© 2016 IJSRSET | Volume 2 | Issue 3 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

An Improved Microarray Gene Expression Classification Using Fuzzy Expert System and Ant Bee Algorithm

S. Deepakkumar*, M. Mohankumar, Dr. P. Murugeswari

Computer Science and Engineering, Sri Vidya College of Engineering and Technology, Virudhunagar, Tamilnadu, India

ABSTRACT

Medical diagnosis and dealing of micro array data can be done effectively by means of fuzzy expert system. The main intention of using fuzzy system is high perfection and minimum complexity in dealing with classification of medical data. The existing GSA (Genetic Swarm Algorithm) enables high precision in classification on fuzzy expert system with desired cost on medical research. The main drawback of GSA is if-then rules which is multifaceted and protracted that's highly complicated, those are difficult for a physician to understand. In order to address the interpretability-accuracy tradeoff, a development is made in presenting the rule set by the combination of integer numbers and the task of rule generation. Ant colony optimization (ACO) generates simple rule set according to the gene expression values by which fuzzy partition is applied. But it still suffers from addressing the formless and continuous expression values of a gene. In this paper, we propose artificial bee colony (ABC) algorithm based on mutual information. This mutual information effective in analyzing the informative genes and the improved proposed mechanism hybrid Ant Bee Algorithm (ABA) using fuzzy-II logic is computed with six gene expression data sets in order to examine its effectiveness. The results shows our improved proposed mechanism achieves more perfection in fuzzy system by the combination of highly interpretable and compact rules among all the data sets thus proves its performance is far better than other traditional mechanisms.

Keywords: Medical Diagnosis, Fuzzy expert system, Micro array data, artificial bee colony, ant colony optimization and mutual information.

I. INTRODUCTION

Data mining is a wide area for the researchers in dealing with various topics according to the real time environment. In which recently medical diagnosis is one of the global segment which grabs the attention of the researchers analyzing about symptoms and signs of diseases. To do this micro array data plays a vital role which is a collection of spots extracted from a solid surface. Analyzing of these data's is a tedious process, especially DNA micro arrays. It is a most famous technology dealing with study of gene expression which has thousands of genes for a sample cell. These are data's are usually in image which can be expressed in the form of matrices by row representing genes and column representing tissues.

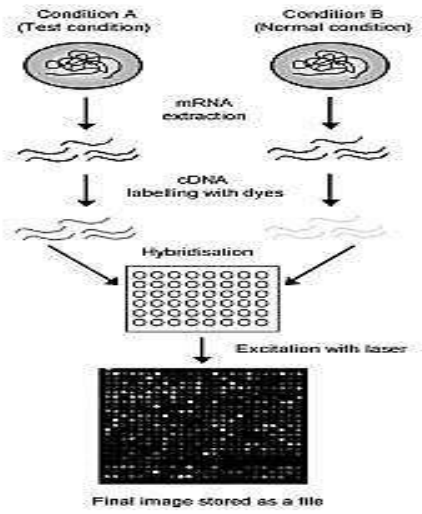

Figure 1. Sample DNA Microarray data

The above fig 1 shows how the sample micro array data forms; it is a thousand of spots copies of same DNA

represent a gene of an organism. Initially these spots arranged to an order with two conditions such as reference condition and test condition as described in fig1 as A & B. Then the RNA is extracted tagged with different dyes which is then hybrid and stored as final image for other research process. Thus it is more complex for a physician to work on these images. In order to resolve it various research scholars presented their work to find a simple mechanism in analyzing these micro array data, thus the growth and some of the best algorithm for these approach which takes minimum computation is discussed in the below sections.

