Classification of Gene Expression Data by Gene Combination using Fuzzy Logic

Daxa Ghadiya[1], Prof. Ankit Kharwar[2], Prof. Monali gandhi[3]

[1] *Student of M.Tech Computer Engineering in Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli*

[2] *Assistant Professor, Department of Computer Engineering, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli*

[3] *Assistant Professor, Department of Computer Engineering, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli*

ABSTRACT

The goal of microarray experiments is to identify genes that are differentially transcribed with respect to different biological conditions of cell cultures and samples. Among the large amount of genes presented in gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. Hence, one of the major tasks with the gene expression data is to find groups of co regulated genes whose collective expression is strongly associated with the sample categories or response variables. A framework is improved/ modified in this report to find informative gene combinations and to classify gene combinations belonging to its relevant subtype by using fuzzy logic. The genes are ranked based on their statistical scores and highly informative genes mare filtered. Such genes are fuzzified to identify 2-gene and 3-gene combinations and the intermediate value for each gene is calculated to select top gene combinations to further classify gene lymphoma subtypes by using fuzzy rules. Finally the accuracy of top gene combinations is compared with clustering results. The classification is done using the gene combinations and it is analyzed to predict the accuracy of the results. The work is implemented using java language.

Keyword: - *Feature selection, classification, clustering, T-test, Cancer classification, gene expression, fuzzy, neural networks, supports vector machines.*

1. INTRODUCTION

In today's world people are flooded by data like scientific data, medical data, financial data and marketing data. However the Human capacity is limited and a common human cannot deal with such massive amount of data to extract the information. The solution is to develop automatic techniques and systems – to analyse the data, to classify the data, to summarize it, to discover and characterize trends in it.

Data mining is the extraction of interesting relations and patterns hidden in the datasets. It combines database technologies, statistical analysis and machine learning .Today data mining techniques are innovatively utilized in numerous fields like industry, Commerce and Medicine. [1] The data results generated by data mining techniques are highly priced by the professionals. For Instance in this dissertation, medical data is used as a fuzzy logic, classification algorithm.

Genes are fundamental physical and functional inheritance units of every living organism. The coding genes are templates for synthesis of proteins. Other genes might specify RNA templates as machines for production of different types of RNAs.

The process in which DNA is transcribed into mRNA and proteins are produced by translation represents the well-known central dogma in molecular biology. The first stage is transcription, then the second stage – translation of mRNA into a sequence of amino acids that compose the protein. When a protein is produced, the corresponding coding gene is expressed.

The gene expression levels indicate the approximate number of produced RNA copies from corresponding gene, which means that gene expression level corresponds to the amount of produced proteins. DNA microarray technology is used to obtain gene expression data experimentally.

One of the most important regulatory functions of proteins is transcription regulation. Proteins, which bind to DNA sequences and regulate the transcription of DNAs and gene expression, are called transcription factors (TFs). TFs can inhibit or activate gene expression of the target genes [3]. Besides gene expression data, other data such as protein-DNA, protein-protein interaction data and microRNAs should be considered for revealing gene regulatory mechanisms.

2. LITERATURE SURVEY

Microarrays are capable of profiling the gene expression patterns of tens of thousands of genes in a single experiment. Gene expression data can be a valuable source for understanding the genes and the biological associations between them. It has high dimension, small samples and the gene selection which means that feature selection is very important to determine the classification accuracy. The dataset utilized for this work is called Lymphoma Dataset which includes 4026 gene expression values with its subtypes [1].

There are three types of lymphomas such as diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), and chronic lymphocytic leukemia (CLL) [1]. The entire data set includes the expression data of 4,026 genes each measured using a specialized cDNA microarray with its relevant Genbank accession number, Name and Clone IDs. A part of the dataset is chosen for the proposed work to classify lymphoma subtypes consists of hundred genes with gene expression values of 62 samples, with a total of 6200 samples and it is called as the Test dataset.

Table 1. A Sample data from Lymphoma Dataset

GENE ID	NAME	VALUES	VALUES	VALUES
GENE 312 9X	Anticrime motility factor receptor Clone=1072873	-0.3000	0.3000	0.5900
GENE 312 6X	2B catalytic subunit Clone=627173	-0.2200	-1.2100	1.4100
GENE 307 2X	APC Clone=125294	-0.0400	0.1500	0.6800
GENE 306 7X	Probable ATP Clone=1350869	0.4100	-0.3400	-0.1800
GENE 400 6X	6XSRC-like adapter protein Clone=701768	1.7600	1.2100	0.9900

Now, next phase is Preprocessing phase. Data pre-processing is an often neglected but important step in the data mining process. Preprocessing is the process of removal of noisy data and filtering necessary information.

