Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 1, January 2014, pg.309 – 318

RESEARCH ARTICLE

Fuzzy Mining Approach for Gene Clustering and Gene Function Prediction

Akhil Kumar Das[1], Debasis Mandal[2], Mainak Adhikary[3], Amit Kumar Sen[4]

[1]Computer Science & Engineering, IMPS College of Engineering & Technology, Malda, West Bengal, India

[2]Computer Science & Engineering, IMPS College of Engineering & Technology, Malda, West Bengal, India

[3]Computer Science & Engineering, IMPS College of Engineering & Technology, Malda, West Bengal, India

[4]Computer Science & Engineering, IMPS College of Engineering & Technology, Malda, West Bengal, India

[1] dasakhi@gmail.com ; [2] debimps07@gmail.com ; [3] onlinemainak@yahoo.com, [4] amitsen41@gmail.com

Abstract— 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. After genome sequencing, DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. Biologists are vastly overwhelmed by the enormous amount of unique qualities of genome-wide data produced by the DNA Microarray experiment. Clustering is the process of grouping data objects into set of disjoint classes called clusters so that objects within a class are highly other classes. Generating high-quality gene clusters and identifying the underlying biological mechanism of the gene clusters are the important goals of clustering gene expression analysis. It is presently the far most used method for gene expression analysis which provides a Fuzzy mining strategy to extract meaningful information from expression profile. In this paper we have used a fuzzy mining approach for Gene Clustering with using different membership function and dividing the available gene expression data for each type of experimental value with four variables for better accuracy. This approach can effectively capture heterogeneity in expression data for pattern discovery. Based on these patterns, it can make accurate Gene Function Predictions and these predictions can be made in such a way that each gene can be allowed to belong to more than one functional class with different degrees of membership.

Keywords— Microarray; Fuzzy Mining; Gene Function Clustering; Gene function Prediction

I. INTRODUCTION

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. In clustering [2], the genes are un-labelled that is we don't have prior knowledge about any of gene's biological functions. Using the expectation those genes with common biological function will have expression profiles that exhibit similar pattern across different experimental conditions. Since cellular processes are naturally complex, heterogeneous expression patterns can always exists among genes in a given functional class behave homogeneously and also gene expression data are noisy and high dimensionality.

© 2014, IJCSMC All Rights Reserved

A. ***Data mining***

Data Mining is the non-trivial process of identifying valid, novel, potentially useful, and Ultimately understandable patterns in data. 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. Data mining is the technique of analyzing datasets (often large) in order to extract implicit, previously unknown and potentially useful information that might otherwise remain unknown.
Data mining techniques are useful in microarray analysis because:

- Data volumes are too large for traditional analysis methods.
- High dimensionality.
- Only a small portion of data is analyzed.
- Decision support process becomes more complex.

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. The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process.

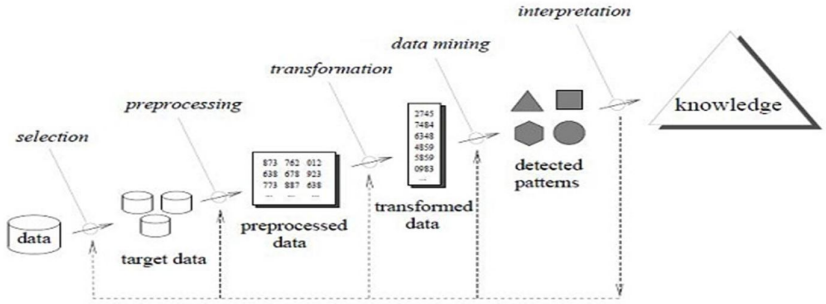


Fig. 1. Steps in KDD

B. ***Bioinformatics***

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. Bioinformatics can be defined as the application of computer technology not only to simulate biological processed and data but includes a study of the inherent genetic information, underlying molecular structure, resulting biochemical functions, and the exhibited phenotypic properties. The main goal of bioinformatics is to enhance the understanding of biological processes. It is the interdisciplinary field of developing and utilizing computer databases and algorithms to accelerate and enhance biological research.

