MICROARRAY GENE EXPRESSION ANALYSIS USING TYPE 2 FUZZY LOGIC (MGA-FL)

V.Bhuvaneswari[1] and S.J.Brintha[2]

[1]Assistant Professor, Department of Computer Application, Bharathiar University, Coimbatore, India

bhuvanes_v@yahoo.com

[2]M.Phil Research Scholar, Department of Computer Application, Bharathiar University, Coimbatore, India

brinthasj@gmail.com

Abstract

Data mining is defined as the process of extracting or mining knowledge from vast and large database. Data mining is an interdisciplinary field that brings together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases. Bioinformatics is defined as the science of organizing and analyzing the biological data. Microarray technology helps biologists for monitoring expression of thousands of genes in a single experiment on a small chip. Microarray is also called as DNA chip, gene chip, or biochip is used to analyze the gene expression profiles. Fuzzy Logic is defined as a multivalued logic that provides the intermediate values to be defined between conventional evaluations like true or false, yes or no, high or low, etc.In this paper, a type 2 fuzzy logic approach is used in microarray gene expression data to convert the numerical values into fuzzy terms. After fuzzification, the fuzzy association patterns are discovered. A framework is proposed to cluster microarray gene data based on fuzzy association patterns. Then the proposed type 2 fuzzy approach is compared with traditional clustering algorithms.

Keywords

Bioinformatics, Microarray, Fuzzy Logic, Association, Clustering.

1. INTRODUCTION

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [2]. Data Mining is an essential step in Knowledge Discovery in Database (KDD) process and it is defined as a process of information extraction. Data Mining is concerned with the algorithmic means by which patterns or structures are enumerated from the data under computational efficiency limitations.

Data mining can be performed on data represented in quantitative, textual, or Multimedia forms. The widespread use of databases and the explosive growth in their sizes, organizations are faced with the problem of information overload [3]. The effective utilization of these massive volumes

DOI : 10.5121/ijcsea.2012.2205

of data is becoming a major problem for all enterprises. Data mining supports automatic exploration of data and evaluates the patterns.

Data Mining is a set of processes related to analyzing and discovering useful, actionable knowledge buried deep beneath large volumes of data sets. This knowledge discovery involves finding patterns or behaviors within the data that lead to some profitable business action. The main aim of data mining is to discover hidden fact in databases. The two primary goals of data mining are prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Description focuses on finding human-interpretable patterns that describes the data.

Bioinformatics is defined as the application of computer technology to the management of biological information. Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules. The main goal of bioinformatics is to enhance the understanding of biological processes [9]. Bioinformatics derives knowledge from computer analysis of biological data. It is the interdisciplinary field of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics also deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, data mining, image processing, etc.

Bioinformatics and data mining together provide exciting and challenging researches in computer science. Advances such as genome-sequencing initiatives, microarrays, proteomics, and functional and structural genomics have pushed the frontiers of human knowledge. Data mining provides the necessary tools for better understanding of gene expression, drug design, and other emerging problems in genomics and proteomics. Knowledge Discovery in Database (KDD) techniques plays an important role in the analysis and discovery of sequence, structure and functional patterns or models from large sequence databases. Data mining approaches are ideally suited for bioinformatics, since it contains rich data.

Data are collected from genome analysis, protein analysis, microarray data and probes of gene function by genetic methods. Maintaining this type of enormous data is a tedious process for most laboratories. Mostly the data are kept on spread sheets for easy usage. Data are arranged in appropriate database format. A database format is important in applied areas such as medical, agricultural and pharmaceutical research.

Gene expression is the conversion of the genetic information that is present in a DNA sequence into a unit of biological function in a living cell. It involves two processes like transcription and translation. The process of converting a gene into RNA is called as transcription. The transcription is followed by the process of translation of the RNA into protein. There are many techniques used for determining and quantifying gene expression, and many of these techniques have substantial statistical components to them. The process of measuring gene expression via cDNA is called gene expression analysis or gene expression profiling. Gene Expression analysis is used to determine whether the particular gene is expressed or not. In gene expression analysis, the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression.

Microarray is also called as DNA chip that is used for analyzing gene expression data. DNA microarrays are becoming a fundamental tool in genomic research. Microarray methods were initially developed to study differential gene expression using complex populations of RNA [10]. A microarray is a huge collection of spots that contain massive amounts of compressed data like DNA, protein, or tissue arranged on array for easy simultaneous analysis. Each spot of a microarray contains a unique DNA sequence, Protein or Tissue. Microarray analysis is a technique that is used to determine whether genes are expressed or not and it also supports the discovery of drug-sensitive genes and the chemical substructures associated with specific genetic responses.

Microarray is a method for profiling gene and protein expression in cells and tissues. As a new technology, microarrays have great potential to provide genome-wide patterns of gene expression, to make accurate medical diagnosis, and to explore genetic causes underlying diseases. Scientists used the DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10^{-12} moles) of a specific DNA sequence. These picomoles are known as probes (or reporters). These probes consist of short section of a gene or other DNA elements that are used to hybridize a cDNA or cRNA sample called target. Several microarray gene expression datasets are publicly available on the Internet. These datasets include a large number of gene expression values and there is a need to have an accurate method to extract knowledge and useful information from these microarray gene expression datasets.