Table 2. Lymphoma Dataset with empty spots

GENE	NAME	VALUES	VALUES	VALUES	VALUES
GENE 1835X	(Clone= 1357915)	-0.1300		-0.2800	0.0400
GENE 1836X	(Clone= 1358277)	-0.3100	0.1600		0.2500
GENE 1865X	(Clone= 1358064)	-0.1200	0.5200		0.8300
GENE 1933X	(Clone= 1358190)	0.0500			0.2800
GENE	(Clone=	-0.2600		-0.0900	0.1500

1932X	1336836)				
GENE 1931X	(Clone= 1336983)	-0.5500			

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

Table 3. Preprocessed Lymphoma Dataset

GENE	NAME	VALUES	VALUES	VALUES	VALUES
GENE 1835X	(Clone=1357915)	-0.1300	-0.2800	-0.2800	0.0400
GENE 1836X	(Clone=1358277)	-0.3100	0.1600	0.1600	0.2500
GENE 1865X	(Clone=1358064)	-0.1200	0.5200	0.5200	0.8300
GENE 1933X	(Clone=1358190)	0.0500	0.0500	0.0500	0.2800
GENE 1932X	(Clone=1336836)	-0.2600	-0.0900	-0.0900	0.1500
GENE 1931X	(Clone=1336983)	-0.5500	-0.5500	-0.5500	-0.5500

Now further we can define ranking if gene. Gene ranking simplifies gene expression tests to include only a very small number of genes rather than thousands of genes. The importance ranking of each gene is done using a feature ranking measure called T-Test which ranks the genes based on their statistical score.

Table 4. List of genes with T-scores

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

After that we can find the finding informative genes and in that it greatly reduces the computational burden and noise arising from irrelevant genes. Gene 1 means the gene ranked first as Shown in following Table. The set of informative genes are passed as input to the next phase for fuzzy classification.

Table 5. Informative genes based on their T-scores

GENEID	T-SCORE	GENE
GENE1943X	0.2047	1
GENE880X	0.1842	2
GENE324X	0.1785	3
GENE1557X	0.1641	4

GENE2231X	0.1598	5
GENE289X	0.1569	6
GENE1792X	0.1559	7
GENE910X	0.1548	8
GENE272X	0.1547	9
GENE692X	0.1541	10

Now, next phase is Fuzzy classification. In this phase the set of informative genes with gene expression data are converted into fuzzy values using Type 1 fuzzy. The first step in fuzzification is to take the crisp inputs and covert to fuzzy values. The second step is to take the fuzzified inputs, and apply them to the antecedents of the fuzzy rules. The fuzzified informative genes are passed as input to the next process to identify various gene combinations.

Specifically Single gene, Two-gene and three-gene combinations are done with the selected informative genes. The Single gene, two gene and three gene combinations are identified to classify the lymphoma subtypes such as DBLCL, FL and CLL. The single gene is identified from the whole informative gene set which consists of 100 genes.

Now apply the intermediate value for the single gene, two gene, three gene combinations and intermediate values are calculated for top gene combinations to frame fuzzy rules and to classify the lymphoma subtypes in the test dataset. With using the fuzzy rules we can classify the test dataset.

Gene Combinations

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

Table 8.Single Gene Classification Accuracy

Gene	**Lymphoma Subtypes**		
	DLBCL	FL	CLL
GENE 50	72%	67%	5%
GENE 55	74%	34%	35%
GENE 84	72%	51%	20%
GENE 96	77%	7%	60%

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

Next is two gene combination classified lymphoma subtypes for several genes in the test dataset.(GENE3,GENE1),(GENE23,GENE40), and (GENE3, GENE67) classified DLBCL subtypes of all the hundred genes within the subtype limit. The two gene classification on test dataset is shown in Table9.

Table 9. Classification Accuracy of two gene combinations

Gene	**Lymphoma Subtypes**		
	DLBCL	FL	CLL
(GENE3,GENE1)	62%	24%	58%
(GENE23,GENE40)	64%	36%	43%

(GENE3,GENE67)	77%	50%	57%

The best two gene combination which classified DLBCL and FLL within subtype limits is(GENE3,GENE67).

The three gene combination classified lymphoma subtypes for several genes in the test dataset. (GENE 98,GENE 89, GENE 32) classified DLBCL and FL accurately within the constraint and it attained 63% accuracy. (GENE1, GENE2, GENE3) classified DLBCL subtypes of all the hundred genes within the subtype limit and attained 62% accuracy. (GENE 59, GENE 71, GENE 94) attained 58% accuracy in classifying DLBCL subtype within the limit. The three gene combinations and its accuracy are as shown in the Table 10.

Table 10. Classification Accuracy of three gene combinations

Gene	**Lymphoma Subtypes**		
	DLBCL	FL	CLL
(GENE59,GENE71,GENE 94)	58%	31%	54%
(GENE98,GENE89,GENE 32)	63%	50%	80%
(GENE1,GENE2,GENE3)	62%	41%	41%

The three gene combination **(GENE 98, GENE 89, and GENE 3**2) is the best one to classify DLBCL and FL within the subtype limits. All other combinations can be used to classify DLBCL subtype.

