ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Statistical Measures and Genetic Algorithm for Gene Selection in Breast Cancer

Dr. E.S. Samundeeswari [1], C.Sathya[2]

Associate Professor of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

Research scholar, Dept. of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

ABSTRACT: In recent years breast cancer can be classified by gene expression profiling. Gene expression profiling is measuring the activity of thousands of genes simultaneously. DNA Microarray based gene expressions profiling is used to identify which genes expressions changed due to pathogens or other organisms by comparing gene expression in infected with that of uninfected cells. Microarray is a high dimensional dataset with small sample size, containing relevant as well as irrelevant and redundant genes. Relevant gene selection is an important task in microarray data classification. In this work, top ranking relevant genes are selected using statistical measures such as Two-sample test, Entropy test and Bhattacharyya test. The top ranked genes thus selected are further filtered using Genetic Algorithm with SVM used for fitness evaluation.

KEYWORDS: Microarray gene expression profiling, Two sample t-test, Entropy test, Bhattacharyya Test, Genetic Algorithm, Support Vector Machine.

I. INTRODUCTION

Breast cancer is the most common cancer found in women all over India and accounts for one fourth of all cancers in women in Indian cities. Efficient and accurate prediction of breast cancer based on clinical and histopathological data is difficult, since 90% of cancer occurs due to genetic abnormalities. Gene expression arrays give the information on how the genes are affected by external stimuli, environmental factors, drugs, and drug dosages. DNA microarrays are used in the medical field, forensics, agriculture and toxicology. In medical field, it is used to identify the cause of a disease, disease prediction and disease treatments. Gene expression profiling is also an efficient way to classify the subtypes of breast cancer.

Accurate analysis and classification of gene expression profiles lead to more reliable tumor classification, better prognostic prediction, diagnosis and selection of more appropriate treatments.

Reducing the dimensionality of gene space is important in microarray data classification. Gene space can be reduced by statistical methods. Most informative genes can also be selected by optimization techniques.

II. RELATED WORK

Gene selection plays a vital role in constructing efficient classifiers for microarray data classification. It is the process of identifying a small subset of features that contains the most discriminative information from the original features to improve the classification performance. An effective gene selection method is necessary to remove the irrelevant genes as it increases the classification accuracy.

Common feature selection methods are Filter and Wrapper methods. In Filter method, all features are ranked with their goodness. The top ranked genes are chosen for the classification. In Wrapper method, gene selection technique is combined with a classifier to evaluate classification performance of each gene subset.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Aigco et al[2] proposed a Partial Least Square[PLS] method which is a non-parametric, multivariate statistical analysis technique. PLS ranks all the features by calculating the projection importance of each feature and selects features that are ranked. To avoid the redundancy, the authors combined PLS with recursive elimination which is more time consuming. So Partial Least Square based recursive feature elimination is integrated with Simulated Annealing and Square root to speed up the selection process.

HalaM.Alshamalan[1] found a new method called as Genetic Bee colony algorithm. GBC is an integration of Genetic Algorithm operators with Artificial Bee Colony algorithm, deriving a modified ABC based algorithm for constrained optimization. The GBC algorithm shows superior performance with highest classification accuracy with lowest average number of selected genes. GBC selects fewer genes and gives more classification accuracy.

Thanh Nguyen et al[3] proposed a new approach, Modified Analytic Process. The top ranking genes are selected using different statistical measures like Two-sample test, Entropy test, ROC curve, Wilcoxon method and Signal to Noise Ratio. The modified analytic hierarchy process is then applied on the selected genes and classification accuracy measured.

In this proposed work top ranked genes are selected by Two sample t-test, Entropy test and Bhattacharyya Test. The informative genes from these selected genes are filtered by Genetic Algorithm. The top ranked genes are the initail popualtion for Genetic Algorithm. Support Vector Machine is used as a classifier for evaluating the fitness function. The informative genes from the Genetic Algorithm classify the samples into relapse or non-relapse class.

III. METHODOLOGY

Feature selection methods are used to select the most relevant genes from microarray expression. Various statistical methods are used to select the most relevant genes from gene expression profile are described below.

A. Two sample t-test

It is used to determine differences between gene expressions of relapse and non-relapse cell tissues and finds the top relevant genes from the expression. The large t values represent the important of the genes .