II. METHODS AND MATERIAL

1. Related Works

There are various research work has done in the field of disease classification system in which the statistical methodologies such as weighted voting approach , the nearest neighbor classification, discrimination approach , least square, logistic regression methods and naive bayes approach are utilized in classifying the gene expression data's [1] -[5]. These approaches were not flexible in classifying the system according the gene expression from its predefined profile. These statistical methods can't able to maintain a stable memory resistant for the overall systems. In order to overcome these issues machine learning approaches like support vector [6] and artificial neural network [7] has evolved and successfully processed for classifying micro array data but it's also lack with accuracy as well as interpret problems. Then the decision tree [8] method is applied it is a constructed rule based classifier as it is more sensitive in rules and biological terms its showing large difference in tree structure on classifications. On this way the next approach is symbolic machine learning approach [9] it has some classification rules from decision tree according to human understandable formats but it not shows effective results. In this order symbolic manipulations, rule based classifiers [10] [11] [12] were discussed but those more complex to process. The classification of micro array data involves decision making which has lot of uncertainties according to the information present in the gene expressions. Difficulties in prediction give rise to fuzzy logic which is dealing specially with uncertain situation and vagueness [13] [14]. Fuzzy based classifier follows two logics such as pure fuzzy classifiers and fuzzy rule based system or simply fuzzy expert system. In which pure fuzzy based on fuzzy pattern matching [15], fuzzy clustering [16], and fuzzy integral [17], fuzzy rule based system [18] which are similar to the decision making as human knowledge. The drawback in fuzzy based system is huge number of input genes which makes difficult to compute a definite rule set. A Knowledge acquisition for a fuzzy [19] system is developed in finding the optimal location to overcome the search problems in dealing high dimensional space but the rule set not supports accordingly. Next one is hybrid fuzzy (HF) [20] is developed to formulate the compact rule based rules which is fail to compute the rules at the same time with the membership function and rule set. Some of the recent works such as [21][22][23][24][25] which are applicable for small volume of genes and helpful in obtaining the better diagnostic accuracy but in case of huge volume showing varied information in genetic mechanisms as not results in robustness, scalability and good empirical successes.

2. Background

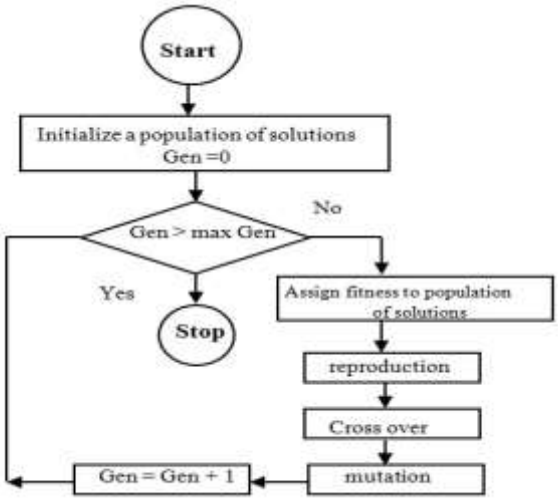

Figure 2: Flow chart for existing system

The existing system is the hybrid combination of Genetic algorithm with Particle Swarm Optimization (PSO) [26][27][28] and it is a heuristic search method moves from a one set of points to another set by the combination of deterministic and probabilistic rules. In general PSO is a stochastic optimization technique used for random solutions and searches. The PSO is well known for solving optimization problems it is computed with GA in order to overcome space problem as result finding a better solution for a classification system. The

above fig2 shows how the existing system works starts with random initialization then swarming the behaviour of the particles towards the location finding a fitness solution. But these rules are based on the binary strings and floating point numbers which uses if-then rule produced by GSA that are hard for the physician to understand and complex too.

3. Proposed System

The proposed type-2 fuzzy system combined with hybrid ant bee algorithm (ABA) to maintain the accuracy in classification with minimum computation cost. In general fuzzy logic same as human reasoning, it process the solution based on the possibilities by the digital values YES and NO. Before getting to type-2, let's discuss how it differs from type-1 fuzzy logic. It is a type reduction and de-fuzzification method in which type-2 is an extended version of type-1 defuzzification method. It captures more information from the set rules and defuzzified value results minimum computation cost. The type-2 method follows simple type-reduction computation procedure and showing best results in time-varying channel. In which inference engine combines rules those are easy for understanding and efficient in dealing with the uncertainties**.**

A. Microarray Data Classification Using Fuzzy Expert System: The method of classifying the micro array data is learning from the sample by which predicting the disease similar to linguistic values. It is a non-linear concept mapping between inputs to output spaces.

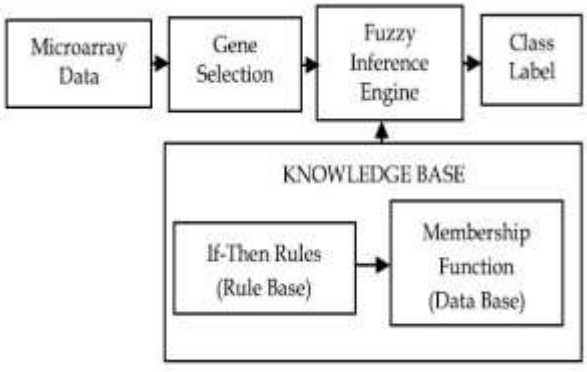

Figure 3. Fuzzy expert system for microarray data classification

The fig 3 shows the working mechanism of fuzzy system, generally micro array data's are high dimensional in nature which need a expert system to deal. After gene selection the inference engine implies mutual information technique in gathering selective information's from the genes. The label class is for indicating the predicted class and the fuzzy system has collection of if-rules for qualitative reasoning. Then the fuzzy relation is constructed by input and output values based on if-rules. Further for the optimized solution the space complexity is solved by ABA with maximum accuracy compared to the other existing methods like GSA.