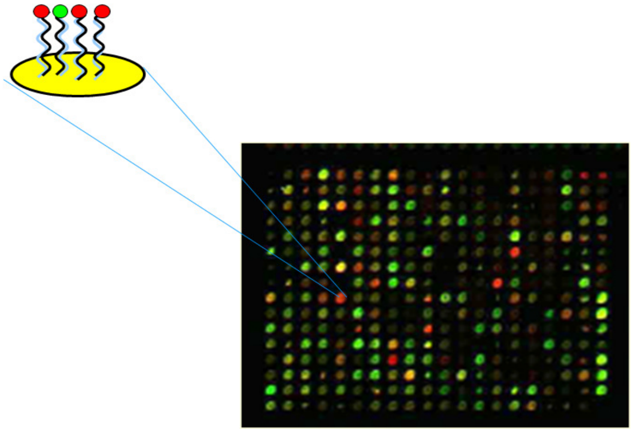


Figure 1. Microarray Gene Expression Analysis

Fuzzy logic is a multivalued logic that differs from "crisp logic", where binary sets have two-valued logic. Fuzzy logic variables have truth value that ranges in degree between 0 and 1. Fuzzy logic is a superset of conventional Boolean logic that has been extended to handle the concept of partial truth.

A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. Fuzzy Logic consists of Type 1 and Type 2 fuzzy. A type-2 fuzzy set contains the grades of membership that are themselves fuzzy. Type-1 Fuzzy Logic is unable to handle rule uncertainties. Type-2 Fuzzy Logic can handle rule uncertainties effectively and efficiently [8]. A Type-2 membership grade can be any subset in the primary membership. For each primary membership there exists a secondary membership that defines the possibilities for the primary membership.

Type 1 fuzzy contains the constant values. A Type-2 Fuzzy Logic is an extension of Type 1 Fuzzy Logic in which the fuzzy sets comes from Existing Type 1 Fuzzy. Type 2 Fuzzy sets are again characterized by IF–THEN rules [14]. Type-2 Fuzzy is computationally intensive because type reduction is very intensive. Type-2 fuzzy is used for modeling uncertainty and imprecision in a better way. The type-2 fuzzy sets are called as “fuzzy fuzzy” sets where the fuzzy degree of membership is fuzzy itself that results from Type 1 Fuzzy [7]. The gene expression levels are qualitatively classified into low, medium and high states to a varying degree based on a set of membership functions.

Gene subset selection is important for both classifying and analyzing the microarray gene expression data. It is a very difficult process because the gene expression data contain redundant information and noises. To solve this problem, fuzzy logic based pre-processing approaches are used. Fuzzy inference rules are used to transform the gene expression levels of a given dataset into fuzzy values. Then the associations (similarity relations) to these fuzzy values are applied to define fuzzy association patterns. Each fuzzy equivalence group and association patterns contain strongly similar genes.

The presence of duplicate information in microarray makes the classification and clustering task more difficult. A fuzzy logic based approach is used for eliminating the redundancy of information in microarray data. This technique is easy to understand and can be used for a biological interpretation. The aim of the paper is to propose a data mining technique to apply fuzzy logic and for clustering microarray data to group genes with similar functionalities to analyze the gene expression profiles.

The paper is organized as follows. Section II provides the literature study of the association rule mining, Type 2 fuzzy logic for microarray. In Section III the MGA-FL methodology for gene expression analysis is explained. Section IV explores the implemented results for the yeast dataset. The final section draws the conclusion of the paper, the major strength of this work and direction for future research.

2. REVIEW OF LITERATURE

Micro array technology is one of the latest advanced technologies in molecular biology. It enables monitoring the expression thousands of genes (virtually the entire genome) simultaneously. A wide range of different methods have been proposed for the analysis of gene expression data including hierarchical clustering, self-organizing maps, and k-means approaches.

In [23], Zuoliang Chen et al., (2008) has proposed an associative classification approach namely Classification with Fuzzy Association Rules (CFAR). Here the Fuzzy logic technique is used for partitioning the domains. The experimental results inferred that CFAR generates better understandability than the traditional approaches. In this paper Nenad Jukic et al., [13] (2011) has proposed qualified association rules mining. This qualified association rule mining is an extension of the association rules data mining method, which discovers previously unknown correlations under some circumstances. This method aims at improving the action results. Various experiments are carried out to illustrate how the qualified rules increase the effectiveness when comparing with standard association rules.

In [22], Yuchun Tang et al., (2008) proposed a new Fuzzy Granular Support Vector Machine Recursive Feature Elimination algorithm (FGSVM-RFE). This algorithm eliminates the irrelevant, redundant, or noisy genes in different granules at different stages and selects highly informative genes with potentially different biological functions. In [21], Yan-Fei Wang et al., (2011) has proposed a Type-2 fuzzy membership test for disease-associated gene identification on microarrays to improve traditional fuzzy methods. In [16], Peter J. Woolf et al., (2000) developed a novel algorithm for analyzing the gene expression profiles. This algorithm used the fuzzy logic to transform gene expression values into qualitative descriptors.

In [1], AnirbanMukhopadhyay et al., (2009) has proposed a novel multiobjective variable string fuzzy clustering scheme for clustering microarray gene expression data. The proposed technique evolves the total number of clusters along with the clustering result. The proposed method is applied on three publicly available real life gene expression datasets.

