Data Analysis on DNA Microarray Expression Values using Self Organizing Map

Krishnaveni.S
Research scholar in Computer Science,
Ayya Nadar Janaki Ammal College
Sivakasi

Lawrance.R
Director, Department of Computer Application,
Ayya Nader Janaki Ammal College
Sivakasi

ABSTRACT
There is a vast need to develop analytical attitude to analyze and to make use of the information contained in gene expression data. A narrative approach for cancer prediction (similarity gene) from DNA microarray data, First apply feature selection using parametric and nonparametric feature selection methods to extract features and select exact feature by combining both methods then apply principal component analysis in microarray data to reduce dimensionality Then the selected principal components are clustered and classified using self organizing map and compare the results.

Keywords
Data Mining, Clustering, Microarray data, Feature Selection, SOM.

1. INTRODUCTION
The DNA microarray can be called "DNAChip" or "Gene Chip"[5]. DNA micro array technology allows us to continuously monitor the expression of enormous amount of genes. Micro array promised to revolutionize our understanding of complex diseases. The large number of genes and the complex biological structure greatly increases the challenges of analyzing and interpreting the huge amount of data.

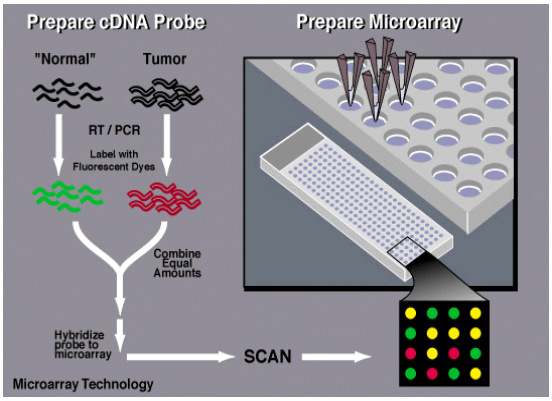

Fig. 1 Microarray

Microarray composes of controlled samples and tested samples. Controlled samples are normal samples and tested samples are diseased samples.

Dimensionality reduction in gene expression data can be critical for a number of reasons. First, for large number of genes or feature set, the processing of all available genes may be computationally infeasible. Second, many of the available features may be redundant and noise-dominated or irrelevant to the classification task at hand. Third, high-dimensionality is also a problem if the number of variables is much larger than the number of data[1] points avail-able. In such a scenario, dimensionality reduction is crucial in order to overcome the curse of dimensionality [6],[7]and allow for meaningful data analysis. For the above reasons, feature selection is important for gene expression data analysis[8].

In this paper it has been applied parametric (ttest) and nonparametric (wilcoxon) are applied to select exact genes then use principal component analysis for dimensionality reduction and apply SOM to cluster selected genes and identified the similar gene hits on same neuron and select significant features.

2. METHODS
2.1 Feature Selection
Rank each feature according to some univariate metric and select the highest ranking features. Filter approaches are used to select the significant features. It is fast and accurate. Parametric test makes inferences about the parameters of the distribution. Nonparametric method not based on parameterized families of probability distribution. Parametric model has a fixed number of parameters; non-parametric model grows the number of parameters with the amount of training data.

2.2 t-test
t-test is a statistical method. t-test compares the mean value of two groups.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where $\bar{X}_1$ is Mean of first set of values, $\bar{X}_2$ is Mean of second set of values, S_1^2 Standard deviation of first set of values S_2^2 Standard deviation of second set of values n_1 Total number of values in first set n_2 Total number of values in second set. t value calculated using t-test method.

t-test value calculated using[1]average value of X1,X2 are in upper part and standard deviation values are used in lower part of the formula. Based on the t-test value, p-values are calculated. Higher t-values or lower p-values are determined the significant features [12].

2.3 Wilcoxon test
The Wilcoxon rank-sum test is a nonparametric method. Wilcoxon test is valid for data from any distribution and less sensitive to outliers. nA + nB observations of the combined

sample. Each observation has a rank: the smallest has rank 1, the 2nd smallest rank 2, and so on[13].

wA -denote the observed rank sum.

WA - sum of the ranks for observations from A.

$\mu A = \frac{nA(nA+nB+1)}{2}$

$\sigma A = \frac{nAnB(nA+nb+1)}{nA}$

$\mathrm{pr}(W_A > wA) \approx pr(Z > z)$

$= \frac{WA-\mu A}{\sigma A} z = \frac{WA-\mu A}{\sigma A}$

μA=115, μA = 15.16575, P= 0.024

Using feature selection methods ttest and wilcoxon 100, informative features are selected from each datasets. The selected top 10 features are listed below.

