# Microarray gene expression ranking with Z-score for Cancer Classification

M .Yasodha ,
Research Scholar
Government Arts College ,
Coimbatore, Tamil Nadu, India

Dr P Ponmuthuramalingam
Head and Associate Professor
Government Arts College ,
Coimbatore, Tamil Nadu, India

***Abstract***--Over the past few decade there has been explosion in the amount of genomic data available to biomedical engineer due to the advantage of biotechnology. For example using microarray it is possible to find out a persons gene expressions profile more than 30000 genomes. Among this one of the most important gene selection problem is gene ranking. Here we will describe Z-score ranking for microarray gene expression selection. In that technique it choose the gene and then applied the Z-Score Ranking technique and then divides the genes into subsets with Successive Feature selection and then finally LDA Applied for the result. With this Z-score ranking technique we will get the accurate results and less effort. The Lymphoma and Leukemia dataset genes are utilized. The proposed technique shows capable classification accuracy for the whole test data sets.

*****

## 1. INTRODUCTION

Cancer has become the leading cause of death worldwide. The most effective way to reduce cancer deaths is to detect it earlier. Though cancer research is generally clinical and biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques have to be applied. Prevention of tobacco-related and cervical cancers and earlier detection of treatable cancers would reduce cancer deaths in India. The absolute number of cancer deaths in India is projected to increase due to population growth and increasing life expectancy.

Previously, cancer classification has always been morphological, and clinical based. This work ranked the whole set of 7129 genes according to their Z-scores (ZSs) in the training data set. Then, take out the top 100 genes with the highest ZSs. The rest of this work ordered as follow: Section 2 offers the environment material about microarray gene expression outline. Section 3 demonstrate and classify the major advances that have been utilized newly for cancer microarray gene expression profile. Section 4, offers discussion and study about the large amount of capable approaches that are offered through out the work. Lastly, Section 5 concludes the work.

## 2. PROPOSED METHODOLOGY

The proposed method is containing of two steps. A rank of every genes in the training data set using a scoring method. Then, retain the genes with high scores. The performance of t-test can be done to minimize the size of the gene and to select the top 100 genes.

### 2.1 Gene Importance Ranking

In this work, compute the significance ranking of every gene using a feature ranking measure, two of which are described below.

***A) Z-Score Methodology***

The classic method for "standardizing" a set of values (finding a common metric or scale) is calculation of z-scores. Instead of translating data to a fixed range as with percentile rescaling, z-scores are anchored by the mean and standard deviation of the original values, and rescaled such that the new mean is 0 and the new standard deviation is 1. The resulting z-scores correspond to points on the standard normal curve, with a theoretical range of approximately -3 to +3. The actual range for each indicator, however, will be different.

Z-scores can be calculated for individual level or aggregate level data. In either case, each value is treated as an "individual" member of a sample. The sample size, then, is equal to the total number of "individuals". When z-scores are calculated for aggregate data, such as county rates and percents, the county is the "individual" (unit of analysis), the group of counties is the sample, and the sample size is therefore the total number of counties. The formula for calculating z-scores is as follows:

**Individual Level Data**

$$z_i = \frac{X_i - \overline{X}}{\sqrt{\frac{\sum_{1-1}^{n}(X_i - \overline{X})^2}{n-1}}}$$

where $X_i$ = the original value for individual i

and $\overline{X}$ = the mean of the distributi on of individual s

and n = the total number of individual s

 ISSN: 2321-8169
2480 – 2484

Aggregate Level Data

$$z_i = \frac{p_i - \bar{p}}{\sqrt{\frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{(n-1)}}}$$

where $p_i$ = the original proportion for county i (or region, census tract, etc.)

and $\bar{p}$ = the mean of the distributi on of counties (or regions, census tracts, etc.)

and n = the total number of counties (or regions, census tracts, etc.)