In [4], the author Dongguang Li (2008) presented information about how to discover useful information based on the DNA microarray expression data collected from mouse experiments for the leukemia research. Many methodologies involving fuzzy sets, neural networks, genetic algorithms, rough sets, wavelets, and their hybridizations, are suggested to provide approximate solutions. In [18], Qilian Liang et al., (2000) has proposed a method for designing interval Type-2 fuzzy logic systems. The concept of upper and lower membership functions is introduced and defined in this paper.

In [19], Raed I. Hamed et al., (2010) developed a fuzzy reasoning model based on the Fuzzy Petri Net. In [17], Pradipta Maji (2011) has proposed a clustering algorithm called fuzzy rough supervised attribute clustering (FRSAC). The effectiveness of the FRSAC algorithm is compared with existing supervised and unsupervised clustering algorithms like hierarchical clustering, the *k*-means algorithm, SOM and PCA.

In [15], Pablo Martín-Muñoz et al., (2010) presented a new algorithm FuzzyCN2 for extracting conjunctive fuzzy classification rules. The algorithm introduces the use of linguistic hedges and produces more compact rules. In the article [12], Muzeyyen Bulut Ozek et al., (2007) has presented a new Type-2 Fuzzy Logic Toolbox in MATLAB programming language. The primary goal is to help the user to understand and implement Type-2 fuzzy logic systems easily.

In [14], Nilesh N. Karnik et al., (1999) has introduced a Type-2 fuzzy logic system that can handle rule uncertainties. Type-2 Fuzzy logic is applied to time-varying channel equalization and it is proved that the Type 2 fuzzy logic system provides better performance than the existing Type-1 Fuzzy and nearest neighbor classifier.

In [11], Mohammad Mehdi Pourhashem et al., (2010) has proposed a new method based on fuzzy clustering and genes semantic similarity to estimate missing values in microarray data. In the proposed method, microarray gene data are clustered based on genes semantic similarity and their gene expression values. In [5], Edmundo Bonilla Huerta et al., (2008) introduced a fuzzy logic based pre-processing approach composed of two main steps. First, the fuzzy inference rules are used to transform the gene expression levels of a given dataset into fuzzy values. Then a similarity relation is applied to these fuzzy values to define fuzzy equivalence groups.

In [20], Vincent S. Tseng et al., (2006) has proposed two fuzzy data mining approaches for microarray analysis called Fuzzy Associative Gene Expression (FAGE) and Ripple Effective Gene Expression Rule (REGER) algorithms. Both techniques transform the microarray gene data into fuzzy items. Then the fuzzy operators and specially designed data structures are used to discover the relationships among genes.

3. PROBLEM FORMULATION AND METHODOLOGY

The main objective of this research work is to propose a framework to classify and analyze Microarray Gene data by using data mining and fuzzy logic. The specific objective of the work is to cluster the microarray gene data based on fuzzy association patterns and to compare the proposed work with existing traditional algorithms.

Genome sequencing projects have provided static pictures of the genomes of many organisms. Fuzzy logic incorporates a simple rule based approach for solving problems rather than attempting to model a system mathematically. Linguistic variables are the input or output variables of the system whose values are words or sentences from a natural language, instead of numerical values. The applied fuzzy logic consists of a set of fuzzy if-then rules that enable accurate nonlinear classification of input patterns. Fuzzy logic transforms quantitative expression values into linguistic terms that are able to uncover hidden fuzzy sequential associations between genes.

The proposed framework given in Figure 2 is used for analyzing associations of microarray gene data using fuzzy logic and clustering approach. The proposed framework consists of three phases. In the first phase the preprocessing and fuzzification of microarray data is done. In the second phase the fuzzy association pattern of genes are discovered and the microarray gene data is grouped according to association patterns. Two types of clustering are done within the third phase and the results are compared with the proposed work for finding accuracy.

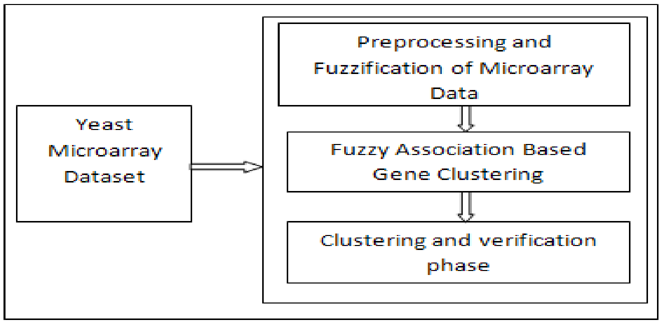

Yeast Microarray Dataset
Preprocessing and Fuzzification of Microarray Data
Fuzzy Association Based Gene Clustering
Clustering and verification phase

Figure 2. Framework (MGA-FL)

3.1 Preprocessing and Fuzzification

The Yeast Saccharomyces Cerevisiae Dataset is downloaded from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) website contains about 6240 Genes with their corresponding Yeast values (0.04481, 0.725). Microarray gene data contains noisy and inconsistent data. Preprocessing is the process of removal of inconsistent data and to extract necessary information. Table 1 show the microarray gene data downloaded consists of empty spots. The preprocessing phase is used for extracting needed information from the data that consists of two main processes.

Table 1. Microarray gene data with empty spots

Gene Id	Value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
2808052		2.250	100	74	88
28084699	-0.02025	0.785	110	152	248
28084700		0.875	100	72	86
28081779	-0.03629	1.008	110	181	302

In the preprocessing step the empty spots are replaced with null values using the isempty method. The empty spots are replaced by unique elements in dataset using unique method. Then the null values in the dataset are replaced by the maximum unique elements by using max method. Table 2 shows the preprocessed data. After replacing all the null values in the microarray gene data, the preprocessed gene expression values are given as input for the next process, the fuzzification.

