative genes be IG1, IG2…IGn, the standard averages is calculated for all subtypes of a gene as $\bar{x}$1, $\bar{x}$2, $\bar{x}$3.

Table 8. Intermediate Value Calculation

Informative	Intermediate Values for IG	Test Genes
IG1	$\bar{x}$1, $\bar{x}$2, $\bar{x}$3	TG1,TG2,
IG2…IGn	$\bar{x}$1, $\bar{x}$2, $\bar{x}$3….$\bar{x}$1n, $\bar{x}$2n,	TG1,TG2,

The intermediate values $\bar{x}$1, $\bar{x}$2, $\bar{x}$3 are used as ranges to classify all the subtypes of the test genes such as TG1,TG2,..TGN in the test dataset. The intermediate values for the single gene, two gene and three gene combinations are calculated in this process to classify the lymphoma subtypes. Table 9 shows intermediate values calculated for individual informative gene.

Table 9. Intermediate values for Single Gene

SINGLE	**INTERMEDIATE**		
Gene1	**0.1755**	**-0.1118**	**0.1237**
Gene 2	**0.0983**	**-0.1089**	**-0.6307**
Gene 3	**-0.0108**	**-0.5238**	**-0.3781**
Gene 4	**0.2922**	**-0.3718**	**0.2934**
Gene 5	**-0.1798**	**0.6440**	**0.7270**

A sample of intermediate values calculated for two gene combinations is shown in Table 10 and for three gene combinations it is shown in Table 11.

Table 10. Intermediate values for 2GC

2G	INTERMEDIATE VALUE		
(1,2)	0.1369	-	-0.2535
(2,4)	0.0824	-	-0.1272
(3.7)	0.0738	-	-0.5266
(4,11)	0.1045	-	0.1788
(23,40)	0.0608	-	-0.2647

Table 11. Intermediate values for 3GC

3GC	INTERMEDIATE VALUE		
(1,2,3)	0.0877	-0.2482	-0.2950
(2,13,38)	0.1832	-0.1300	0.2570
(3,27,69)	0.1255	-0.1543	0.0274
(4,7,59)	0.2225	-0.1150	0.4751
(59,71,94)	0.1224	-0.1541	-0.1615

The intermediate values are calculated for top gene combinations to frame fuzzy rules and to classify the lymphoma subtypes in the test dataset.

3.3.4 Fuzzy Rules

A fuzzy rule involves a fuzzy condition and a fuzzy conclusion. The test dataset consists of hundred random genes and it is selected from the whole dataset of 4026 genes with its samples. It is converted to fuzzy values as shown in Table 12. The genes included in the test dataset are not selected as top genes in informative genes set.

Table 12. Sample Test Data from Lymphoma Dataset

GENEID	**VALUES**	**VALUES**	**VALUES**	**VALUES**
GENE143X	-0.5224	-0.1563	-0.3851	-0.1929
GENE141X	-0.5407	-0.1425	-0.3989	0.4155
GENE3844X	-0.0922	-0.1975	-0.0876	1.0000
GENE1400X	0.0568	0.5622	-0.0327	-0.0373
GENE137X	-0.4858	-0.4675	-0.4080	0.9208

The three lymphoma subtypes are identified by a specific fuzzy rule by assigning intermediate value ranges. A single gene, 2GC and 3GC classifies all genes included in the test dataset and individual count of the relevant lymphoma subtypes is displayed as it is shown in Figure 2. The subtypes not classified under mentioned lymphoma subtypes is grouped under other subtypes.