C. ***Soft Computing in Bioinformatics***

As soft computing are considered to handle imprecision, uncertainty and near optimality in large and complex search spaces use of soft computing tools for solving bioinformatics problems have been gained the attention of researchers. Our literature survey of recent research papers shows role of soft computing in modelling various aspects of bioinformatics, it involves genomic sequence, protein structure, gene expression microarray, and gene regulatory networks. Most of the researches are Woven around the tasks of pattern recognition and data mining like clustering, classification, feature selection, and rule generation, while classification pertains to supervised or unsupervised learning, clustering corresponds to unsupervised self organization into homologous partitions. Feature selection techniques aim at reducing the number of irrelevant and redundant variables in the dataset. Rule generation enables efficient representation of mined knowledge in human understandable form.

D. Gene Expression Analysis

© 2014, IJCSMC All Rights Reserved

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. 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.

Microarrays provide an extremely powerful way to analyze gene expression. Using a microarray, it is possible to examine the expression levels of thousands of genes across different developmental stages, clinical conditions or time points. It helps in understanding gene functions, biological processes, gene networks, effects of medical treatments, etc. 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.

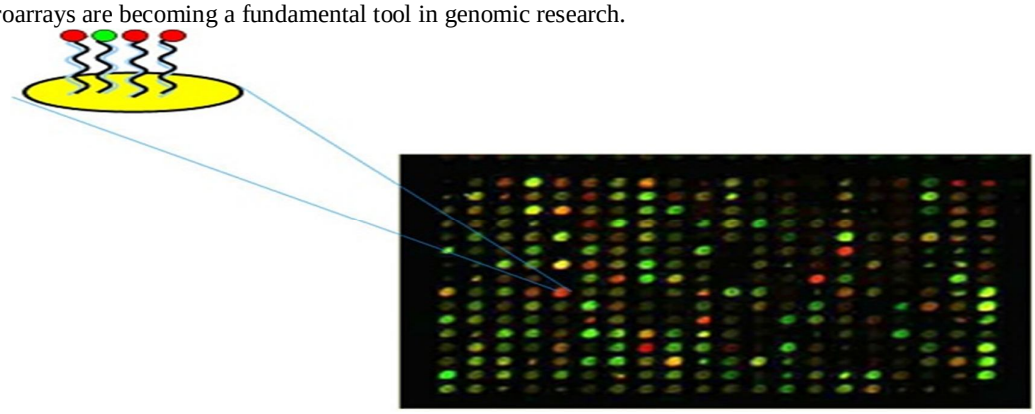


Fig. 2. Microarray Gene Expression Analysis

E. ***Fuzzy Logic***

Fuzzy logic is that it accepts the uncertainties that are inherited in the realistic inputs and it deals with these uncertainties in their affect is negligible and thus resulting in a precise outputs. Fuzzy Logic reduces the design steps and simplifies complexity that might arise since the first step is to understand and characterize the system behavior by using knowledge and experience. FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. It mimics human control logic. Fuzzy logic-based and fuzzy rule-based models can control and analyze processes and diagnose and make decisions in biomedical sciences. A fuzzy logic based approach is used for eliminating the redundancy of Information in microarray data. 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. This technique is easy to understand and can be used for a biological interpretation.

II. PROPOSED SCHEME

A. Objective:

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. 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.

B. Algorithm:

Step 1 : In this step the different gene expression data available from different experimental conditions are being fuzzified with the help of different membership function. Different states are defined. The whole data space is divided with the use of four variables. i.e. Numeric values are transformed into different fuzzy states.

© 2014, IJCSMC All Rights Reserved

Step 2 : Here fuzzy association patterns are discovered. gene positions corresponding to a particular association pattern are grouped together .
Step 3 : In this step weight is assigned to each fuzzy association pattern discovered in the previous step. These helps in characterization of a particular cluster.
Step 4 : Now the accuracy between the genes are predicted using calculated fuzzy association patterns in new gene expression data with the patterns in the previous data set. Same occurrences of particular association are grouped together and are stored for each cluster.