Table 2. Preprocessed Microarray gene data

Gene Id	Value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
2808052	2	2.250	100	74	88
2808469	-0.02025	0.785	110	152	248
2808470	2	0.875	100	72	86

Gene expression data is quantitative and it contains numeric values. In Type 2 fuzzy ranges are given for the fuzzy values itself. The membership function for Type 2 fuzzy is given in Figure 3. Gene expression levels are qualitatively classified into up regulated (U), down regulated (D), low (L), medium (M) and high (H) states to a varying degree based on a set of membership functions in Type 2 Fuzzy logic.

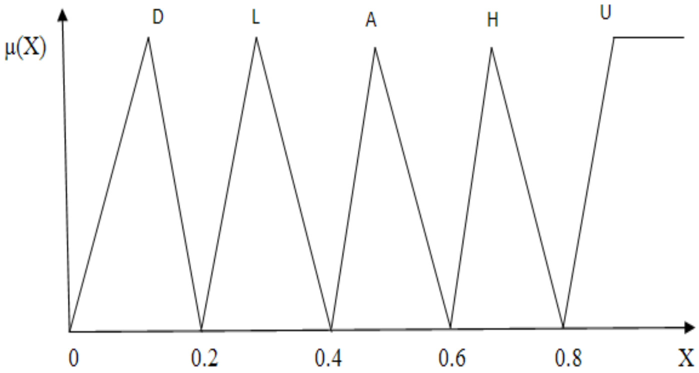


Figure 3. Membership Function for Type 2 Fuzzy

The calculated fuzzy numeric values are shown in Table 3. Then the numeric values are converted into fuzzy linguistic terms are shown in Table 4.

Table 3. Gene data with Fuzzy values

Gene Id	Value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
2808052	-0.04578	0.99997	1	0.991927	0.9920
2808469	0.76562	0.99995	1	0.97578	0.97504
2808470	0.87553	0.99996	1	0.966735	0.96640
2808177	0.93850	0.99997	1	0.969313	0.969085

Table 4. Gene data with Fuzzy Terms

Gene Id	Value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
2808052	'D'	'U'	'U'	'U'	'U'
28084699	'H'	'U'	'U'	'U'	'U'
28084700	'U'	'U'	'U'	'U'	'U'
28081779	'U'	'U'	'U'	'U'	'U'

3.2 Fuzzy Association Based Gene Clustering

The numeric quantitative values of gene data are converted into fuzzy terms using fuzzy logic. After fuzzification, the fuzzy values are given as input for the next phase, the finding of gene association. In the second phase, to find the fuzzy association pattern lpq⟷ljk the association between the linguistic terms lpq and ljk are discovered.The microarray gene data with three states contains nine possible associations according to the gene expression states. The gene data contains fuzzy association patterns like

$$L \rightarrow L, L \rightarrow H, L \rightarrow A, A \rightarrow L, A \rightarrow A, A \rightarrow H, H \rightarrow L, H \rightarrow A, H \rightarrow H$$

The gene positions corresponding to a particular association are grouped as shown in Table 5. For example 51, 69, 3, 5, indicate the gene position numbers in which the associations are present.

Table 5. Association patterns and its Gene numbers

Rule No	Associations	Gene numbers
1	L→ A	51,69,165,177,281,317,654,655,703
2	L→ H	1,2,3,4,5,6,7,8,9,10,11,12,13
3	A→ H	2308,4272,4273,4274,4630,5850
4	A →L	17,18,19,20,21,22,23,24,25,26,27,28
5	H →L	5755,5756,5757,5758,5759,5760
6	H→A	284,285,286,287,288,289,290,291
7	L→L	0
8	A →A	0
9	H→H	0

The total numbers of genes for particular association are also given Table 6. In the below Table the count 6512 indicates the total number of genes for the association L→A.L→ A represents the association Low →Average.

Table 6. Total number of genes for Association patterns

Association Patterns	Total no of genes
L→A	65
L→H	6152
H→L	6

The uncertainty associated with the fuzzy association patterns can be modeled by using confidence measure. The confidence measure is defined as the probability of the pattern Pr(lpq↔ljk). The weight of evidence measure W (lpq ↔ ljk) is calculated to handle the uncertainties.The weights calculated for five clusters are shown in Table 7.

Table 7. Calculated Weights for Five Clusters

Associations	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
L→A	6.4013	2.0717	6.2697	2.0719	6.5422
L→H	7.0953	2.0620	6.5270	2.0625	6.5425
A→H	2.4775	2.3393	2.8394	2.3377	6.0861
A→L	2.7859	2.3297	3.0213	2.3286	6.0864
H→L	3.0213	2.3201	3.1758	2.3191	6.0867
H→A	2.4788	2.6912	2.5847	2.6887	5.7730
L→L	2.6477	2.6818	2.7068	2.6793	5.7727
A→A	2.0371	4.5420	2.0625	4.5660	5.3364
H→H	2.0794	6.3173	2.0794	6.5525	5.2742