Classifying Gene	Gene to be Classified	Lymphoma Subtypes			
		DBLCL	FL	CLL	Other
Single Gene					
2GC	Gene x1...	Count(DBLCL)	Count(FL)	Count(CLL)	Count(Other)
3GC	Gene xn	…..n	….n	….n	….n

Figure 2. Classifying test dataset using fuzzy rule

4. IMPLEMENTATION RESULTS AND DISCUSSION

According to the subtype limits given in base paper [14] in the Lymphoma dataset there are 62 samples for a gene, out of which 42 samples are of DLBCL, 9 samples are FL and 11 samples are CLL. A single informative gene is used to classify subtypes in the test dataset. A single gene GENE3 classified the subtypes of each gene in the dataset, and the count displayed under the subtypes is the count of DLBCL's, FL's and CLL's classified in the total expression values of a specific gene in the test dataset. The gene expression values which are not classified as DLBCL, FL and CLL are classified into other lymphoma subtypes. The single gene GENE3 classification on test dataset is shown in Table 13.

Table 13. Single Gene Classification in Test Dataset

SINGLE GENE- GENE 3				
GENE TO BE CLASSIFIED	**DLBCL**	**FL**	**CLL**	**OTHER SUBTYPES**
GENE3852X	32	**9**	19	2
GENE3844X	34	**9**	19	0
GENE3845X	34	16	**11**	1
GENE3846X	33	**9**	20	0
GENE1126X	**41**	13	8	0
GENE1127X	**40**	14	8	0

A single gene GENE3 classified the whole test dataset out of which the DLBCL subtype classification on GENE3846X and GENE1126X were nearest to the subtype limit. The FL subtype classification on GENE3852X, GENE3846X and GENE3844X is equal to the subtype limit. The CLL subtype classification on GENE3845X was also equal to the subtype limit. The single gene GENE3 classified all the lymphoma subtypes and the classification of DLBCL subtype of all 100 genes in the test dataset is within subtype limit i.e. DLBCL count of all individual genes in the test dataset were not above the subtype limit 42. The total subtype's classification in the entire test dataset i.e. for 6200 samples by list of single informative genes is shown in Table 14.

Table 14. Classification in Test Dataset by Single Gene

GENE	LYMPHOMA SUBTYPES			
	DLBCL	FL	CLL	OTHER
GENE 3	**3074**	1995	946	185
GENE	**3041**	**2818**	156	185
GENE	**3108**	1448	1459	185
GENE	**3041**	2123	851	185
GENE	**3249**	263	2503	185

The single gene GENE3 classified the entire test dataset with 3074 DLBCL subtypes, 1995 FL subtypes and 946 CLL subtypes from the total of 6200 samples. Most of the single genes classified DLBCL subtype within the subtype limit. The single gene GENE50 also classified DLBCL and FL subtype within the subtype limit. It classified DLBCL and also FL with good accuracy. The single gene GENE50 classified DLBCL and FL subtype within the subtype limit and GENE3 classified DLBCL within subtype limit. The comparison of single gene GENE50 and GENE3 is pictorially represented in Figure 3.

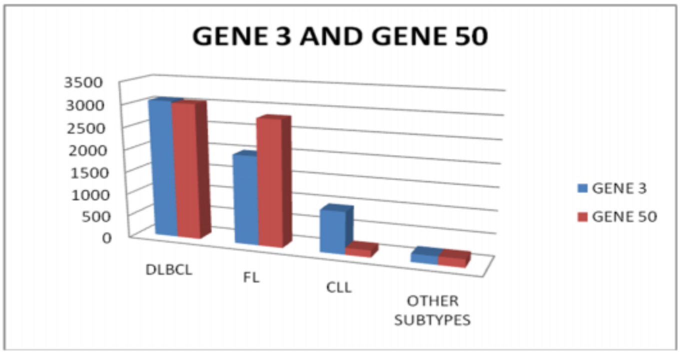


Figure 3. Comparison of single genes GENE3 and GENE50

The single gene GENE67 classified only the DLBCL's, CLL's and was unable to classify FL subtype in the test dataset. The single gene GENE67 classified GENE141X expression values into 31 DLBCL's, 0 FL's , 22 CLL's, 9 other subtypes and also classified other genes in the test dataset as shown in Table 15.