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

Finally gene combinations are verified and its correlation is compared with hierarchical clustering approach by grouping the entire informative genes. Then the classification accuracy of the gene combination is analyzed based on its efficiency of subtype's classification such as DLBCL, FL and CLL of the test dataset.

3. Proposed Approach

In this project we are presenting the hybrid framework for classification of microarray gene expression data using the gene combination with fuzzy logic and support vector machine (SVM) with goal of improving both classification accuracy and scalability. Here we are using the fuzzy logic approach while feature selection. This is presented to find informative gene combinations and to classify gene combinations belonging to its relevant subtype by using fuzzy logic. The genes are ranked based on their statistical scores and highly informative genes are filtered. Such genes are fuzzified to identify 2-gene and 3-gene combinations and the intermediate value for each gene is calculated to select top gene combinations. Still to this we can say it as construction of feature descriptors for classification. The final step in this project is to present the scalable classification framework in which we prepare training and test datasets using above feature construction method with identified gene combinations. For classification we are going to use efficient classification method SVM to claim efficiency of our proposed approach.

4. CONCLUSIONS

We can conclude that among the large amount of genes present in gene expression data, only a small fraction of them is effective for performing classification. Such informative genes are retained by a process called feature selection. The proposed 2-gene and 3-gene combinations are verified with a clustering approach called Hierarchical clustering which proved that gene combination taken are good combinations in classifying lymphoma subtypes. The classification accuracy of gene combination is verified in the final phase. From the experimental results it was found

that from the top hundred genes ranked based on T-scores, single gene selected classified 77% of DLBCL subtypes, 67% of FLL subtypes for all hundred genes in the test dataset, two gene combinations was found to have 77% of DLBCL subtypes, 50% of FL subtypes for all hundred genes in the test dataset and three gene combinations was found to have 63% of DLBCL subtypes, 50% of FL subtypes for all hundred genes in the test dataset. The evolutionary approaches such as optimization methods can be used to generate best gene combinations to achieve higher level classification accuracy.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere thanks and deep sense of gratitude to my guide for imparting me valuable guidance. They helped me by solving many doubts and suggesting many references. They helped me by giving valuable suggestions and encouragement which not only helped me in preparing this thesis but also in having a better insight in this field.

REFERENCES

[1] V.Bhuvaneswari,Vanitha, Classification of Microarray Gene Expression Data by Gene Combinations using Fuzzy Logic (MGC-FL), International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.4, August 2012.

[2] Jin-Hyuk Hong and Sung-Bee Cho, Cancer Classification with Incremental Gene Selection based on DNA Microarray Data, Computational Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB '08. IEEE Symposium on 15-17Sept. 2008.

[3] Jin-Hyuk Hong and Sung-Bee Cho, Cancer Classification with Incremental Gene Selection based on DNA Microarray Data, Computational Intelligence in Bioinformatics and Computational Biology, 2008. CIBCB '08. IEEE Symposium on 15-17 Sept. 2008.

[4] Lipo Wangand Feng Chu, Extracting Very Simple Diagnostic Rules from Microarray Data, IEEE, Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference on Aug. 31 2010-Sept. 4 2010.

[5] Vaishali P Khobragade, Dr.A.Vinayababu, A Classification of Microarray Gene Expression Data Using Hybrid Soft Computing Approach, International Journal of Computer Science, Engineering and Applications (IJCSEA) ,Vol.9, No.2, November 2012.

[6] S.J.Brintha ,V. Bhuvneswari, Clustering Microarray Gene Expression Data Using Type 2 Fuzzy Logic, International Journal of Computer Science, Engineering and Applications (IEEE), March 30-31 2012.

[7] Chhanda Ray, Cancer Identification and Gene Classification using DNA Microarray Gene Expression Patterns, International Journal of Computer Science, Engineering and Applications (IEEE) Vol.8, Issue.2, March 2011.

[8] Lipo Wang, Feng Chu, and Wei Xie, Accurate Cancer Classification Using Expressions of Very Few Genes, Computational Biology and Bioinformatics, IEEE/ACM
Transactions on 20 February 2007.

[9] Yu, S., Zhang, Y., Song, C., Chen, K.: "A security architecture for Mobile Ad Hoc Networks", Proceedings of Asia-Pasic Advanced Network(APAN), 2004.

[10] UI. Partisans, A survey of models for inference of gene regulatory networks, Non-linear Analysis: Modeling and Control Vol.18, No.4, Sep. 25 2013.

[11] Sung- Bae Cho, Hong-Hee , Won, Machine Learning in DNA Microarray Analysis for Cancer Classification.