B. Ant Bee Algorithm with Fuzzy – Type-2 Logic: For the best result we design fuzzy type-2 with Ant Bee Algorithm it works with the mechanism of three phases such as Employed bee phase, Onlooker bee phase and Scout bee phase. The ABA is strengthening by combing ACO and ABC algorithms. ABA is like maintaining colony of ants and permissible range values of desire variables. The ABA rules has linguistics values such as low, medium and high, based on the resultant values classes labeled as normal and diseases.

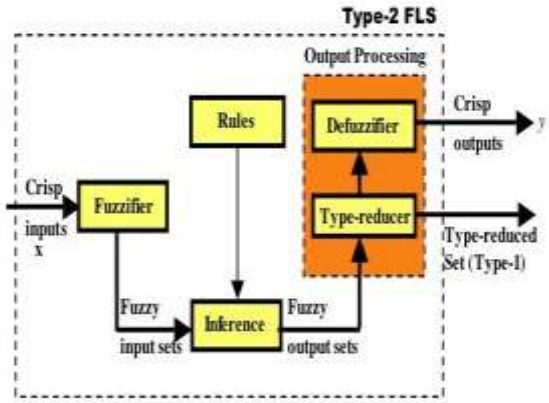

Figure 4. Fuzzy Type-2 Architecture

The type-2 FLS is similar to type-1 FLS in which the fuzzier points the input in the fuzzy set then inference IF- rules and the structure remains the same. The inference process rule set which consist of three sections namely rule selection, input variables and output variables. The fitness section is calculated by formulating maximizing the correctly classified data with minimum difference between total numbers.

Trade off = Linguistic Fuzzy Modeling / Precise Fuzzy Modeling

III. RESULTS AND DISCUSSION

Performance Comparison

The accuracy of the proposed system is based on the Learning ability and generalization ability in which learning ability uses all samples of training patterns. To find a solution for space maximum of 10 rules are generated for understanding the linguistic values " low" represented by 1, "medium" represented by 2 and " high" represented by 3. Each value in the rule set is initialized and probability matrix is constructed. The performance analysis is based on the dataset such as Type 2 Diabetes, Colon cancer (Col), Leukemia (Leu), Lymphoma (Lym), Rheumatoid Arthritis versus Osteoarthritis (RAO) and Rheumatoid Arthritis versus Controls. The six gene expression data sets are extracted from [31][32][33][34][35][36]. These dataset are process under four rules and each result is tabulated as shown below;

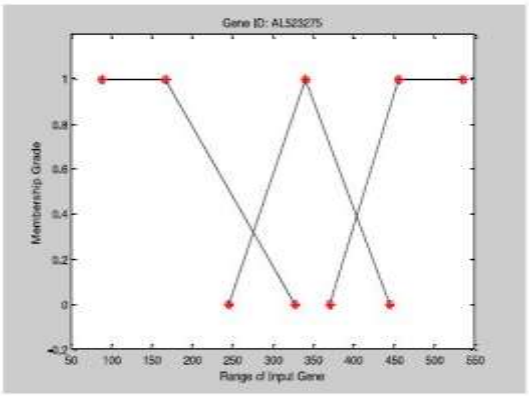

Figure 5. Optimal membership function by ABA

Inference Rules:
Rule 1: If NM_021133.1 is medium and NM_0223491.1 is low then it is NGT.
Rule 2: If AW291218 is low and AL523575 is high then it is NGT
Rule 3: If NM_022349.1 is medium and BC000229.1 is low then it is DM2.
Rule 4: If NM_005260.2 is high and BG339560 is medium then it is DM2

In which the input sets are labeled as $a_{1\ldots}a_6$

Then a= {(x,μ_a (x))|xƐX

Where $\mu_a(X)$ is called the membership function of x in a;

Min f=(S-Cc) + (k*SN_R)

Here k is the constant SNR is the selected number of rules s represents the difference and Cc represents correctly classified data.

Fitness = K/f, where K is another constant.