Table 15. Classification in Test dataset by single gene GENE67

SINGLE GENE - GENE 67				
GENES TO BE	**DLBCL**	**FL**	**CLL**	**OTHER SUBTYPES**
GENE141X	31	0	22	9
GENE1127X	47	0	15	0
GENE1126X	42	0	20	0
GENE1583X	38	0	22	2

The same gene is combined with other genes to find out whether it classifies all lymphoma subtypes. The single gene GENE67 is combined with another gene i.e. in a two gene combination in the next process. The single gene combined with another gene may or may not classify the lymphoma subtypes due to the cooperation of the genes. The single gene GENE67 is combined with GENE3 to find out whether it classifies all lymphoma subtypes as shown in Table 16.

Table 16. Two gene combination which classified Lymphoma subtypes

GENE	GENES TO BE	LYMPHOMAS			
		DLB	FL	CL	OTHER
[GENE	GENE1126	42	6	14	0
	GENE137X	40	5	17	0
	GENE1127	41	1	20	0
	GENE142X	30	7	20	5

The gene GENE67 which was unable to classify FL subtype as a single gene, classified it when combined with another gene GENE3 which is one of the best gene in classifying all subtypes. The single gene GENE3 is combined with another gene GENE1 to classify the test dataset. The purpose of combination is to find out whether it will classify all lymphoma subtypes. The classification of subtypes in test dataset by GENE3 and GENE1 is shown in Table 17.

Table 17. Classification in Test Dataset by two genes

TWO GENE COMBINATION [GENE 3, GENE 1]				
GENE TO BE CLASSIFIED	**DLBCL**	**FL**	**CLL**	**OTHER SUBTYPES**
GENE3851X	30	5	27	0
GENE3844X	32	7	23	0
GENE3845X	32	5	24	1
GENE1128X	34	7	21	0
GENE1129X	32	7	23	0
GENE1583X	36	2	22	2
GENE1125X	36	5	21	0

The two gene combination GENE3 and GENE1 classified the whole test dataset out of which the DLBCL subtype classification on GENE1583X and GENE1125X were nearest to the subtype limit 42. The FL subtype classification on GENE3850X, GENE3844X, GENE1128X and GENE1129X is nearer to the subtype limit i.e. 9. The two gene combination was able to classify CLL subtype but not within the subtype limits.

The two genes GENE3 and GENE1 classified all the lymphoma subtypes and the classification of DLBCL subtype of all 100 genes in the test dataset is within subtype limit i.e. DLBCL counts of all individual genes in the test dataset were not above the subtype limit 42. The total subtype classification in the entire test dataset (i.e. for 6200 samples) by two gene combinations (GENE3, GENE 1), (GENE3, GENE67) and (GENE23, GENE40) is shown in Table 18.

Table 18. Gene Classification in test dataset by two gene combinations

GENE	LYMPHOMA SUBTYPES			
	DLB	FL	CLL	OTHER
(GENE3,GENE1	2585	987	2443	185
(GENE23,	2674	1525	1816	185
(GENE3,	3249	371	2395	185

The two gene combinations (GENE3, GENE1), (GENE3, GENE67) and (GENE23, GENE40) are compared and (GENE3, GENE67) combination is considered to be best because of its high DLBCL classification when compared to other combinations and it is graphically depicted in Figure 4.