C. Flow of the proposed work

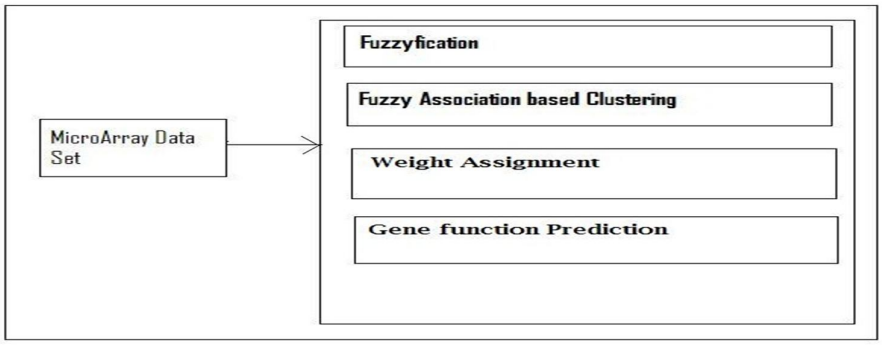


Fig. 3. Fuzzy Mining Technique Steps

D. *Explanation*:
It has four steps first Fuzzyfication, Second Association Based Clustering , then Weight Assignment And last Gene Function Prediction .

D.1 Fuzzyfication:
In this phase , we define different set for different conditions used in Fuzzyfication. We have N genes namely $G_1,G_2,…..,G_N$ in the set of gene expression data denoted by $G = \{ G_1,G_2,…..,G_N \}$.We have considered M experimental conditions denoted by $E_1,E_2,…….E_M$ in the set of experiments $E=\{ E_1,E_2,…….E_M \}$.Now the values $e_{i1},e_{i2},…….e_{ij}$, …..e_{iM} represent the expression value of the i^{th} gene in under the experimental condition $E_1,E_2,…..,E_j,….E_M$ Where e_{ij} □ domain(E) and also each gene is preclassified into one of the known functional classes. If the class information of the dataset is not available then a two phase clustering approach can be used ,in the first phase any clustering algorithm can be used to group genes into a set of initial clusters then we can apply our algorithm to the cluster discovered. Microarray gene data contains noisy and inconsistent data. Preprocessing is the process of removal of inconsistent data and to extract necessary information. In the preprocessing step the empty spots are replaced with null values using the is empty 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. To minimize the impact of noisy data ,we can also represent these quantitative gene expression data in linguistic variables and terms using the concept of fuzzy set.

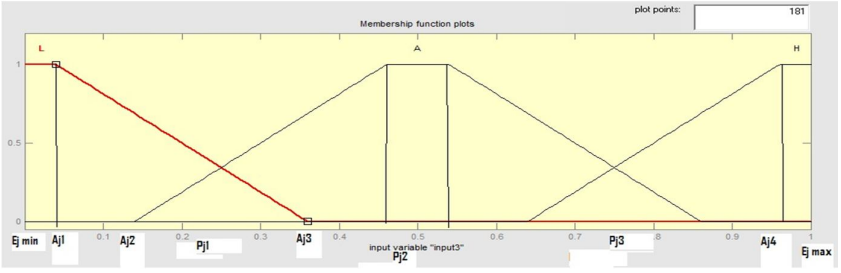


Fig. 4. Membership Function

In the figure Ej_{max} and Ej_{min} denote the maximum and minimum values of the quantitative attribute Ej, where j=1,2,…..M. We have assumed that the values of Ej are sorted in ascending order and let Pj1 be the value of Ej

© 2014, IJCSMC All Rights Reserved

that exceeds one-fourth of the measurements and is less than remaining three-fourth, similarly Pj2 be the value of Ej that exceeds half of the measurements and less than the remaining half of the measurements. Also Pj3 be the value that exceeds three-fourth of the measurements and is less than remaining one-fourth.
To calculate the value of Pj1,Pj2 and Pj3 ,the measurements are divided into a number of small class intervals nc of equal width ∂ (i.e., nc = 10 as suggested in [29]), and counted the corresponding class frequencies fi where i = 1, 2, . . . , n_c .