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2 + S_y^2}{n+m}}}$$

where $\bar{x}$, $\bar{y}$ are the means of relapse and non-relapse classes
Sx and Sy are the standard deviations of relapse and non-relapse classes
n and m are the sample sizes of relapse and non-relapse classes

B. Entropy Test

Relative entropy, also known as Kullback-Leibler distance or divergence measures uncertainty associated with a random variable. The entropy value for each gene is calculated with following expression

$$e = \frac{1}{2}\left[\left(\frac{\sigma 1^2}{\sigma 2^2} + \frac{\sigma 2^2}{\sigma 1^2} - 2\right) + \left(\frac{1}{\sigma 1^2} + \frac{1}{\sigma 2^2}\right)(\mu 1 - \mu 2)^2\right]$$

μ1 and μ2 are mean of the relapse and non-relapse classes.
σ1 and σ2 are variance relapse and non-relapse of the class.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

C. Bhattacharyya Test

Bhattacharyya test determines the relative closeness of two samples. The distance between the two classes is calculated by extracting the mean and variance of the two classes

$$D_{B(p,q)} = \frac{1}{4} In\left(\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2\right)\right) + \frac{1}{4}\left(\frac{\left(\mu_p - \mu_q\right)^2}{\sigma_p^2 + \sigma_p^2}\right)$$

μp and σp mean and variance of the relapse class.
μq and σq mean and variance of the non-relapse class.

D. Genetic Algorithm

Genetic algorithm belongs to the class of Evolutionary algorithm. It is used to find the optimal solution from a set of solutions. The solution gets improved over the generations based on the mechanisms of natural selection. An initial population is created consisting of randomly generated rules. Each rule is represented as a string of bits. Left most bits represent attributes. Right most bit represents the class.

Steps in Simple Genetic Algorithm

1. Create random chromosome for initial population.
2. Calculate the fitness of each chromosome in the population.
3. Repeat the following steps until no offspring have been created:
    - 3.1 Select a pair of parent chromosomes from the current population.
    - 3.2 With probability pc, crossover the pair at a randomly chosen point to form two offspring.
    - 3.3 Mutate the two offspring at each locus with probability pm.
4. Replace the current population with new population.
5. Test the stopping criteria.
6. Loop: continue step 2-5 until criteria is satisfied

The basic operators in GA are encoding, selection, recombination and mutation operators.
Encoding is a process of representing individual genes. The process can be performed using bits, numbers, trees, arrays, lists or any other objects. The common way of encoding is binary string with 1s and 0s. Selection is the process of choosing two parents from the population for crossing. The various selection methods are Roulette wheel selection, random selection, Rank selection, Tournament selection and Boltzmann selection. Roulette selection is the traditional GA selection technique. Crossover is the process of taking two parent solutions and producing a child from them. The recombination operation is performed between the selected parents and it produces the new offspring . There are different types of crossover techniques like Single-point crossover, Two-point crossover, Multipoint crossover, Uniform crossover and Three-point crossover. Mutation prevents the algorithm trapped in a local minimum. For binary representation, a simple mutation inverts the value of each gene with a small probability.

E. Support Vector Machine

Support Vector Machine is a promising method for the classification of both linear and nonlinear data. It transforms the original training data into a higher dimension and it separates the dimension by hyperplane. The hyperplane is a decision boundary separating the tuples of one class from another class. The SVM finds the hyperplane with the help of support vectors and margins. The best hyperplane is achieved by maximum marginal hyperplane. Support Vector Machine is regarded as an effective classification method in Microarray gene expression profile cancer classification Carlos et al[4]. The points lying above the separating line represent non-relapse class and the points lying below the line represent relapse class.

IV. ExPERIMENT RESULT ANALYSIS

In the proposed approach, top ranking genes are selected by statistical methods such as Two sample t-test, Entropy test and Bhattacharyya test. The first 50 and 100 ranking genes are selected for classification. Two feature sets consisting of the first 50 and 100 top ranking genes are created. The feature set further filtered using GA-SVM. Fitness value is

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

calculated for all chromosomes and based on the fitness value the chromosomes are selected by Roulette wheel selection method. Single point crossover is applied on selected paired string and it creates new chromosome.

