Classification of Micro Array Gene Expression Data using Factor Analysis Approach with Naïve Bayesian Classifier

Tamilselvi Madeswaran[#1] G.M.Kadhar Nawaz[*2]

Research Scholar, Anna University of Technology, Coimbatore, Tamilnadu, India

Director & Professor , Department of Computer Application, Sona College of Technology, Salem, Tamilnadu, India

***Abstract*—Microarray data studies produce large number of data and in order to analyze such large micro array data lies on Data mining or Statistical Analysis. Our objective is to classify the micro array gene expression data. Usually before going for the classification the dimensionality reduction will be performed on the micro array gene expression dataset. A statistical approach for the extraction of the gene has been proposed. The drawback in the statistical analysis is that, it doesn't identify the important genes. Here for the classification process we use k-nearest neighbor and SVM and Naïve Bayesian classifiers. From the experimental result our proposed classifiers show increase in the efficiency and accuracy.**

Keywords-Micro array gene expression data, Gene patterns, Statistical approach, Dimensionality reduction, KNN, Naïve Bayesian Classifier.

I. Introduction

Microarray technology is an emerging widely used platform for genomic studies in biomedical fields. Microarray technology yields a systematic way to survey biological data variations. Since large number of data will produce from microarray studies the analysis of such data lies on data mining and statistical analysis. Large amount of data are stored in a database. These data may be come from an organization or some gene expression or log data. From the vast data, researchers are doing research over the extraction of unidentified patterns in the large data. But for the former algorithms and tools make the retrieval in some hours. So we have found that the Data Mining is a optimal solution for identifying the information in the database. Some patterns in the data results in the computational efficiency restrictions in the fundamental steps of Knowledge Discovery in Databases (KDD) [3]. In the former researches they have used many statistical tools and mathematical models [4] [5] for finding the unknown patterns and relationships [7] [8] [9].

Micro array technology is an emerging thing in tracking the genome expressions of genes [15]. The micro array technology may operate the diagnostic task and improve the efficiency of the traditional diagnostic methods. The analysis of huge number of gene expressions can be made from the micro array technologies [17]. Gene expression microchip is the most developing tool in the analysis and monitoring the expression levels of huge amount of genes at the same time [18].

II. Related Works

Li-Yeh Chuang et al. [20] proposed that the Support Vector Machine (SVM) produces same or efficient result than the neural networks in the process of learning. They have implemented SVM to take advantages over the fuzzy logic and statistical theories and group the gene expression profiles. FSVM (Fuzzy Support Vector Machine) used the proposed strategies and outlier detection methods to achieve an equivalent or superior performance over the former methods in differentiating the SRBCT and non-SRBCT samples.

Edmundo Bonilla Huerta et al. [21] proposed the Genetic Algorithm (GA) which integrates the Support Vector Machines (SVM) for the high dimensional micro array data categorization. In this approach a pre-filtering technique based fuzzy logic has been proposed. The gene subset fitness values have been calculated using the SVM classifiers using GA. The most important genes have been identified by the frequency based technique using "good" gene subsets. They have evaluated the proposed methodology with the six existing methods towards the cancer datasets.

Hieu Trung Huynh et al. [22] discussed about the DNA micro array used in the molecular biology and biomedicine. Former methods are used to analyze the result of an arrayed sequence of thousands of microscopic spots of DNA contained. In the recent years, the usage of intelligent computing has been utilized for the analysis of the micro array data. The former researches state a method called Single hidden layer feed forward Neural Network for DNA micro array classification which used SVD (Singular Value Decomposition) for training process. The activation function of the hidden units "tansig" for the classifier for the single hidden-layer feed forward neural network. The evaluation result states that the training procedure and also the network structure of the SVD trained result has been given to the neural network. The proposed approach yields better result when compared to the exiting approach.

Pradipta Maji et al. [23] discussed about the usage of many information measures like entropy, mutual information and f-information has been provided. The proposed methodology has been proved for its success by selecting the relevant and non redundant genes from the high-dimensional micro array data set. While calculating the true density functions and executing the gene expression values, the true marginal and joint distributions of continuous gene expression values have been approximated in introducing the concept of the fuzzy equivalence of the partition matrix. The row of the matrix is calculated by a fuzzy equivalence partition which can be automatically extracted from the specified expressions. The class seperability index and predictive accuracy of the SVM is compared with the existing approach to calculate its performance.