$Pj1 = low_i + ((R_1 - cf_{i-1})x\ \partial)/f_i$

$Pj2 = low_i + ((R_2 - cf_{i-1})x\ \partial)/f_i$

$Pj3 = low_i + ((R_3 - cf_{i-1})x\ \partial)/f_i$

Where low_i is the lower limit of the i^{th} class interval,
$R_k = (Nx\ k)/ N_F$ is the rank of the k^{th} partition value,N_F is the total number of fuzzy sets and N is the total no. of measurements and cf_{i-1} is the cumulative frequency of the preceding class .
We define
Aj1 = (Ej min + Pj1) \ 2
Aj2 = (Pj1+Pj2) \ 2
Aj3 = (Pj2 + Pj3) \ 2
Aj4 = (Pj3 + Ejmax) \ 2
Now we give the degree of membership function of a gene expression value e_{ij} of Ej in Gi to each fuzzy set is as follows

$$\mu_L = \begin{cases} 1 & \text{if } eij \le Aj1 \\ \dfrac{Aj2-eij}{Aj2-Aj1} & \text{if } Aj1 < eij < Aj2 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_A = \begin{cases} 0 & \text{if } eij \le Aj1 \\ \dfrac{eij - Aj1}{Aj2 - Aj1} & \text{if } Aj1 < eij < Aj2 \\ 1 & \text{if } Aj2 \le eij \le Aj3 \\ \dfrac{Aj4 - eij}{Aj4 - Aj3} & \text{if } eij \le Aj4 \end{cases}$$

$$\mu_H = \begin{cases} 0 & \text{if } eij <= Aj2 \\ \dfrac{eij - Aj3}{Aj4 - Aj3} & \text{if } Aj3 < eij < Aj4 \\ 1 & \text{otherwise} \end{cases}$$

D.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 this phase ,we find the fuzzy association pattern lpq $\longrightarrow$ ljk between the linguistic terms ljk and lpq . lpq is the value of the linguistic variable Lp in G.To find the fuzzy association pattern lpq $\longleftrightarrow$ ljk the association between ljk and lpq are discovered and microarray gene data is grouped according to association patterns. In the microarray gene data we have considered three states namely L, A and H .These three states contains 3! = 9 possible associations according to the gene expression states. The gene data contains fuzzy association pattern like
L $\longrightarrow$ L , L $\longrightarrow$ H , L $\longrightarrow$ A , A $\longrightarrow$ L , A $\longrightarrow$ A , A $\longrightarrow$ H , H $\longrightarrow$ L , H $\longrightarrow$ A , H $\longrightarrow$ H

© 2014, IJCSMC All Rights Reserved

D.3 Weight Assignment:

A weight is assigned to each discovered pattern because of Fuzzy association pattern discovered in the previous step is not deterministic. To remove the uncertainty we can use 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.

$W(lpq \longleftrightarrow ljk) = \log (Pr(ljk| lpq) \setminus Pr(ljk))$

D.4 Gene function Prediction:

In the previous step we have calculated the weights. Now a set of gene expression data collected from a set of N' genes from previously unseen gene expression data are collected.

Fuzzy association patterns that are previously discovered in each class is searched to see which patterns match the expression profile for the prediction accuracy. In the N' genes , the weight of evidence W'(lpq ⟷ ljk) supporting the assignment of new class is as [28]:

$W'(lpq \longleftrightarrow ljk) = W(lpq \longleftrightarrow ljk)^* \mu_{Fpq}$

Now all the fuzzy association patterns are gathered together for the assignment of new gene to a new class. Degree of membership for the new gene expression data is calculated. Degree of membership with same value are grouped together in a cluster and with different degree of membership are put into another cluster. To predict the accuracy between genes matching of fuzzy association patterns in the new gene expression data with the previous dataset. Occurrences of particular association pattern are grouped and the data belonging to that association are stored for each cluster and calculated fuzzy values are also grouped for clusters.

E. Implementation:

Here we have a small part from the expression data of the GSE 1039 series ,platform id GPL978.the table contains 23184 rows and 43 columns. We have only taken only six attributes from the table and five rows of it.

Description of the table attributes is given below

ID_REF	the unique identifier of the feature derived from the Array List.
VALUE	Ratio of Medians (635/532)
Log Ratio (635/532)	log (base 2) transform of the ratio of the medians.
Diameter	the diameter in um of the feature-indicator.
F635 Median	median feature pixel intensity at wavelength #1 (635 nm).
F635 Mean	mean feature pixel intensity at wavelength #1 (635 nm).