There are K classes. $\max\{y_k, k = 1,2, \dots K\}$ is the maximum of all $y_k$. $C_k$ refers to class k that contains $n_k$ samples. $x_{ij}$ is the expression rate of gene *i* in sample *j*. $\bar{x}_{ik}$ is the mean expression rate in class *k* for gene *i*. *n* is the whole amount of samples. $\bar{x}$ is the common mean expression value for gene *i*. $s_i$ is the joint within-class standard deviation for gene *i*. In fact, the TS utilized here is a *t-statistic* among the centroid of an exact class and the overall centroid of all the classes. An additional probable model for TS might be a *t-statistic* between the centroid of an exact class and the centroid of all the additional classes. To find the minimum gene subset when after selecting some top genes in the importance ranking list, this attempt to classify the data set with only one gene. This work inputs each selected gene into LDA classifier.

**3.2 Successive Feature Selection**

Successive Feature Selection SFS method (SFS) a set of $x \geq 10$ features is procedure single at a time that the rate of x is taken due to memory constraint and it is experimentally establish that the fitting values of x is equal to or lower than 10. The output is the grade of features. In the successive stage that the feature is reduced once at a time and a subset of features is achieved. That the classification accuracy using classifiers calculate, and the top subset of features is processed to the next level. There might be further than one top subset of features in a given stage. A feature is dropped in level 1 that offers four dissimilar subsets of features. The top set in level 1 {x, x2, x4} is which is chosen for level 2. In a related way a feature is dropped from the best set of features of level 1 into level 2, which provides three different subsets of features. The best sets in level 2 are {x2, x4} and {x1, x2} supposing that their classification accuracies are the similar and are elevated than those of other subsets and the best set in level 3 is {x2} .

This procedure is finished when all the features are ranked. Two ranked sets are achieved in SFS: that is R1= {x2, x4, x1, x3} and R2= {x2, x1, x4, x3}