Venkatesh et al. [24] studied about the genes and their functionality over the genomics. There are several former methods are evaluated for analyzing the thousands of genes in the micro array analysis. In the evaluation they have used Gene samples from the biopsy samples from the color cancer patients. In the recent dimensionality reduction in the gene expression data the results are not so accurate, so our major objective to overcome the drawbacks.

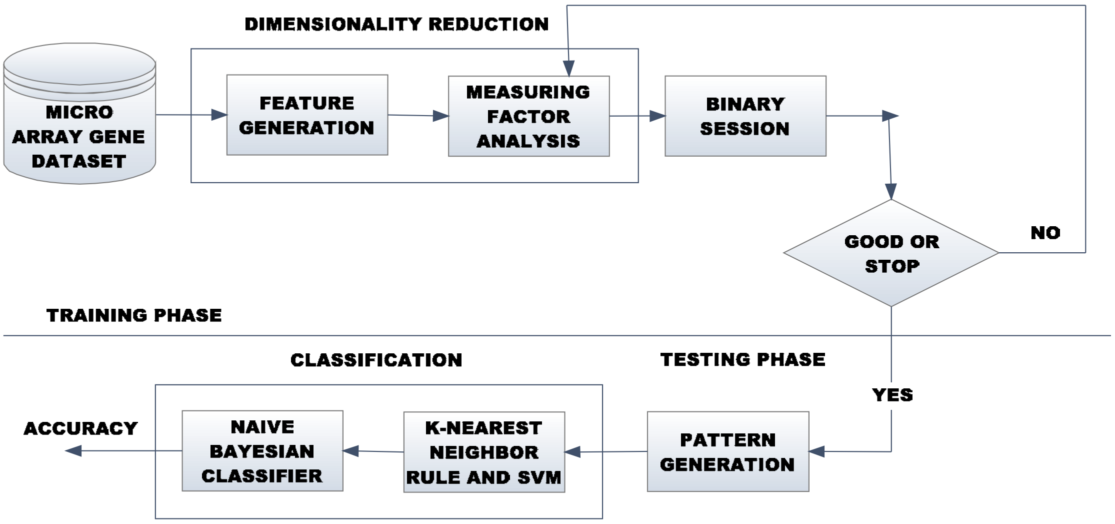

Fig 1: Structure of Our Proposed Classification System for Micro Array Gene Expression Data

In this paper an effective classification technique has been proposed which uses KNN and Naïve Bayesian classifiers. The dimensionality reduction technique can be performed by statistical approaches and the reduced data will be obtained. From the reduced data the important genes are found out. The KNN classifiers and Naïve Bayesian classifiers is developed for the classification. The rest of the paper is organized as follows; section 3 explains about the proposed classification technique with statistical operations. Section 4 tells about the implementation results and section 5 concludes the paper.

III. Classification Technique for Micro Array Gene Expression

In this paper, an efficient technique for the classification of micro array gene expression has been proposed. It consists of four major parts:

- Dimensionality reduction
- Binary Session
- Pattern Generation
- KNN and Naïve Bayesian classifier are for gene classification.

A. Dimensionality reduction using Factor Analysis method

Let us assume the micro array gene expression data as $X_{nm} : 0 \le n \le N, 0 \le m \le M$ where *N, M* represents the number of samples and genes respectively. The gene data can be written as

$$X_{nm} = \begin{bmatrix} x_{(1,1)}x_{(1,2)}\ldots x_{(1,M)} \\ x_{(2,1)}x_{(2,2)}\ldots x_{(2,M)} \\ \ldots.. \\ \ldots.. \\ x_{(N,1)}x_{(N,2)}\ldots x_{(N,M)} \end{bmatrix} \quad (1)$$

The individual row and column is depicted in the equation (2) where N and M is the total number of rows and columns in the dataset.

$$R_n = \{r_1, r_2, \ldots.r_N\}, C_m = \{c_1, c_2, \ldots..c_M\} \quad (2)$$

The gene expression data X_{nm} is having higher degree and it is reduced by the dimensionality reduction. This process can be obtained by the factor analysis method. We can estimate the parameters of a factor analysis model (W and σ_j) by maximum likelihood.

The factor analysis model treats the data in a probabilistic way. Let us consider that each data item $d = (d_1, d_2, \ldots.., d_p)$ is generated with *M* latent variables $z_1, z_2, \ldots.., z_m$. The relation of d with z is the linear one.