Table 1.Selected Features (ttest,wilcoxon)

S.No	Dataset	Ttest	Wilcoxon
1	Leukemia	2020, 5772, 4328, 3320, 6281, 1306, 3847, 2354, 2642, 2759	5772, 2354, 6855, 144, 1630,4535, 6281,2233, 1928, 804
2	Carcinoma	3493, 2, 8, 68, 103, 36,4, 107, 142,111	4, 8,9, 78 107,138, 2 15,36, 68

2.4 Principal component analysis

Principal Component Analysis [14] is an unsupervised Feature Reduction method for projecting high dimensional data into a new lower dimensional representation of the data that describes as much of the variance in the data as possible with minimum reconstruction error[7]. Principal component analysis is a statistical technique for determining key variables in a high dimensional data set to lower dimension dataset without loss of information [3]. Covariance matrix based calculations are used to calculate principal components for selected 100 features [4]. Principal components are the inputs of self organizing map.

2.5 Self organizing map

The Self-Organizing Map (SOM) was developed by Kohonen[7]. A self-organizing map (SOM) is an unsupervised neural network which has been successfully used for the analysis and used to interpret the gene expression data[9]. The SOM can be used to select the informative features from the dataset. The SOM algorithm performs mathematical cluster analysis useful for recognition and classification of features in complex, multidimensional data without using prior knowledge[10]. The two-dimensional relationship of the SOM output nodes provide supplementary information beyond the strict classification of a particular input vector[5]. Therefore SOM is a good tool to do quick analysis of multi-variety data because it is accurate to produce good results[6].

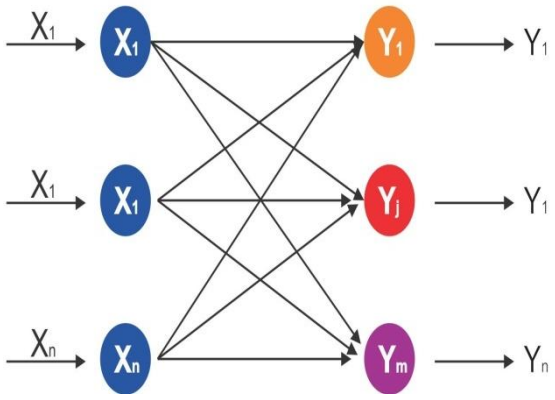

Fig. 2 neural network(SOM)

2.5.1 Algorithm: Self Organizing map

Input:
W, Random weights vectors
D(s), Input Vectors
L, Learning rate

Output:
Clustered gene expression with two dimensional grid

Procedure on SOM

Begin

Step1: *Random weight vector initialization*
Step2: *Input vector with training set*
Step3: *Calculate distance using Euclidean method to select best matching unit with minimum distance*
Step4: *Update the nodes in the neighborhood of the BMU by pulling them closer to the input vector using the following formula*

$W_v(s+1) = W_v(s) + \alpha(s)(D(s) - W_v(s))$

Step5 : Increase s and repeat from step 2 while $s < \lambda$

End

$W(s+1) = w(s) + L(s)(v(s) - w(s))$ (1)

weight updation in self organizing map following formula in 1. Here L(s) is a learning rate. New weight is w(s+1).v(s) is a input vector. w(s) is old weight.

Learning or training is a process by means of which a neural network adapts itself to a stimulus by making proper parameter adjustments, resulting in the production of desired response. The learning rate is denoted by α. it is used to control the amount of weight adjustment at each step of training. The learning rate ranging from 0 to 1, determines the rate of learning at each time step.