Dataset

The breast cancer microarray gene expression profile is obtained from Kent Ridge Biomedical Data Repository. The dataset contains 24481 genes expression of two classes relapse and non-relapse. The dataset contains training and testing sample. Two separate run are made, the first with training sample of 78 and testing sample of 19 and the second with training sample of 70 and testing sample of 27.

Chromosome Representaion

The chromosome represents information about the solution. Encoding is used to define the structure of chromosome. Traditionally solutions are represented by binary string '0' or '1'. The final gene subset obtained by selecting genes with fitness value.

Fitness Function

The fitness function of the chromosome is measured by SVM technique. The fitness function is

$$\text{Fitness}(x) = \text{Accuracy}(x)$$

obtained from SVM. The classification accuracy of Support Vector Machine is calculated by the formula

$$\frac{\text{number of test samples correctly classified}}{\text{Total number of test samples}} \times 100$$

Table 1 shows the parameter values for the Genetic Algorithm. Two separate runs are made with Chromosome size as 50 and 100. The stopping criteria of Genetic Algorithm fixed as 300. The parent selection is made by Roulette wheel selection method. Single point crossover is applied on the selected parents.

Table 1 GA Parameters And Their Values

Parameter	Value
Chromosome size	50 and 100
Population size	50
Maximum no. of Generations	300
Selection Method	Roulette Wheel
Crossover Type	Single point
Crossover Probability	0.6
Mutation Probability	0.1

Table 2 contains the experimental result of two separate runs made on Breast Cancer Dataset. The top 50 and 100 genes are selected using statistical methods Two sample t-test, Entropy test and Bhattacharyya test. Optimization technique Genetic Algorithm is applied and it selects the most relevant genes and provides maximum classification accuracy.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

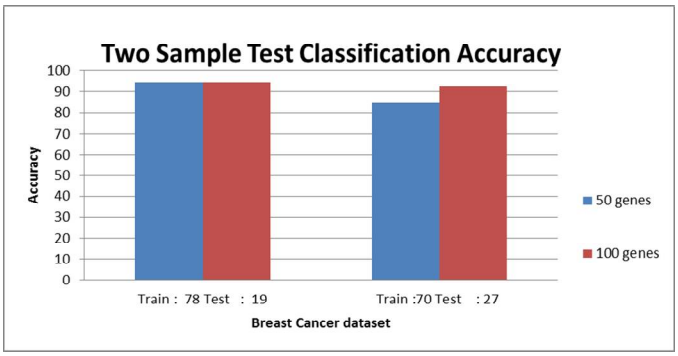
Vol. 5, Issue 3, March 2016

Table 2 Experimental Results

Breast Cancer Dataset	Gene Selection Method	No of Genes	Accuracy
Train Sample : 78 Test Sample : 19	Two Sample Test	50	94.73
		100	94.73
	Entropy	50	94.73
		100	100
	Bhattacharyya	50	89.47
		100	100
Train Sample : 70 Test Sample : 27	Two Sample Test	50	85.18
		100	92.59
	Entropy	50	96.29
		100	88.88
	Bhattacharyya	50	88.88
		100	92.59

The Fig. 1 shows the GA classification accuracy for the gene sets selected by Two-sample t-test. It is found that Two-sample t-test performed well when the training & testing samples are 70 and 27 respectively and for a gene set of 100.

Fig. 1 Two sample t-test classification Accuracy



The Fig.2 shows classification accuracy for the combined GA and Entropy test. The accuracy of classification varies irrespective of sample size and training environment.

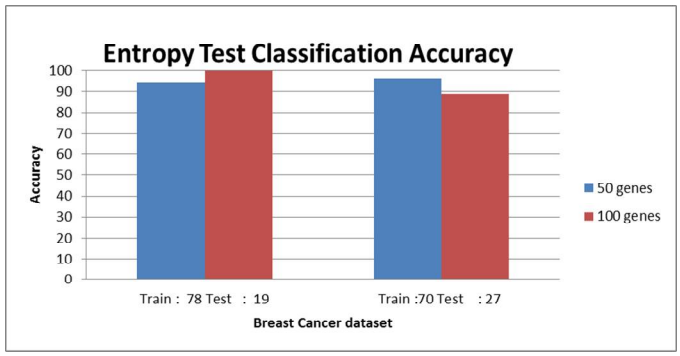
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