The observed data points d can be obtained by the following:

$$d = \mu + W_z + \in \qquad (3)$$

where μ is the mean vector for the *p* components of d, W is a *p x M* matrix. Є is a vector of *p* residuals which is considered as independent and comes from the Gaussian distribution with mean zero. The variance can be obtained by $Є_j$ is σ_j^2 .

B. *Binary Session*

Binary session is carried out by threshold value. In the binary process, the matrix *M, D, PS and SS* will be involved and these matrixes rows and columns are depicted as:

$$R^{(a)}_{n`} = \{r^{(a)}_1, r^{(a)}_2, \ldots r^{(a)}_{R`}\}, C^{(a)}_{m`} = \{c^{(a)}_1, c^{(a)}_2, \ldots c^{(a)}_{c`}\} \, (4)$$

$$R^{(e)}_{n`} = \{r^{(d)}_1, r^{(d)}_2, \ldots r^{(d)}_{R`}\}, C^{(d)}_{m`} = \{c^{(d)}_1, c^{(d)}_2, \ldots c^{(d)}_{c`}\} \, (5)$$

$$R^{(ps)}_{n`} = \{r^{(ps)}_1, r^{(ps)}_2, \ldots r^{(ps)}_{R`}\}, C^{(ps)}_{m`} = \{c^{(ps)}_1, c^{(ps)}_2, \ldots c^{(ps)}_{c`}\} \, (6)$$

$$R^{(ss)}_{n`} = \{r^{(ss)}_1, r^{(ss)}_2, \ldots r^{(ss)}_{R`}\}, C^{(ss)}_{m`} = \{c^{(ss)}_1, c^{(ss)}_2, \ldots c^{(ss)}_{c`}\} \, (7)$$

The threshold value in the binary session is used to change the values in the matrixes. It is also used to reduce the execution complexity; binary session makes the classification easier in the further gene expressions. The binary session can be explained in the below pseudo code.

```
Input: Matrices M, D, PS and SS
Output: Binarized Matrices M`, D`, PS` and SS`
Step 1: select column from each matrix M, D, PS and SS
Step 2: Determine
    $L_{m'} = \min(C^{(a)}_{m`})$
    $H_{m'} = \min(C^{(a)}_{m`})$
Step 3: Find threshold
```

$$T_{m'} = \frac{L_{m`} - H_{m`}}{2}$$

Step 4: Modify $C_{m`}^{(a)}(x_{(n`,m`)})$ with

$C_{m`}^{(a)}(x_{(n`,m`)})$ ={0; if $x_{(n`,m`)} < T_{m`}$

$C_{m`}^{(a)}(x_{(n`,m`)})$ ={1; if $x_{(n`,m`)} > T_{m`}$

Step 5: repeat steps 2 to 6 until all binarized matrixes are obtained for all matrixes *M, D, PS* and *SS*.

Step 6: return the matrixes *M`, D`, PS`* and *SS`*

C. Pattern Generation

The binary session process gives the results as patterns. In the binary session each binary elements are obtained as the set of matrix with each matrix. In the set of matrix the first element of the each matrix is considered as one gene pattern, the second element is considered as another gene and this process is repeated for the last value of the given matrix.

The generated gene pattern is given as:

$$P = \begin{bmatrix} p_{11}p_{12}\cdots p_{1h} \\ p_{21}p_{22}\cdots p_{2h} \\ ..... \\ ..... \\ p_{g1}p_{g2}\cdots p_{gh} \end{bmatrix} \quad (8)$$

Where *g, h* is the number of columns and rows.

$$p_{gh} = (M'x_{(g,h)}D'x_{(g,h)}PS'x_{(g,h)}, SS'x_{(g,h)}) \quad (9)$$

In our process we used two types of cancer datasets for the gene. The generated pattern P contains column g from 1:38, whereas ALL dataset contains 1:26 and AML dataset contains 27:38. If the gene pattern contains 1011 like pattern then the first, second, third and fourth element will be obtained from the matrixes *M`, D`, PS` and SS`*. Then the pattern selection will be done with the help of the following pseudo code.

Input: Pattern Matrix *P*

Output: Select gene patterns *P`*

Step 1: select p_{gh}

Step 2: Count p_{gh} in *P*

Step 3: Sort *P* in ascending order based on count // row wise sorting.