3. GENE SELECTION USING SOM

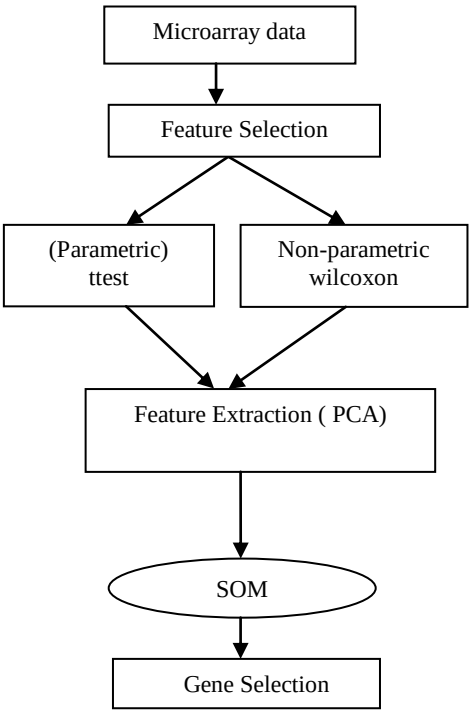

Fig. 3 Analysis of Gene Expression Using SOM

4. EXPERIMENTAL RESULTS

The experimental results by using DNA microarray databases of leukemia, carcinoma cancer are the following

4.1 Carcinoma dataset

Data compose of total 7,457 genes for a total of 36samples. There were 18 cancer samples and 18normal samples [2]. Samples have paired characteristics. The experiment results indicated the numbers of genes selected (top 10) are indicated in Table 1.

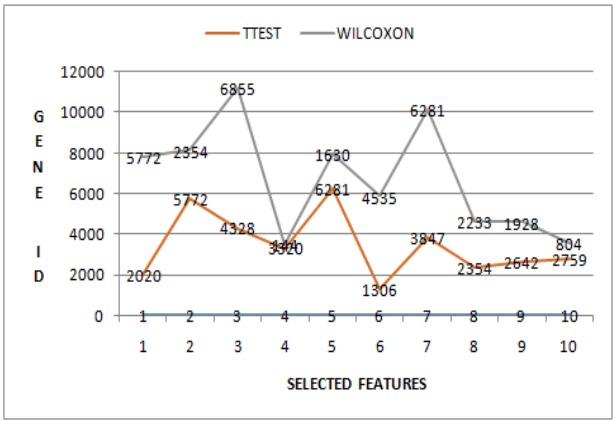

Fig. 4.Compartive analysis of feature selection on leukemia

4.2 Leukemia dataset

Data compose of 7,129 genes for a total of 72samples[2]. There were 47 Actual Lymphoblastic Leukemia (ALL) samples and 25 Actual Myeloid Leukemia (AML) samples [21]. Data are independent from each other (unpaired data).top 10 selected features are listed in Table 1.

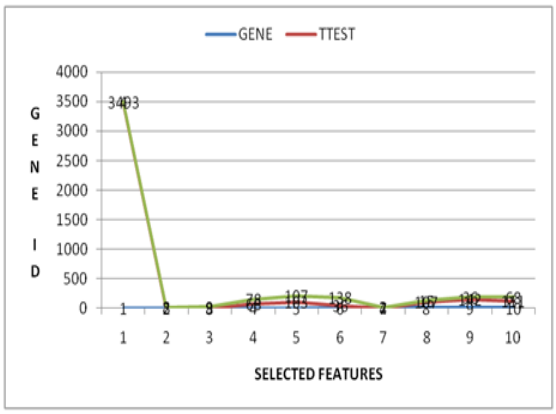

Fig. 5.Compartive analysis of feature selection on carcinoma

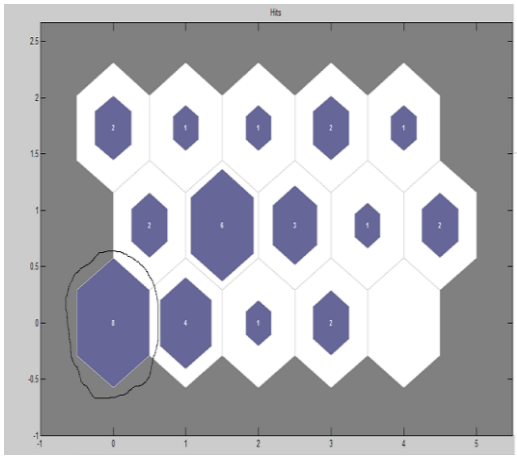

Fig. 6. Neurons hits on som in leukemia dataset

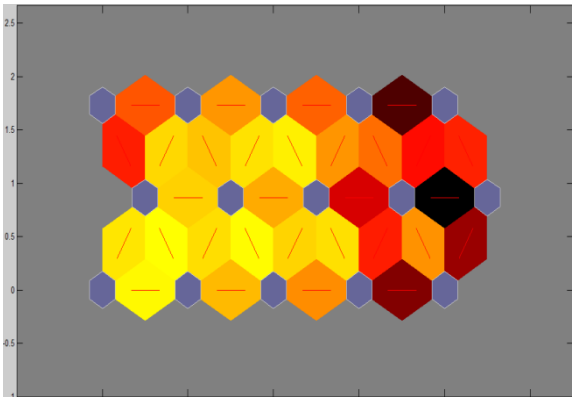

Fig. 7.distance map for leukemia dataset

Table 2 Top five Selected features carcinoma(SOM)