Data Set	Rule Set	N covers	N corrects	Coverage	Accuracy
Col	R1	28	27	45.16	96.43
	R2	25	23	40.32	92
	R3	13	10	20.97	76.92
Lym	R1	25	18	55.56	72
	R2	12	10	26.67	83.3
	R3	10	9	23.22	90
	R4	11	7	24.44	63.64
Leu	R1	25	18	34.72	72
	R2	32	26	44.44	63.64
RAC	R1	21	17	60	80.95
	R2	23	16	65.71	69.57
RAO	R1	22	15	70.97	68.18
	R2	20	15	64.52	75
T2D	R1	20	16	58.82	80
	R2	12	8	35.29	66.67
	R3	19	5	33.33	55.56
	R4	8	4	29.63	50

The obtained results of ABA are computed on CPU consumption which compared with the other existing systems. The resultant graph on below fig6 shows the effective performance of our proposed algorithm. Thus the proposed system achieves the goal of classifying micro array data with minimum computation and high accuracy in an easy understandable procedure.

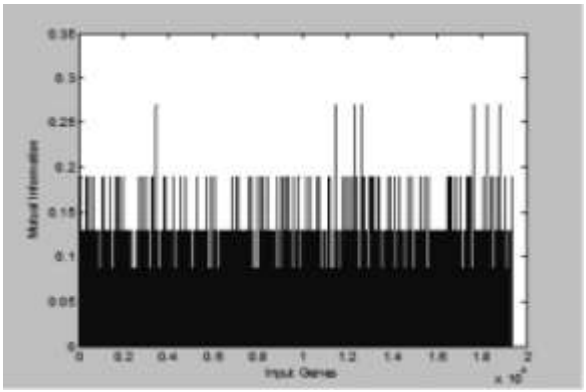

Figure 6. Mutual information for the input genes of T2D

IV. CONCLUSION

Overall, the Type-2 fuzzy expert term examine the best rules by executing the knowledge acquisition in microarray data classification using if-then rules and membership function. The ABA algorithm maximizes the Accuracy-Interpretability tradeoff by means of rule sets represent in integer numbers. The performance comparison of six micro array data set and the ABA results shows how it is effective than the other traditional systems. In the proposed system the length and complexity of the rule is minimized and computes the prediction in a short span of time.

V. REFERENCES

[1] T. R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, pp. 531-537, 1999.

[2] L. Li, "Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the ga/knn Method," Bioinformatics, vol. 17, pp. 1131-1142, 2001.

[3] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," J. Am. Statistics Assoc., vol. 97, no. 457, pp. 77- 87, 2000.

[4] G. Fort and S.L. Lacroix, "Classification Using Partial Least Squares with Penalized Logistic Regression," Bioinformatics, vol. 21, no. 7, pp. 1104-1111, 2005.

[5] L. Fan, K.L. Poh, and P. Zhou, "A Sequential Feature Extraction Approach for Na€ıve Bayes Classification of Microarray Data," Expert Systems with Applications, vol. 36, no. 6, pp. 9919-9923, 2009.

[6] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," Bioinformatics, vol. 16, pp. 906-914, 2000.

[7] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," Nature Medicine, vol. 7, pp. 673-679, 2001.

[8] A.C. Tan and D. Gilbert, "Ensemble Machine Learning on Gene Expression Data for Cancer Classification," Applied Bioinformatics, vol. 2, pp. 75-83, 2003.

[9] D.E. Johnson, F.J. Oles, T. Zhang, and T. Goetz, "A Decision-Tree- Based Symbolic Rule Induction System for Text Categorization," IBM Systems J., vol. 41, no. 3, pp. 1-10, 2002.

[10] J.S.R. Jang, C.T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computting. Prentice Hall, 1997.

[11] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, and D. Geman, "Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles," Bioinformatics, vol. 21, pp. 3896-3904, 2005.

[12] Y. Yoon, S. Bien, and S. Park, "Microarray Data Classifier Consisting of k-Top-Scoring Rank-Comparison Decision Rules with a Variable Number of Genes," IEEE Trans. Systems, Man, and Cybernetics- Part C: Applications and Rev., vol. 40, no. 2, pp. 216-226, Mar. 2010.

[13] P. Woolf and Y. Wang, "A Fuzzy Logic Approach to Analyzing Gene Expression Data," Physiological Genomics, vol. 3, pp. 9-15, 2000.

[14] S. Vinterbo, "Small, Fuzzy and Interpretable Gene Expression Based Classifiers," Bioinformatics, vol. 21, no. 9, pp. 1964-1970, 2005.