Gene data with values

Gene Id	value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
1	10	3	120	205	213
3	20	0.788	120	225	235
4	76	0.322	100	312	317
5	93	-0.604	110	282	293
10	98	0.251	110	409	419

Table 1 : Gene data with values

This table converts the original data to the corresponding Fuzzy Values.

Gene Id	value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
1	0	1	0.98	0.15	0.113
3	0.3758	0.328	0.98	0.224	0.219
4	0.7568	0.14	0.95	0.661	0.682
5	0.824	0	0.975	0.554	0.597
10	1	0.113	0.975	0.997	0.995

Table 2 : Gene data with Fuzzy values

© 2014, IJCSMC All Rights Reserved

Here different States are given according to the membership function.

Gene Id	value	Raw Log Ratio (635/532)	Diameter	F635 Median	F635 Mean
1	L	H	H	L	L
3	A	A	H	L	L
4	H	L	H	A	A
5	H	L	H	A	A
10	H	L	H	H	H

Table 3 : Gene data with Fuzzy Terms

The gene positions corresponding to a particular association are grouped

Rule No	Associations	Gene Number
1	L ⟶ L	0
2	L ⟶ H	1,2,3,4,5,6,7,13
3	L ⟶ A	69,165,,154,155,103
4	A ⟶ L	18,19,20
5	A ⟶ A	0
6	A ⟶ H	108,172,150
7	H ⟶ L	156,160
8	H ⟶ A	85,86,87,88
9	H ⟶ H	0

Table 4 : Association patterns and its Gene numbers

As we have experimented and compared the table with less number of gene expression data .therefore actual no of genes in any association pattern is not given here. The experiment has been done with total 200 genes (approx).The following table indicate that probable number of genes in the following association pattern.

Association Patterns	Total no of Genes
L ⟶ A	26
H ⟶ L	11
L ⟶ H	128
A ⟶ H	7
H ⟶ H	0

Table 5 : Total no of Genes for Association Patterns

The weights calculated are shown in the following table

Associations	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
L ⟶ A	4.9013	1.57174	4.7585	1.582	5.0324
L ⟶ H	5.5873	1.601	4.856	1.482	5.041
L ⟶ L	1.131	1.201	1.198	1.166	4.538
A ⟶ L	1.285	1.08	1.548	1.601	4.652
A ⟶ A	1.0371	3.542	1.062	3.566	4.336
A ⟶ H	1.652	1.5624	1.682	1.553	5.086
H ⟶ L	2.021	2.320	2.175	1.319	5.086
H ⟶ A	1.278	1.491	1.384	1.487	4.55
H ⟶ H	1.08	5.297	1.079	5.452	4.198

Table 6 : Weight of the following clusters

Cluster No	Association Rule Number
1	1,2,3,4 ,5
2	5,9,
3	6,1
4	2,7
5	8,4,7,5

Table 7 : Fuzzy Association Rules for five cluster

© 2014, IJCSMC All Rights Reserved

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.

F. Output From The Simulator:

For implementation in Matlab 6.1 , We are giving the following screenshots that was used in Fuzzy Toolbox or Fuzzy Inference Engine .