S.No	Selected features	Gene Accession No	Description
1	2	M83670	"Human carbonic anhydrase IV mRNA, complete cds"
2	4	M97496	"Homo sapiens guanylin mRNA, complete cds"

3	8	T965448	ye49f12.s1 Homo sapiens cDNA clone 121103 3' similar to gb:X16940 ACTIN, GAMMA-ENTERIC SMOOTH MUSCLE (HUMAN)
4	36	J02854	Human 20-kDa myosin light chain (MLC-2) mRNA, complete cds
5	107	T64297	"yc48a10.s1 Homo sapiens cDNA clone 83898 3' similar to gb:M10050 FATTY ACID-BINDING PROTEIN, LIVER (HUMAN);"

Table 3. Neuron hits on leukemia data

Index	Hits
(1,1)	2
(2,1)	1
(4,1)	3
(5,1)	1
(6,1)	2
(7,1)	6
(8,1)	1
(10,1)	2
(11,1)	8
(12,1)	4
(13,1)	2
(14,1)	4

5. CONCLUSION

The method stated the various parts of a new approach for the selection of genes by using parametric and nonparametric methods used and taking informative genes clustering by SOM. The method can extract knowledge from the number of more than 7,000 genes in the form of most similar genes from the som neuron hits .The experiment also received a small numbers of genes with 11 and 2 genes, respectively. From the experimental results statistical method and Self-Organizing Map are methods of neural network that is used for gene selection. These techniques show the result of gene selection with the simple and small numbers of iterations. The experiment indicates that the novel approach can be used to predict similar genes with disease nature and can be applied for some other diseases with DNA microarray.

6. REFERENCES

[1] Sathishkumar, E. N., K. Thangavel, and T. Chandrasekhar.”A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction."arXiv preprint arXiv:1306.2118 (2013).

[2] Kent Ridge Bio-medical Data Set. [On-line] Available:http://sdmc.lit.org.sh/GEDatasets/Data sets.html

[3] Kim, H., Ahn, J., Park, C., Yoon, Y., & Park, S. “ICP: A novel approach to predict prognosis of prostate cancer with inner-class clustering of gene expression data.” *Computers in biology and medicine*,2013, vol*43*(10),pp: 1363-1373

[4] Chen, Chien-Hsing. "A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection." Applied Soft Computing 20 (2014): 4-14.

[5] Shieh, Shu-Ling, and I-En Liao. "A new approach for data clustering and visualization using self-organizing maps." Expert Systems with Applications 39.15 (2012): 11924-11933

[6] Tasdemir, Kadim, Pavel Milenov, and Brooke Tapsall. "Topology-based hierarchical clustering of self-organizing maps." *Neural Networks, IEEE Transactions on* 22.3 (2011): 474-485.

[7] Dash, B., et al. "A hybridized K-means clustering approach for high dimensional dataset." International Journal of Engineering, Science and Technology 2.2 (2010): 59-66.

[8] GaneshKumar, Pugalendhi, et al. "Hybrid ant bee algorithm for fuzzy expert system based sample classification." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 11.2 (2014): 347-360.

[9] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 96: 2907–2912.

[10] Lalinka de compos Teixeira Gomes, FernandoJ. Von Zuben, Pablo Moscato, A Proposal forDirect-Ordering Gene Expression Data bySelf-Organising Maps, Elsevier, (2004), pp:11-21.

[11] Aleshunas, John Joseph, Daniel C. St Clair, and William E. Bond. "Classification characteristics of SOM and RT2." Proceedings of the 1994 ACM symposium on Applied computing. ACM, 1994.

[12] Rapaport, Franck, et al. "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data." Genome Biol 14.9 (2013): R95.

[13] Reczko, Martin, et al. "Functional microRNA targets in protein coding sequences." Bioinformatics 28.6 (2012): 771-776.

[14] Kumar.S.,Thangavel.kand Chandrasekhar.T. "A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction",*arXiv preprint arXiv:1306.2118*, 2013.