Step 4: Compare p_{gh}

$$(p_{1h} \,\&\, p_{27h}, p_{2h} \,\&\, p_{28h} \cdots p_{26h} \,\&\, p_{38h})$$

Step 5: *P`={P_{gh}: if $P_{1h} \neq P_{27h}$….$P_{26h} \neq P_{38h}$*

P`={0; otherwise

Step 6: return *P`*

After this Pattern Matrix has been found we would go for the gene classification part.

D. Classification of Micro gene expression using KNN and Naïve Bayesian Classifiers

1. KNN Rule

The k-NN rule is used to introduce notation. In the *k*-NN rule the patterns which we are going to classify are symbolized as vectors in a d-dimensional Euclidean space E^d. From the set of Training samples and query patterns, the *k*-NN rule finds the *k* nearest neighbors.

$$d(\overrightarrow{X}, \overrightarrow{X_i}) = \left(\sum_{j=1}^{d} | X^j - X_i^j |^2 \right)^{1/2} \quad (10)$$

$$d(\overrightarrow{X},\overrightarrow{X_i}) = \left(\sum_{j=1}^{d} | X^{j} - X_i^{j} | \right) \qquad (11)$$

The similarity can be found by Euclidean measures and Manhattan measures and this can be depicted from the equations (10), (11).

2. Naïve Bayesian Classifier

Naïve Bayesian classifier is a simple probabilistic classifier model which is based on the Bayes Theorem with independence assumptions. The probabilistic model of the classifier is the “independent feature model”.

Since the probability model is a unique in nature, Naïve Bayes classifiers could be trained effectively. In several applications parameter estimation for Naïve Bayesian model uses the maximum likelihood method.

Naïve Bayesian classifier joins with the decision rule. The corresponding classifier “Classify” function can be defined as follows:

$$\begin{aligned} &classify\,(f_1,\ldots., f_n) = \\ &\arg\max_{c}\ p(C=c)\prod_{i=1}^{n} p(F_i = f_i \mid C = c) \end{aligned} \qquad (12)$$

Where $f_1,\ldots., f_n$ represents the feature variables n. C is the dependent class variable and c represents a class.

Naïve Bayes classifier needs only a small amount of training data to calculate the parameter which are needed for the classifications. This is the major advantage over the Naïve Bayes classifier.

IV. Result and Discussion

We use MATLAB platform 7.8 version for the implementation and the evaluation results are shown as follows. We have used the micro array gene expression data which is having the dimension of X = 7192 and Y = 38. After using the factor analysis the dimension of the data has been reduced to X = 100 and Y = 38 is achieved.

The performance can be shown with the following table. This table compares our previous results and our proposed approach yields better result.

Statistical Measures	Fuzzy neural Network PCA	Fuzzy neural Network ProbPCA	Fuzzy Genetic System	Fuzzy Inference System	KNN and Naïve Bayesian
TP	13	12	7	4	3
TN	17	15	18	25	26
FP	8	10	7	0	0
FN	0	1	6	9	8
Sensitivity (%)	100	92	54	31	28
FPR (%)	32	40	28	0	
Accuracy (%)	79	71	66	76	92
Specificity (%)	68	60	72	100	100
PPV (%)	62	55	50	100	100
NPV (%)	100	94	75	74	76
FDR (%)	38	45	50	0	
MCC	1.68	0.5	0.25	0.48	

Our proposed KNN rules and Naïve Bayesian classifier is showing results better than the previous approaches. TP, TN, FP, FN represents here as True Positive, True Negative, False Positive, False Negative.

V. Conclusion

A novel method of KNN rule and Naïve Bayesian classifier is proposed in this approach. For the dimensionality reduction we use factor analysis. The dimensionality reduction yields better result and so the classification process provides better result when compare to the earlier fuzzy genetic systems and fuzzy inference system, etc.,

References

[1] Osmar, “Introduction to Data Mining”, In: Principles of Knowledge Discovery in Databases, CMPUT690, University of Alberta, Canada, 1999

[2] Kantardzic and Mehmed, “Data Mining: Concepts, Models, Methods, and Algorithms”, John Wiley & Sons, 2003

[3] Umarani and Punithavalli, "A Study on Effective Mining of Association Rules from Huge Databases", International Journal of Computer Science and Research, Vol. 1, No. 1, pp. 30-34, 2010

[4] Chieh-Yuan Tsai and Min-Hong Tsai, " A dynamic Web service based data mining process system", In Proceedings of the 5th IEEE International Conference on Computer and Information Technology, pp. 1033-1039, 21- 23 September, 2005