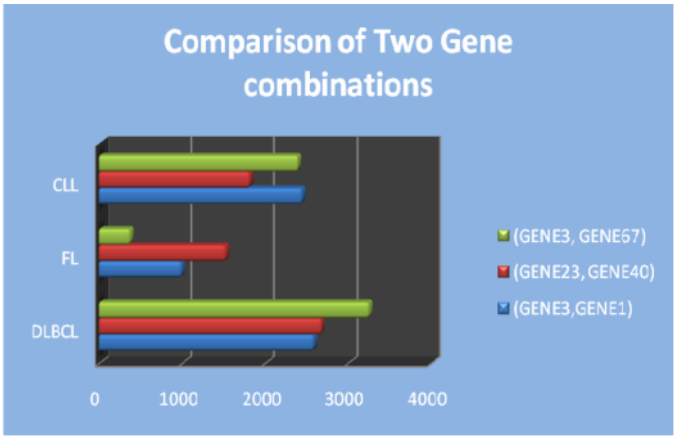


Figure 4. Comparison of two gene combinations

When the two gene combinations are unable to classify all subtypes, the reason may be due to poor cooperation of genes and as a try three gene combinations are made to classify all subtypes within subtype limits. The three genes (GENE1, GENE2, and GENE3) are used as a three gene combination to classify lymphoma subtypes in the test dataset. Table 19 shows a snapshot of classification of subtypes in test dataset by GENE1, GENE2 and GENE3.

Table 19. Classification of Test dataset by three gene combinations

THREE GENE COMBINATION [GENE1, GENE2,				
GENE TO BE	**DLBCL**	**FL**	**CLL**	**OTHER SUBTYPES**
GENE3850X	29	9	24	0
GENE3852X	29	7	24	2
GENE3844X	32	7	23	0
GENE1126X	34	17	11	0
GENE1128X	34	17	11	0
GENE1583X	36	6	18	2
GENE1125X	36	13	13	0

The three gene combination GENE1, GENE2 and GENE3 classified the whole test dataset out of which the DLBCL subtype classification on GENE1583X, GENE1125X were nearest to the subtype limit 42. The FL subtype classification on GENE3850X, GENE3844X and GENE3852X is nearer to subtype limit 9. The CLL subtype classification on GENE1126X, GENE1128X is equal to subtype limit 11. When compared to other combinations the three genes combination classified CLL subtype accurately for some genes in the test dataset.

The total subtype classification in the entire test dataset (i.e. for 6200 samples) by three gene combinations (GENE59, GENE71, GENE94), (GENE98, GENE89, GENE32) and (GENE1, GENE2, GENE3) is shown in Table 20. The total count of DLBCL's, FL's and CLL's in the test dataset is classified and displayed under relevant subtype columns.

Table 20. Three Gene Classifications on Total Test Dataset

GENE	LYMPHOMA SUBTYPES			
	DLBCL	FL	CL	OTHERS
(GENE 59, GENE 71, GENE 94)	**2449**	1301	2265	185
(GENE 98, GENE 89,GENE 32)	**2632**	42	3341	185
(GENE1,GENE2,GENE3)	**2585**	1712	**1718**	185

The three gene combinations (GENE59, GENE71, GENE94), (GENE98, GENE89, GENE32) and (GENE1, GENE2, GENE3) are compared and it is graphically depicted in Figure 5.

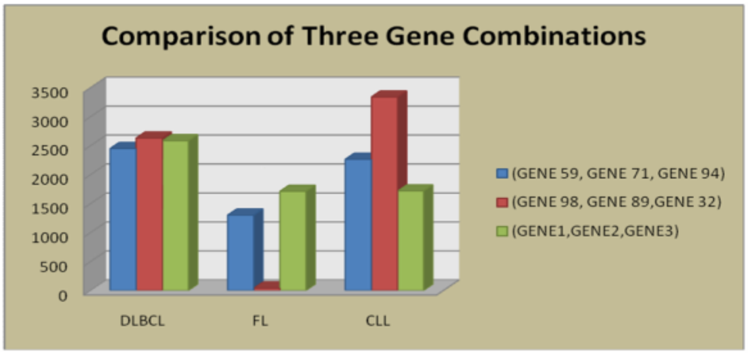


Figure 5. Comparison of three gene combinations

The single gene, two gene and three gene combinations is taken to the next process to find correlation between gene combinations. Finally in the accuracy verification phase the 2GC and 3GC are verified and its correlation is compared with hierarchical clustering approach by grouping the entire informative genes. Clustering is the process of organizing objects into groups whose members are similar in some aspects. Here the gene combinations such as 2GC and 3GC are grouped into a set of disjoint classes, called clusters as shown in Figure 6.