After the calculation of weight, a set of gene expression data collected from a set of N' genes from previously unseen gene expression data are collected. To predict the accuracy the fuzzy association patterns previously discovered in each class can be searched to see which patterns can match with the new expression profile. The weight of evidence W' (lpq↔ljk) supporting the assignment of new class is defined as:

$$W'(lpq \leftrightarrow ljk) = W(lpq \leftrightarrow ljk) * \mu Fpq$$

Finally the degree of membership that the new gene expression data belongs to each class is calculated. The accuracy between the genes can be predicted by calculating how the fuzzy association patterns in new gene expression data can match with the patterns in previous dataset. The occurrences of the particular association are grouped and the data belonging to that association are grouped and stored for each cluster. The calculated fuzzy values are also grouped for clusters. The rules generated for five clusters are displayed separately in Table 8. The table contains the cluster number and rule number. As given in the Table 8 the following association patterns are found for the clusters identified. The numbers like 2, 3, 1, 6, and 5 indicates the association rule number. Some clusters contain few association rules. It indicates that the other rules are not present within the cluster.

Table 8.Fuzzy Association Rules for five clusters

Cluster no	Association Rule patterns
Cluster 1	2,3,4,5,6,7,8,9
Cluster 2	1,7,9
Cluster 3	1,2,3,4,8
Cluster 4	3,9
Cluster 5	1,6,7,8

The occurrences of the association are grouped for each cluster. Each association rule is represented by some constant values. The original data and the fuzzy values belonging to the particular association are grouped in clusters.

3.3 Clustering and Verification Phase

The fuzzy association patterns are discovered and the accuracy is calculated in the second phase. After the pattern discovery the clustering is done in third phase. Two types of clustering algorithms are used. One is k-means and other is hierarchical clustering. After clustering the fuzzy logic technique is applied for the cluster results. A comparison is made with the results from kmeans without fuzzy logic and kmeans with fuzzy logic. The comparison is done for hierarchical clustering similar to kmeans. The comparison is done for both Type 1 and Type 2 fuzzy.

After comparison, the results are plotted and verified for both types of clustering. After the verification of results it is observed that the accuracy is increased by using fuzzy logic technique. The accuracy values for kmeans clustering without applying fuzzy logic and with applying Type 1 fuzzy logic are shown in Table 9. Table 10 shows the accuracy values for hierarchical clustering.

Table 9. Accuracy Values for Kmeans and Type 1 Fuzzy

Kmeans without fuzzy	Kmeans+Type 1 Fuzzy
50.09	60.19
58.94	65.25
14.55	60.31
50.57	80.51
44.23	57.08

Table 10. Accuracy values for Hierarchical and Type 1 Fuzzy

Hierarchical without fuzzy	Hierarchical + Type 1 fuzzy
5.30	5.60
3.25	6.79
8.01	11.19
9.47	10.93
11.33	13.07

Similar to Type 1 fuzzy the comparison is made for Type 2 fuzzy using kmeans and hierarchical clustering algorithms. The accuracy values using Type 2 fuzzy and Kmeans clustering are shown in Table 11. Similar to kmeans the accuracy values are calculated using hierarchical clustering and it is shown in Table 12.

Table 11.Accuracy Values for Kmeans and Type 2 Fuzzy

Kmeans without fuzzy	Kmeans+Type 2 Fuzzy
47.92	81.18
52.99	72.72
55.79	79.41
51.008	70.59
53.56	68.07
56.47	82.25

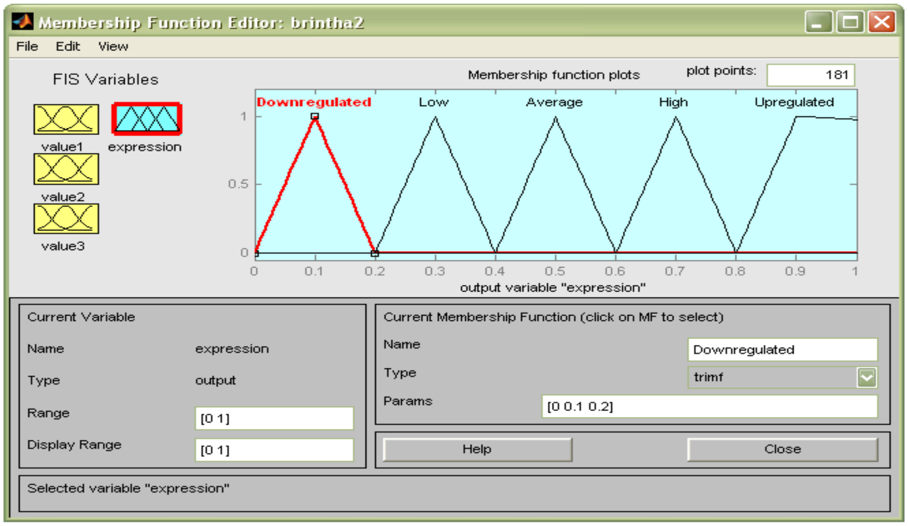
Table 12. Accuracy values for Hierarchical and Type 2 Fuzzy

Hierarchical without fuzzy	Hierarchical + Type 2 fuzzy
5.29	5.90
5.59	7.59
3.12	7.53
7.71	10.27
8.68	11.51
9.97	13.83

The results of clustering without using fuzzy technique are compared with the results of fuzzy technique. Then the proposed Type 2 approach (MGA-FL) is compared with the existing Type 1 approach. From the Tables 9,10,11,12 it is inferred that the proposed Type 2 approach (MGAFL) deals with more uncertainties than existing Type 1 fuzzy. The comparison between the accuracy values of Type 1 and Type 2 fuzzy shows that Type 2 fuzzy is more accurate than the Type 1 fuzzy. The K-Means algorithm and hierarchical clustering algorithms are implemented with fuzzy technique and without fuzzy technique.