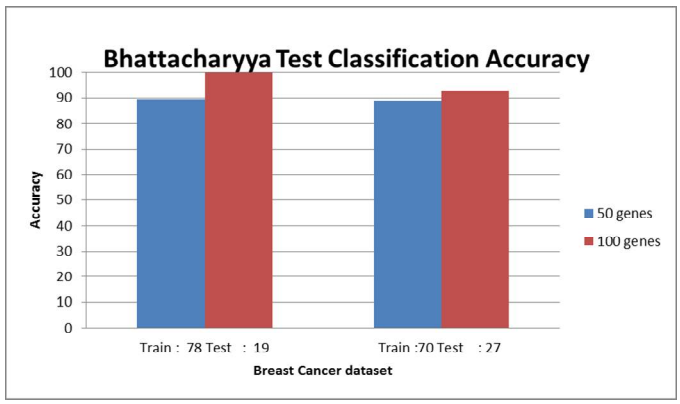
Vol. 5, Issue 3, March 2016

Fig. 2 Classification Accuracy of GA for the input sets selected by Entropy Test



The Fig. 3 shows that Bhattacharyya test performed well for a gne set of 100 in both the training environment. The performance of all could not be generalized. But based on the no.of genes selected by GA, Bhattacharyya test seems to perform well than the other two tests.

Fig. 3 Classification Accuracy of GA for the input sets selected by Bhattacharyya test



It is suggested that the experiments should be carried out rigoursely for more datasets and with other optimization techniques.

V. CONCLUSION

Microarray gene expression data is used to identify the breast cancer biomarkers clearly for relapse and non-relapse classes. Two sample t-test, Entropy test and Bhattacharyya test select the top ranking genes. Genetic algorithm with SVM classifier is applied on top ranking genes. Genetic algorithm selects most informative genes and SVM classification accuracy is considered as the fitness function. This simple model based on various statistical measures and genetic algorithm finds most relevant biomarkers for breast cancer classes.

VI. ACKNOWLEDGEMENT

This paper was funded by UGC-MRP under grant- No.4937/14(SERO/UGC). The authors, therefore acknowledge with thanks UGC, India.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

REFERENCES

[1] Hala M.Alshamalan, GhadaH.Badr, Yousef A.Alohali, "Genetic Bee Colony algorithm: A new gene selection method for microarray cancer classification", Computational Biology and Chemistry, Vol.56, pp. 49-60, 2015.

[2] Aigco Wang,Ning An,Guilin Chen,Lian Li,Gil Alterovitz, "Improving PLS-RFE based gene selection for microarray data classification", Computers in Biology and Medicine, Vol.62, pp.14-24 ,2015.

[3] Thanh Nguyen , Abas Khosravi, Douglas Creighton, Saeid Nahavandi, "A Novel aggregate gene selection method for microarray data classification", Elsevier, Pattern Recognition Letters, Vol. 60-61, pp. 16-23, 2015.

[4] Carlos J. Alonso-Gonzalez, Q.Isaac Moro-Sancho,Arancha Simon-Hurtado, Ricardo Varela-Arrabal , "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification method", Expert Systems with Applications, vol. 39, pp. 7270-7280, 2012.

[5] Chellamuthu Gunavathi and KandasamyPremalatha, "A Comparative Analysis of Swarm Intelligence Techniques for Feature Selection in Cancer Classification", The Scientific World Journal Volume, Hindawai Public Corporation , vol. 2014 , pp. 12 ,Article ID 693831, 2014.

[6] Li-Yeh Chuang et al, "Gene selection and classification using Taguchi chaotic binary particle swarm optimization", Expert Systems with Applications ,Vol.38, pp. 13367-13377, 2011.

[7] Cruz, Joseph A., David S., Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", Cancer Informatics, Vol.2, 59-77, 2006.

[8] S.N.Sivanandam, S.N.Deepa, "Principles of Soft Computing", Second Edition, Wiley India Pvt Ltd. 2011.

[9] Sumeet Dua, Pradeep Chowriappa, "Data Mining for Bioinformatics", Indian Edition, CRC press Taylor and Francis Group. 2013.