Clusters	Genes	Count(Genes)
1	x1…. xn	Count(x1…. xn)
2	x2…. xn	Count(x2…. xn)
…n	Xn….xn	Count(Xn….xn)

Figure 6. Clustering

The informative genes are grouped under 10 clusters. The entire informative gene dataset is passed as input to clustering to compare gene combinations. Table 21 gives a snapshot of the total number of clusters and informative genes included in each clusters.

Table 21. Clusters and Informative Genes Included in Each Cluster

Cluster No	No. of Genes	Genes
1	1	25
2	89	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24, 26,30,32,33,34,37,38,39,40,41,43,44,45
3	1	27
4	2	28,31
5	1	35
6	1	61
7	1	58
8	2	42,47
9	1	36
10	1	93

The two gene combinations such as [23, 40], [3, 67] ,[3,1] and three gene combinations such as [98, 89,32] , [59, 71,94] and [1,2,3] that are selected as gene combinations for classifying lymphoma subtypes belongs to the same clusters and because of its correlation it was able to classify all subtypes and Gene 3 and Gene 67 classified DLBCL subtypes within the subtype limits. The three gene combination [1, 2, and 3] classified DLBCL and CLL subtypes for some genes within the subtype limits. The gene combination which was taken in the proposed work to classify the lymphoma subtypes belongs to the same clusters and it proved its correlation. There are some genes which was unable to classify all subtypes because of its poor cooperation. The clusters and the total number of genes included inside the cluster is pictorially depicted in Figure 6.

.

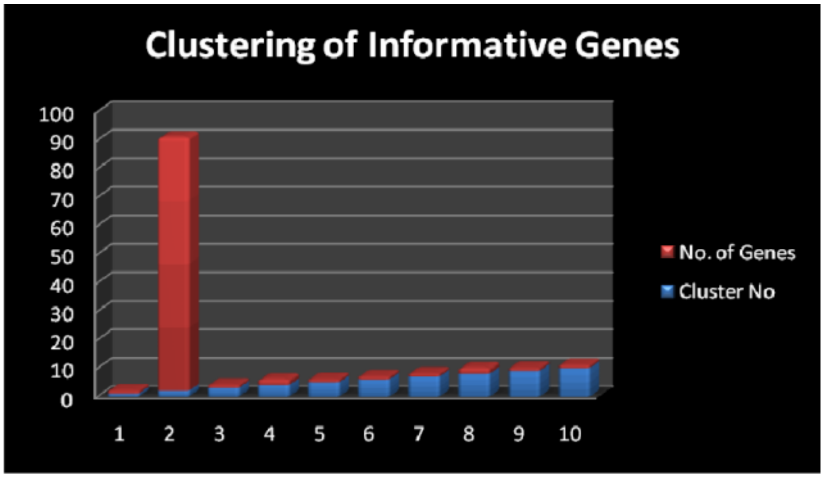


Figure 6. Clustering

A single gene, 2GC and 3GC classification accuracy is verified in this final phase. The accuracy is calculated based on the gene’s classification ability to classify all subtypes of lymphomas. Most of the informative genes are able to classify all lymphomas.

The single informative gene classified most of the genes in the dataset and some genes were unable to classify all subtypes. So probably 2GC and 3GC were used to find out whether it classifies all the subtypes of lymphoma. The single gene which was unable to classify all subtypes when combined with another gene as a 2GC may classify all lymphoma subtypes. The accuracy of classification is verified for all single genes, 2GC and 3GC respectively based on the subtypes classification. Some genes from the selected gene combinations accurately classified DLBCL not crossing its sample limits.

The single gene classified lymphoma subtypes for several genes in the test dataset. Genes such as GENE1, GENE2, GENE3, GENE4, GENE6, GENE7, GENE8, GENE9, GENE50, GENE55, GENE84 and GENE96 classified DLBCL subtypes of all the hundred genes within subtype limit.GENE96 attained 77% accuracy in classifying DLBCL subtype for Gene (GENE1126X) in the test dataset which is equal to subtype limit 42. Similarly GENE55 attained 74% accuracy in classifying DLBCL subtype for Gene (GENE1126X) in the test dataset which is very nearer to subtype limit 42. Table 22 shows list of single genes with their classification accuracy.