4. IMPLEMENTATION RESULTS AND DISCUSSION

The experimental results and comparative study of the Fuzzy techniqyes and two algorithms are presented in this section. The membership functions are represented in fuzzy logic toolbox. The membership function for Type 2 fuzzy is shown in Figure 4. The membership function for Type 2 fuzzy is represented by trimf function type.



plot points: 181

Current Variable

Name expression

Type output

Range [0 1]

Display Range [0 1]

Current Membership Function (click on MF to select)

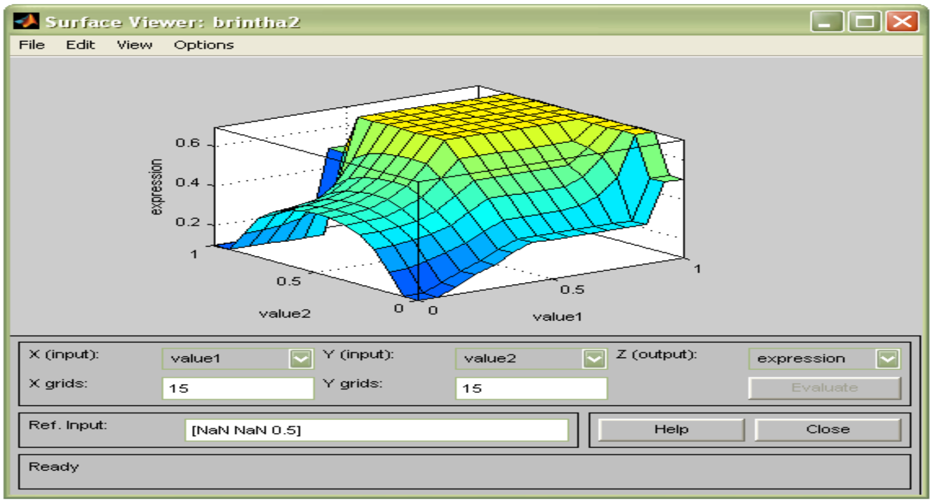
Name Downregulated

Type trimf

Params [0 0.1 0.2]

Selected variable "expression"

Figure 4. Membership Function for Type 2 fuzzy



X (input): value1
Y (input): value2
Z (output): expression
X grids: 15
Y grids: 15
Ref. Input: [NaN NaN 0.5]
Ready

Figure 5.Surface Viewer for Type 2 fuzzy

The Surface Viewer in the fuzzy logic toolbox has a special capability that is very helpful in cases with two (or more) inputs and one output. When opening the surface viewer a three dimensional curve that represents the mapping of genes to expression is shown. Figure 5 shows the surface viewer for the proposed work.

The cluster numbers and the gene numbers corresponding to the association patterns are shown in Table 13 .

Table 13. Cluster numbers and Gene numbers for Association Pattern

Association Patterns	Cluster Numbers	Gene Numbers
L → A	4,3,1	51,69,165,177,281,317,654
L → H	4,5,3	1,2,3,4,5,6,7,8,9,10,11,12,13
A → H	5,1,1	2308,4272,4273,4274,4630
A → L	5,1,4	17,18,19,20,21,22,23,24
H → L	3,4,1	5755,5756,5757,5758,5760
H → A	2,5,4	284,285,286,287,288,289,290

The number of clusters and the total number of genes associated with a particular cluster is shown in the form of bar chart in Figure 6. Based on the results from the bar chart it is analyzed that the maximum numbers of genes fall under the cluster 5 for fuzzy associations A → H, H → A, H → H, A → A.

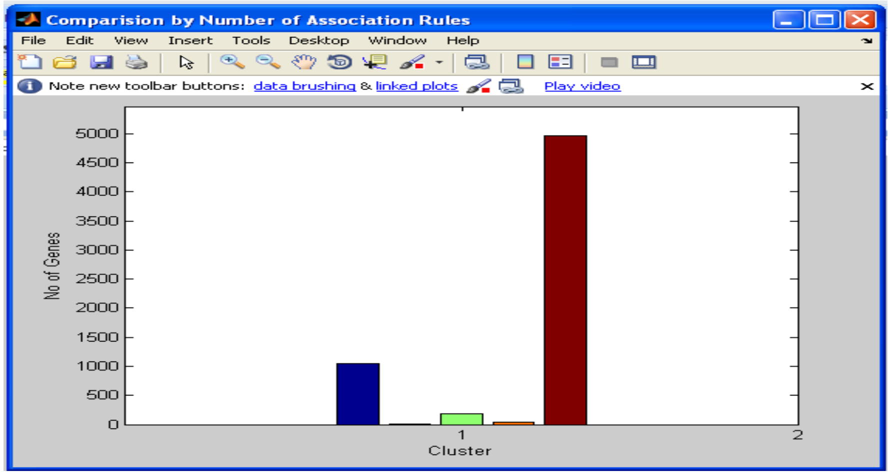


Figure 6. Clusters and its total number of Genes

The bar diagram for the different association patterns and total number of genes is represented in the Figure 7. Based on the analysis of results from the Fig 27, it is inferred that the association patterns like A → H, H → A, A → A and H → H occurred frequently in most of the clusters.