First the picture of Fuzzy Inference system is being given

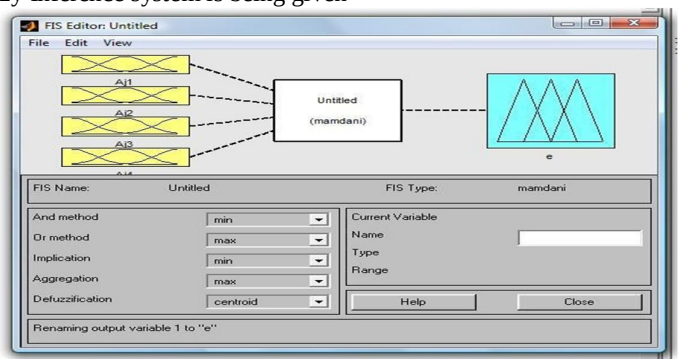


Fig. 5. Fuzzy Inference system

Membership Function of the variable is being given bellow

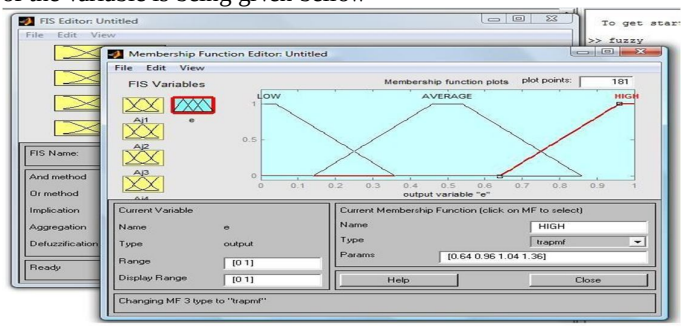


Fig. 6. Membership Function the variable

The Rule Viewer of Four Variables is given below

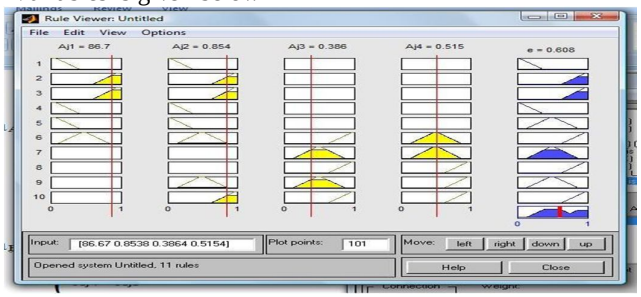


Fig. 7. Rule Viewer of Four Variable

Fig. 8. Rule Viewer of Four Variable

© 2014, IJCSMC All Rights Reserved

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. It shows the surface viewer for the proposed work.

Surface Viewer : Aj2 Aj3 Variables have been considered

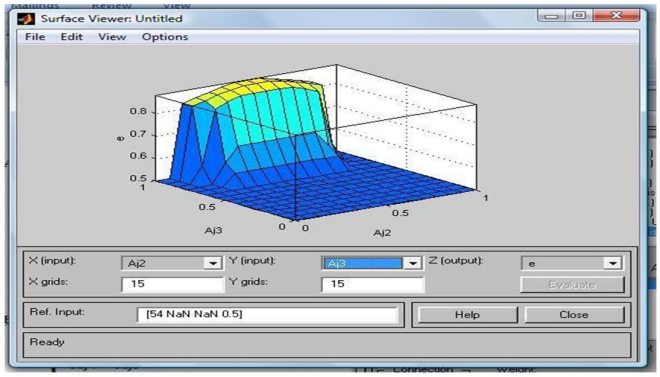


Fig. 9. Surface Viewer : Aj2 Aj3 Variables

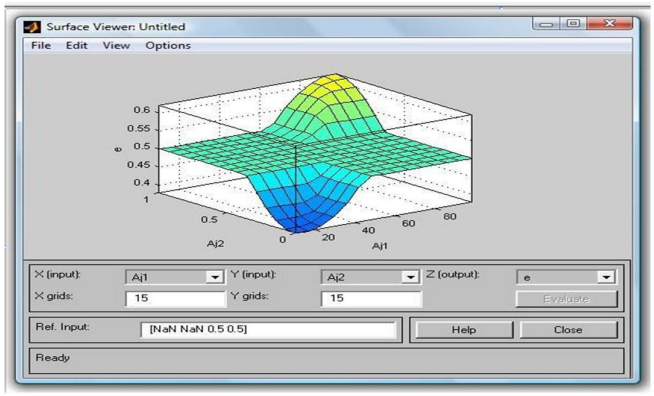
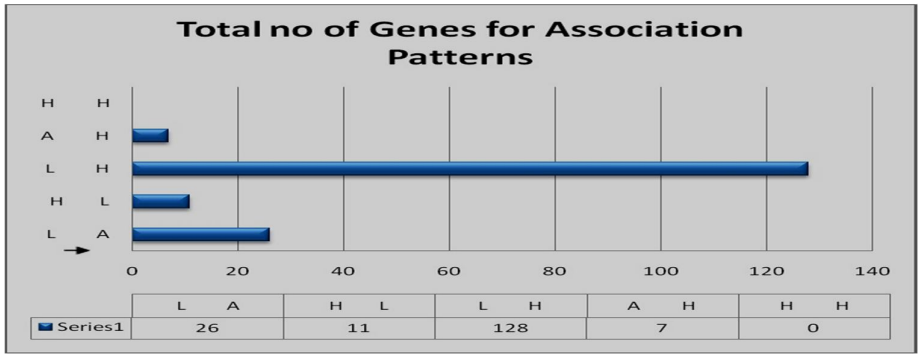