[5] Lamine M. Aouad, Nhien-An Le-Khac and Tahar M. Kechadi, "Distributed Frequent Itemsets Mining in Heterogeneous Platforms", Journal of Engineering, Computing and Architecture, Vol. 1, No. 2, 2007

[6] J. Han and M. Kamber, “Data Mining: Concepts and Techniques. Morgan Kaufman, San Francisco, 2000

[7] Bigus, “Data Mining with Neural Networks”, McGraw-Hill, 1996

[8] Klaus Julisch,” Data Mining for Intrusion Detection -A Critical Review”, In Proceedings of the IBM Research on application of Data Mining in Computer security, Chapter 1 , 2002

[9] Hewen Tang, Wei Fang and Yongsheng Cao, “A simple method of classification with VCL components”, In Proceedings of the 21st international CODATA Conference, 2008

[10] Umarani and Punithavalli, "A Study on Effective Mining of Association Rules From Huge Databases", International Journal of Computer Science and Research, Vol. 1, No. 1, pp. 30-34, 2010

[11] Yendrapalli, Basnet, Mukkamala and Sung, “Gene Selection for Tumor Classification Using Microarray Gene Expression Data", In Proceedings of the World Congress on Engineering, London, U.K., Vol. 1, 2007

[12] Sandrine Dudoit, Jane Fridlyand and Terence P. Speed, “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data”, Journal of the American Statistical Association, Vol. 97, pp. 77-87, 2002

[13] Peterson and Ringner, “Analyzing Tumor Gene Expression Profiles”, Artificial Intelligence in Medicine, Vol. 28, No. 1, pp. 59-74, 2003

[14] Anandhavalli Gauthaman, "Analysis of DNA Microarray Data using Association Rules: A Selective Study", World Academy of Science, Engineering and Technology, Vol.42, pp.12-16, 2008

[15] Chintanu K. Sarmah, Sandhya Samarasinghe, Don Kulasiri and Daniel Catchpoole, "A Simple Affymetrix Ratio-transformation Method Yields Comparable Expression Level Quantifications with cDNA Data", World Academy of Science, Engineering and Technology, Vol. 61, pp.78-83, 2010

[16] Khlopova, Glazko and Glazko, “Differentiation of Gene Expression Profiles Data for Liver and Kidney of Pigs”, World Academy of Science, Engineering and Technology, Vol. 55, pp. 267-270, 2009

[17] Ahmad m. Sarhan, "Cancer classification based on microarray gene expression data using DCT and ANN", Journal of Theoretical and Applied Information Technology, Vol. 6, No. 2, pp. 207-216, 2009

[18] Ying Xu, Victor Olman and Dong Xu, "Minimum Spanning Trees for Gene Expression Data Clustering", Genome Informatics, Vol. 12, pp. 24–33, 2001

[19] Lucila Ohno-Machado, Staal Vinterbo and Griffin Weber, "Classification of Gene Expression Data Using Fuzzy Logic", Journal of Intelligent & Fuzzy Systems, Vol. 12, No. 1, pp. 19-24, January 2002

[20] Li-Yeh Chuang, Cheng-Hong Yang and Li-Cheng Jin, "Classification Of Multiple Cancer Types Using Fuzzy Support Vector Machines And Outlier Detection Methods", Biomedical Engineering applications, Basis and Communications, Vol. 17, No. 6, pp. 300-308, December 2005

[21] Edmundo Bonilla Huerta, Beatrice Duval and Jin-Kao Hao, "A hybrid GA/SVM approach for gene selection and classification of micro array data", In Lecture Notes in Computer Science, pp. 34-44, Springer, 2006

[22] Hieu Trung Huynh, Jung-Ja Kimand Yonggwan Won, "Classification Study on DNA Micro array with Feed forward Neural Network Trained by Singular Value Decomposition", International Journal of Bio- Science and Bio- Technology Vol. 1, No. 1, pp. 17-24, December, 2009

[23] Pradipta Maji and Sankar K. Pal, "Fuzzy–Rough Sets for Information Measures and Selection of Relevant Genes from Micro array Data", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 40, No. 3, pp. 741-752, June 2010

[24] Venkatesh and Thangaraj, "Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer", International Journal on Computer Science and Engineering, Vol. 02, No. 06, pp. 2114-2116, 2010.