[15] G. Schaefer, "Thermography Based Breast Cancer Analysis Using Statistical Features and Fuzzy Classification," Pattern Recognition, vol. 42, no. 6, pp. 1133-1137, 2009.

[16] X. Zong, Z. Yong, J. Li-min, and H. Wei-li, "Construct Interpretable Fuzzy Classification System Based on Fuzzy Clustering Initialization," Int'l J. Information Technology, vol. 11, no. 6, pp. 91- 107, 2005.

[17] Y. Hu, "Fuzzy Integral-Based Perceptron for Two-Class Pattern Classification Problems," Information Sciences, vol. 177, no. 7, pp. 1673-1686, 2007.

[18] Z. Wang and V. Palade, "A Comprehensive Fuzzy-Based Framework for Cancer Microarray Data Gene Expression Analysis," Proc. IEEE Int'l Conf. Bioinformatics and Bioeng., pp. 1003-1010, 2007.

[19] S.M. Chen and F.M. Tsai, "Generating Fuzzy Rules from Training Instances for Fuzzy Classification

Systems," Expert Systems with Applications, vol. 35, no. 3, pp. 611-621, 2008.
[20] G. Schaefer and T. Nakashima, "Data Mining of Gene Expression Data by Fuzzy and Hybrid Fuzzy Methods," IEEE Trans. Information Technology in Biomedicine, vol. 14, no. 1, pp. 23-29, Jan. 2010.
[21] P. Maji, "f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data," IEEE Trans. Biomedicine Eng., vol. 56, no. 4, pp. 1063-1069, Apr. 2009.
[22] P. Maji and S.K. Pal, "Fuzzy-Rough Sets for Information Measures and Selection of Relevant Genes from Microarray Data," IEEE Trans. Systems, Man and Cybernetics, vol. 40, no. 3, pp. 741-752, June 2010.
[23] Y. Leung and Y. Hung, "A Multiple-Filter-Multipe-Wrapper Approach to Gene Selection and Microarray Data Classification," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 7, no. 1, pp. 108-117, Jan./Feb. 2010.
[24] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," , vol. 9, no. 4, pp. 1106-1119, 2012.
[25] Sharma, S. Imoto, and S. Miyano, "A Top-r Feature Selection Algorithm for Microarray Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754-764, May/June 2012.
[26] P. Maji, "f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data," IEEE Trans. Biomedicine Eng., vol. 56, no. 4, pp. 1063-1069, Apr. 2009.
[27] P. Maji and S.K. Pal, "Fuzzy-Rough Sets for Information Measures and Selection of Relevant Genes from Microarray Data," IEEE Trans. Systems, Man and Cybernetics, vol. 40, no. 3, pp. 741-752, June 2010.
[28] Y. Leung and Y. Hung, "A Multiple-Filter-Multipe-Wrapper Approach to Gene Selection and Microarray Data Classification," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 7, no. 1, pp. 108-117, Jan./Feb. 2010.
[29] Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," , vol. 9, no. 4, pp. 1106-1119, 2012.
[30] Sharma, S. Imoto, and S. Miyano, "A Top-r Feature Selection Algorithm for Microarray Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754-764, May/June 2012.
[31] V.K. Mootha, C.M. Lindgren, K.F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, et.al, "PGC-1a Responsive Genes Involved in Oxidative Phosphorylation are Coordinately Down Regulated in Human Diabetes," Nature Genetics, vol. 34, no. 3, pp. 267-273, 2003.
[32] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proc. Nat'l Academy of Sciences USA, vol. 96, no. 12, pp. 6745-6750, 1999.
[33] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick et.al, "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," Nature, vol. 403, no. 3, pp. 503-511, 2000.
[34] T.C. Kraan, F.A. Gaalen, P.V. Kasperkovitz, N.L. Verbeet, T.J. Smeets, M.C. Kraan, "Rheumatoid Arthritis Is a Heterogeneous Disease: Evidence for Differences in the Activation of the STAT-1 Pathway between Rheumatoid Tissues," Arthritis and Rheumatism, vol. 48, no. 8, pp. 2132-2145, 2003.
[35] V.H. Teixeira, R. Olaso, M.L.M. Magniette, S. Lasbleiz, L. Jacq, C.R. Oliveira, P. Hilliquin, "Transcriptome Analysis Describing New Immunity and Defense Genes in Peripheral Blood Mononuclear Cells of Rheumatoid Arthritis Patients," PLoS ONE, vol. 4, no. 8, p. e6803, 2009.