Table 22.Single Gene Classification Accuracy

GENE	LYMPHOMA SUBTYPES		
	DLBCL	FL	CLL
GENE50	72%	67%	5%
GENE55	74%	34%	35%
GENE84	72%	51%	20%
GENE96	77%	7%	60%

The best genes are **GENE96** and **GENE55** in classifying DLBCL subtype and **GENE50** attained 67% accuracy in classifying FLL subtype for Gene (GENE100X) in the test dataset which is nearer to subtype limit 9.

The two gene combination classified lymphoma subtypes for several genes in the test dataset. (GENE3,GENE1),(GENE23,GENE40), and (GENE3, GENE67) classified DLBCL subtypes of all the hundred genes within the subtype limit. (GENE3, GENE67) attained 77% accuracy in classifying DLBCL subtype for Gene (GENE1126X) in the test dataset which is nearer to subtype limit 42 and also classified FLL subtype within the limit. Similarly (GENE23, GENE40) attained 64% accuracy in classifying DLBCL subtype and (GENE3, GENE1)attained 62% classification accuracy as shown in Table 23.

Table 23. Classification Accuracy of two gene combinations

GENE	LYMPHOMA		
	DLB	FL	CL
(GENE3,GENE	62%	24%	58
(GENE23,	64%	36%	43
(GENE3,	77%	50%	57

The best two gene combination which classified DLBCL and FLL within subtype limits is (GENE3,GENE67) and other combinations is also best in classifying DLBCL subtype. The three gene combination classified lymphoma subtypes for several genes in the test dataset. (GENE 98, GENE 89, GENE 32) classified DLBCL and FL accurately within the constraint and it attained 63% accuracy. (GENE1, GENE2, GENE3) classified DLBCL subtypes of all the hundred genes within the subtype limit and attained 62% accuracy. (GENE 59, GENE 71, GENE 94) attained 58% accuracy in classifying DLBCL subtype within the limit. The three gene combinations and its accuracy are displayed in the Table 24.

Table 24. Classification Accuracy of three gene combinations

GENE	LYMPHOMA		
	DLBC	FL	CL
(GENE 59, GENE 71,	58%	31%	54
(GENE 98, GENE	63%	50%	80
(GENE1,GENE2,GENE	62%	41%	41

The three gene combination **(GENE 98, GENE 89, and GENE 3**2) is the best one to classify DLBCL and FL within the subtype limits. All other combinations can be used to classify DLBCL subtype. The classification accuracy of single gene, two gene and three gene combination is graphically depicted in Figure 7.