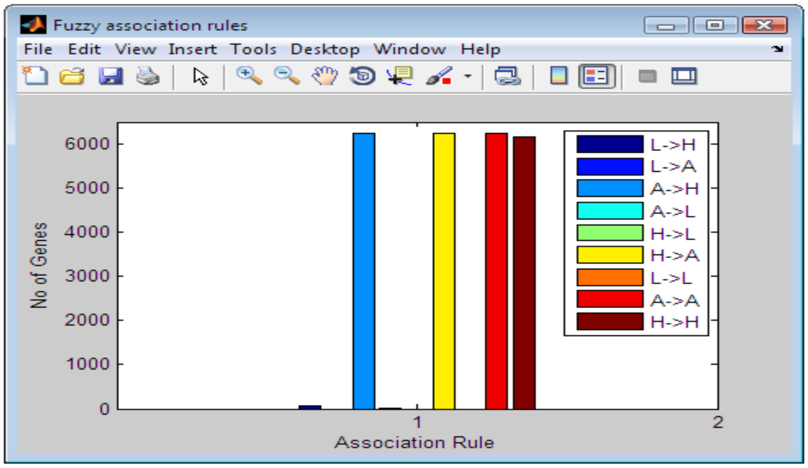


Figure 7. Gene Cluster for Association Patterns

The comparison of clustering algorithm accuracies for Type 1 and Type 2 fuzzy is done and it is shown in the Table 14.

Table 14. Comparison of accuracy values in Type 1 and Type 2 fuzzy

Type1		Type 2	
Kmeans	Hierrachical	Kmeans	Hierrachical
60.19	5.60	92.27	5.90
65.25	6.37	71.02	7.59
60.31	6.79	81.18	7.53
80.51	11.19	72.72	10.27
57.08	10.93	79.41	11.51
68.93	11.70	70.59	13.83
66.54	13.07	68.07	13.88
77.06	12.15	82.25	14.15

The result for Type 1 fuzzy and Type 2 fuzzy are analyzed and compared and discussed in the form of line chart. The accuracy values for kmeans and hierarchical clustering are plotted in line chart and it is shown in Figure 8 and Figure 9. The analysis of the clustering results infer that the accracy values are increased by using fuzzy logic technique when compared to the traditional clustering algorithms. Based on the implementation results the accuracy can be further increased by using the proposed Type 2 fuzzy approach (MGA-FL) compared to the existing Type 1 fuzzy approach.

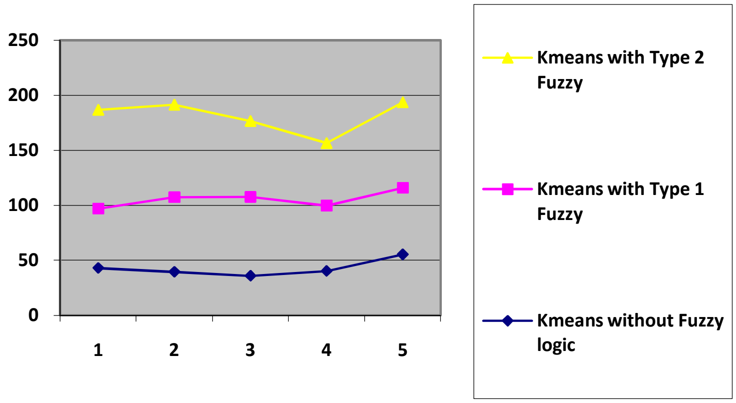


Figure 8.Accuracy results for Kmeans

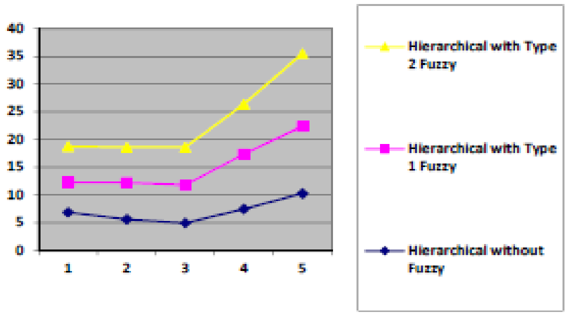


Figure 9.Accuracy results for Hierarchical clustering

The technique proposed in the research work using Type 2 fuzzy for finding association patterns to cluster microarray data is found to have highest accuracy than Type 1 fuzzy and other traditional clustering algorithms.

5. CONCLUSION AND FUTURE DIRECTIONS

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. We have proposed a methodology consisting of a framework (MGA-FL) for improving the proposed idea. The main idea of our proposed work is to apply fuzzy logic to microarray gene data to find fuzzy association patterns. The association uncertainties are handled by calculating weight and the accuracy of gene clusters are predicted. The fuzzy associations are proposed for both Type 1 and Type 2 fuzzy. In this study the microarray gene dataset is implemented and the clustering is done using fuzzy logic associations. The various snapshots obtained in the section IV prove that the proposed idea is found to have more accuracy than existing clustering algorithms K-Means and Hierarchical clustering. We have also compared the proposed work with existing Type 1 fuzzy and it is found that the accuracy results of the proposed Type 2 fuzzy logic is increased when compared to Type 1 fuzzy approach. The proposed approach clusters microarray gene data based on fuzzy association rules. The same can be implemented for classifying microarray gene data and compared with famous algorithms like K-NN and SVM. The proposed work can be combined with other machine learning approach like Neural Networks.