Fig. 10. Surface Viewer : Aj1 Aj2 Variables

Chart for Total no of Genes for Association Patterns



Comparison of Different classification Algorithms Genome wide Functional Prediction, Prediction Accuracy

Algorithm	Fuzzy mining	K-NN	SVM
Average	59	49	42

This prediction accuracy can be increased if the experiment is being done on very large data set. As the number of attributes taken in consideration increase in number then the prediction accuracy increases slowly . If this

© 2014, IJCSMC All Rights Reserved

fuzzy mining approach is combined with traditional clustering algorithms then this approach gives better accuracy results with respect to only using any fuzzy mining approach.

III. CONCLUSION

In this thesis paper We have developed a fuzzy mining technique for gene clustering and gene function prediction in gene expression data . Gene expression data are usually lowly expressed, averagely expressed and highly expressed across conditions. For gene expression data ,capturing this expressed information is more important. The main objective of this work is to apply fuzzy logic to microarray gene data for fuzzyfying the expression data and to find the association pattern for gene clustering with the help of weight assignment .It transforms continuous gene expression data into Linguistic values that can express that data in qualitative terms as Low, average and High.

In the Fuzzyfication We have used Trapezoidal membership function in gene expression data and the total data set has been divided with the help of four variables for better accuracy of result. Association uncertainties are handled by calculating weight and the accuracy of gene clusters are predicted.

IV. REFERRENCE

[2]M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome- wide expression patterns," in Proceedings of National Academy of Sciences, vol. 95, 1998, pp. 14 863–14 868.

[3]U. Alon, N. Barkai, Notterman, D. A., K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array," in Proceedings of Natl. Acad. Sci, vol. 96(12), USA, 1999, pp. 6745–6750.

[4]A. Bhattacharya and R. De, "Divisive correlation clustering algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," Bioinformatics, vol. 24(11), pp. 1359–1366, 2008.

[5]M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A densitybased algorithm for discovering clusters in large spatial databases with noise," in Proceedings of International
Conference on Knowledge Discovery in Databases and Data Mining (KDD-96), Portland, Oregon, 1996, pp. 226–231.

[6]D. Jiang, J. Pei, and A. Zhang, "DHC: a density-based hierarchical clustering method for time series gene expression data," in Proceedings of BIBE2003: 3rd IEEE International
Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, 2003, p. 393.

[7]R. Jarvis and E. Patrick, "Clustering using a similarity measure based on shared nearest neighbors," IEEE Transactions on Computers, vol. 11, 1973.

[8]A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," J. R. Statist. Soc.,vol. B 39 (1), pp. 1–38, 1977.

[9]J. Travis and Y. Huang, "Clustering of gene expression data based on shape similarity," EURASIP Journal on Bioinformatics and Systems Biology,, vol. 2009(195712), 2009.

[10] M. Beal and Z. Ghahramani, "The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures," in Proceedings of the 7th Valencia International Meeting on Bayesian Statistics, vol. 63(4), Spain, 2003, pp. 453–464.

[11]R. Sharan and R. Shamir, "CLICK: A clustering algorithm with applications to gene expression analysis," in Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology. AAAI Press, 2000.

[12]Ben-Dor, A., Shamir, R. & Yakhini, Z. (1999). Clustering gene expression patterns, Journal of Computational Biology.

[13]Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference)

[14]Kohonen T. Self-Organization and Associative Memory. Spring-Verlag, Berlin, 1984.

[15] Herrero J., Valencia A. and Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126–136, 2001.

© 2014, IJCSMC All Rights Reserved