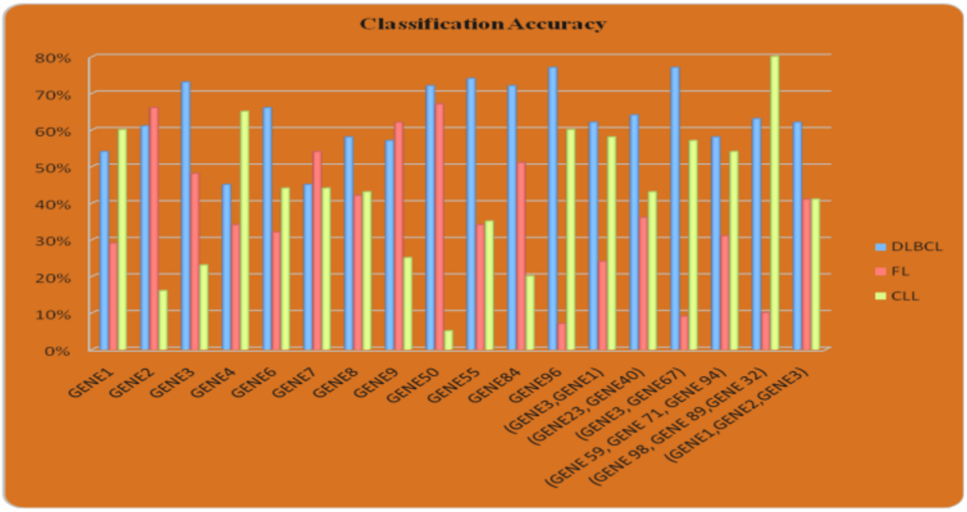


Figure 7. Classification Accuracy of single gene, two genes and three genes

From the experimental results it was found that from the top hundred genes ranked based on T-scores, single gene selected classified 77% of DLBCL subtypes, 67% of FLL subtypes for all hundred genes in the test dataset, two gene combinations was found to have 77% of DLBCL subtypes, 50% of FL subtypes for all hundred genes in the test dataset and three gene combinations was found to have 63% of DLBCL subtypes, 50% of FL subtypes for all hundred genes in the test dataset.

5. CONCLUSION AND FUTURE DIRECTIONS

Bioinformatics and data mining are developing as interdisciplinary science. The proposed methodology consists of a framework (MGC-FL) for improving the proposed idea. The main idea of the proposed work is classification of gene expression data based on subtypes using fuzzy logic. Among the large amount of genes present in gene expression data, only a small fraction of them is effective for performing classification. Such informative genes are retained by a process called feature selection. The proposed 2-gene and 3-gene combinations are verified with a clustering approach called Hierarchical clustering which proved that gene combination taken are good combinations in classifying lymphoma subtypes. The classification accuracy of gene combination is verified in the final phase. From the experimental results it was found that from the top hundred genes ranked based on T-scores, single gene selected classified 77% of DLBCL subtypes, 67% of FLL subtypes for all hundred genes in the test dataset, two gene combinations was found to have 77% of DLBCL subtypes, 50% of FL subtypes for all hundred genes in the test dataset and three gene combinations was found to have 63% of DLBCL subtypes, 50% of FL subtypes for all hundred genes in the test dataset. The evolutionary approaches such as optimization methods can be used to generate best gene combinations to achieve higher level classification accuracy. In this work we tested a sample dataset called test dataset which contained hundred genes, it is considered as the limitation and we move forward to classify the entire dataset with fuzzy logic in future.

6. REFERENCES

[1] Guoyin Wang, Jun Hu, Qinghua Zhang , Xianquan Liu and Jiaqing Zhou "Granular computing based data mining in the views of rough set and fuzzy set", IEEE International Conference on Granular Computing, 978-1-4244-2513-6, 2008.

[2] Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[3] Qinghua Huang, Dacheng Tao, "Exploiting Local Coherent Patterns for Unsupervised Feature Ranking", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, , 1083-4419, 2011.

[4] Lipo Wang and Feng Chu, "Extracting Very Simple Diagnostic Rules from Microarray Data",32nd Annual International Conference of the IEEE EMBS, August 31 - September 4, 2010.

[5] Qingzhong Liu, and, Andrew H. Sung, "Recursive Feature Addition for Gene Selection", International Joint Conference on Neural Networks, Canada, July 16-21, 2006.

[6] Yan-Fei Wang, Zu-Guo Yu1 and Vo Anh, "Type-2 fuzzy Approach for Disease-Associated Gene Identification on Microarrays", International Conference on Bioscience, Biochemistry and Bioinformatics, IPCBEE vol.5, 2011.

[7] Pablo Martín-Munoz and Francisco J. Moreno-Velo, "FuzzyCN2: An Algorithm for Extracting Fuzzy Classification Rule', IEEE World Congress on Computational Intelligence, July 18-23, 2010.

[8] Nilesh N. Karnik, Jerry M. Mendel and Qilian Liang, " Type-2 Fuzzy Logic Systems", IEEE transactions on fuzzy systems, vol. 7, no. 6, December 1999.