6. REFERENCES

[1] AnirbanMukhopadhyay, SanghamitraBandyopadhyayand UjjwalMaulik, “Analysis of Microarray Data using Multiobjective Variable String Length Genetic Fuzzy Clustering”, 978-1-4244-2959, 2009 IEEE.

[2] Arun K Pujari, “Data Mining Techniques”, Universities press (India) Limited 2001, ISBN-81-7371-3804.

[3] Daniel T Larose, “Introduction to Data Mining”, Discovering Knowledge in Data: An Introduction to Data Mining, ISBN 0-471-66657-2, 2005.

[4] Dongguang Li, “DNA Microarray Expression Analysis and Data Mining For Blood Cancer”, 2008 International Seminar on Future BioMedical Information Engineering.

[5] Edmundo Bonilla Huerta, Beatrice Duval, and Jin-Kao Hao, “Fuzzy Logic for Elimination of Redundant Information of Microarray Data”, Vol. 6 No. 2, 2008.

[6] Hellman M., “Fuzzy Logic Introduction”, Info. &Ctl., Vol. 12, 1968, pp. 94-102.

[7] Juan R Castro, Oscar Castillo, Luis G Martinez, “Interval Type-2 Fuzzy Logic Toolbox”, Engineering Letters, 15:1, EL_15_1_14, 2007.

[8] Luis Tari, ChittaBaral, Seungchan Kim, “Fuzzy C-Means Clustering with Prior Biological Knowledge”, Journal of Biomedical Informatics, 2008.

[9] Luscombe N.M, Greenbaum D, Gerstein M., “Bio informatics: A Proposed Definition and Overview of the Field”, IMIA Yearbook of Medical Informatics 2001: Digital Libraries and Medicine, pp. 83–99.

[10] Michael F Ochs and Andrew K Godwin, “Microarrays in Cancer: Research and Applications”, BioTechniques 34:S4-S15 March 2003.

[11] Mohammad Mehdi Pourhashem, ManouchehrKelarestaghi, Mir Mohsen Pedram, “Missing Value Estimation In Microarray Data Using Fuzzy Clustering And Semantic Similarity”, Global Journal of Computer Science And Technology, Page 18,Vol.10,October 2010.

[12] MuzeyyenBulutOzek, ZuhtuHakanAkpolat, “A Software Tool: Type-2 Fuzzy Logic Toolbox”, Wiley Periodicals Inc., 2007.

[13] NenadJukic, SvetlozarNestorov, Miguel Velasco, Jami Eddington, “Uncovering Actionable Knowledge in Corporate Data with Qualified Association Rules”, International Journal of Business Intelligence Research, 2(2), 1-21, April-June 2011.

[14] 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.

[15] Pablo Martin Munoz and Francisco J Moreno Velo, “Fuzzy CN2: An Algorithm for Extracting Fuzzy Classification Rule Lists”**,** WCCI 2010 IEEE World Congress on Computational Intelligence, 18-23, July 2010 - CCIB, Barcelona, Spain.

[16] Peter J Woolf and Yixin Wang, “A Fuzzy Logic Approach to Analyzing Gene Expression Data”, Physiol Genomics3: 9–15, 2000.

[17] PradiptaMaji, “Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data”, IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics, Vol. 41 and No. 1, February 2011.

[18] Qilian Liang and Jerry M. Mendel, “Interval Type-2 Fuzzy Logic Systems:Theory and Design”, IEEE Transactions On Fuzzy Systems, Vol. 8, No. 5, October 2000.

[19] Raed I Hamed1, Ahson S.I, Parveen R., “A New Approach for Modelling Gene Regulatory Networks Using Fuzzy Petri Nets”, Journal of Integrative Bioinformatics, 7(1):113, 2010.

[20] Vincent S Tseng, Yen-Hsu Chen, Chun-Hao Chen, and Shin J.W., “Mining Fuzzy Association Patterns in Gene Expression Databases”, International Journal of Fuzzy Systems, Vol. 8, No. 2, June 2006 .

[21] Yan-Fei Wang, Zu-Guo Yu, “A Type-2 Fuzzy Method for Identification of Disease-Related Genes on Microarrays”, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 1, No. 1, May 2011.

[22] YuchunTang,Yan-Qing Zhang,, Zhen Huang, Xiaohua Hu, and Yichuan Zhao, “Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification”, IEEE Transactions on Information Technology In Biomedicine, Vol. 12, No. 6, November 2008.

[23] Zuoliang Chen, Guoqing Chen, “Building an Associative Classifier Based on Fuzzy Association Rules”, International Journal of Computational Intelligence Systems, Vol.1, No. 3 (August, 2008), 262 – 273.

Authors

Ms V Bhuvaneswari received her Bachelor’s Degree (B.Sc.) in Computer technology from Bharathiar University, India 1997, Master’s 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 credit for her publications in journals, International/ National Conferences.

Ms. S.J.Brintha received her Bachelor’s degree in (B.Sc.) Microbiology from Bharathiar University, India 2007, Master’s Degree (MCA) in Computer Applications from Anna University, India 2010 and M.Phil in Computer Science in 2011 from Bharathiar University, India. She has attended National conferences. She was awarded gold medal for First Rank Holder in Master of Computer Applications. Her research interests include Data Mining, Bioinformatics, and networking.