[9] Jahangheer Shaik and Mohammed Yeasin, "Fuzzy-Adaptive-Subspace-Iteration-Based Two-Way Clustering of Microarray Data", IEEE/ACM transactions on computational biology and bioinformatics, vol. 6, no. 2, april-june 2009.

[10] Keon Myung Lee, Kyung Soon Hwang, and Chan Hee Lee "Fuzzy Set-based Microarray Data Analysis Techniques for Interesting Block Identification", IEEE transactions, 2009.

[11] Zarita Zainuddin, Ong Pauline, "Improved Wavelet Neural Network for Early Diagnosis of Cancer Patients Using Microarray Gene Expression Data", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009.

[12] Wutao Chen,Huijuan Lu,Mingyi Wang , "Gene Expression Data Classification Using Artificial Neural Network Ensembles Based on Samples Filtering" , International Conference on Artificial Intelligence and Computational Intelligence, 2009.

[13] Alizadeh, “Distinct Types of Diffuse Large b-Cell Lymphoma Identified by Gene Expression Profiling,” Nature, vol. 403, pp. 503-511, 2000.

[14] Lipo Wang, Feng Chu, and Wei Xie, “Accurate Cancer Classification Using Expressions of Very Few Genes”, IEEE/ACM transactions on computational biology and bioinformatics, vol. 4, no. 1, january-march 2007.

[15] Patharawut Saengsir and Sageemas Na Wichian “Classification Models Based-on Incremental Learning Algorithm and Feature Selection on Gene Expression Data”, 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 978-1-4577-0425-3 ,pp 426 - 429 , 2011.

[16] Wutao Chen,Huijuan Lu and Mingyi Wang, “Gene Expression Data Classification Using Artificial Neural Network Ensembles Based on Samples Filtering”, International Conference on Artificial Intelligence and Computational Intelligence, 2009.

[17] Alok Sharma, Seiya Imoto and Satoru Miyano, “A top-r Feature Selection Algorithm for Microarray Gene Expression Data”, IEEE, 1545-5963,2011.

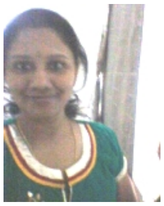
[18] Ming Chen and Zhengwei Yao, “Classification Techniques of Neural Networks Using Improved Genetic Algorithms”, Second International Conference on Genetic and Evolutionary Computing, 978-0-7695-3334-6, pp 115 - 119, 2008.

[19] Mingrui Zhang, Wei Zhang, Hugues Sicotte and Ping Yang, “A New Validity Measure for a Correlation-Based Fuzzy C-means Clustering Algorithm”, 31st Annual International Conference of the IEEE EMBS, USA, September 2-6, 2009.

[20] Yan-Fei Wang, Zu-Guo Yu and Vo Anh, “Fuzzy C-means method with empirical mode decomposition for clustering microarray data”, IEEE International Conference on Bioinformatics and Biomedicine, 2010.

Authors

Ms V Bhuvaneswari received her Bachelor’s Degree (B.Sc.) in Computer technology from Bharathiar University, India 1997 , Masters Degree (MCA) in Computer Applications from IGNOU, India and M.Phil in Computer Science in 2003 from Bharathiar University, India. She has qualified JRF, UGC-NET, for Lectureship in the year 2003. She is currently pursuing her doctoral research in School of Computer Science and Engineering at Bharathiar University in the area of Data mining. Her research interests include Bioinformatics, Soft computing and Databases. She is currently working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, India. She has for her credit publications in journals, International/ National Conferences.



Ms. K. Vanitha received her Bachelor’s Degree (B.Sc.) in Computer Science, Master Degree (MCA) in Computer Applications and MBA in Human Resources from Bharathiar University, India. She is pursuing M.Phil in Computer Science [Part-Time] in School of Computer Science and Engineering, Bharathiar University, India. Her research interests include Data Mining, Fuzzy Logic and Bioinformatics. She is currently working as Assistant Professor in the Department of Computer Applications, Hindusthan College of arts & science, Coimbatore, India. She has for her credit publications in International/ National Conferences. She is the member of IEEE.