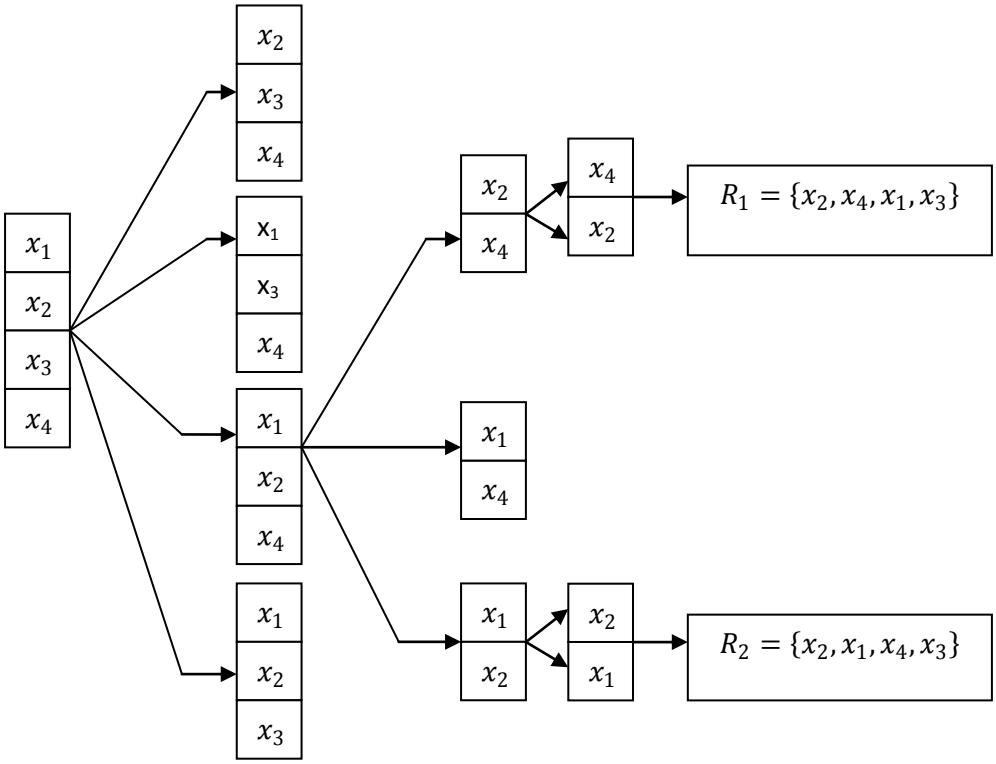


**Figure 1 Successive Feature Selection**

### 3.3 Block Reduction

A d-dimensional feature vector has been partitioned into m roughly equal blocks, $S_j$, for $j = 1 \ldots m$ of size h _ 10. Each block has at least r features. All the blocks have been processed through the SFS procedure one at a time, which yields top-r feature sets, Fj, for j ¼ 1 . . . q. Then, the unique features of two consecutive feature sets, F1 and F2, are used to find the best top-r feature set, Fb. Next, the unique features of Fb and F3 are used to obtain the best set. This process is continued for all the q sets. The obtained best top-r feature set, Fb, from the block reduction procedure is stored for further pruning.

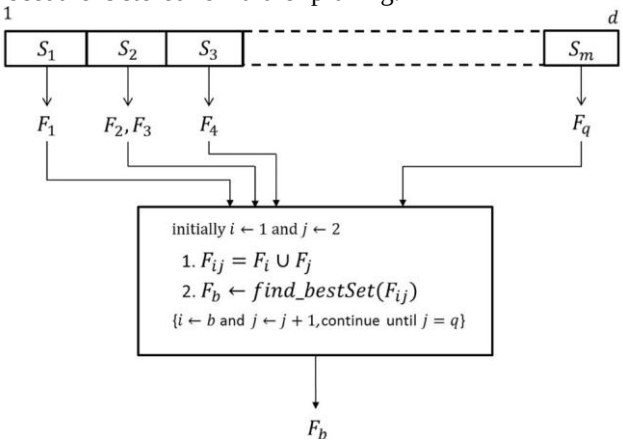


**Figure 2 Block Reduction**

1. Select the *r* number of features to be investigated, where 1< *r* < *h*, and select the block size *h*, where *h* ≤10.
2. Decompose the training samples randomly into a training set *(Tr)* and a validation set (*V)* using a proportionality ratio $p^1$.
3. Partition the features of the sets (*Tr* and *V*) into *m* roughly equal blocks, $S_j$, for $j$ =1…*m*.
4. Apply the Successive Feature Selection (SFS) procedure on each of $S_j$ to get the Z-Score ranked feature set, $F_j$ and its corresponding classification accuracy, $\alpha_j$, for $j$=1…*q*, where $q \geq m$ and $F_j \neq F_l \; \forall \; j \neq l$,
5. Initialize $i \leftarrow 1$ and $j \leftarrow 2$.
6. Find the best features set $F_b \leftarrow find\_bestSet(F_i, F_j)$
7. Terminate the process if $j = q$, or else update $i \leftarrow b$ and $j \leftarrow$j + 1, and go to Step 6.
8. If more than one set of $F_b$ is obtained, then perform cross-validation to get one best set (for cross-validation, decompose training samples randomly *n* times[2] into training sets and validation sets using the proportionality ratio *p* and compute the average classification accuracy for all sets in $F_b$; select a set if $F_b$ for which the average classification accuracy is the highest)
9. Repeat Steps 2-8 for another random decomposition of training samples. Let the new training set and validation set be defined as *Tr** and *V**. This will give a best set $F_b^*$.
10. Find the best set and its corresponding average classification accuracy ($\alpha_b$) using $F_b$ and $F_b^*$; i.e., [$F_b$,$\alpha_b$]← find_set &alpha ($F_b$,$F_b$*)
11. Repeat Steps 9-10 until $\alpha_b$ does not show any improvement.

The dimensionality of the feature space is reduced either through feature selection or through feature extraction. Linear Discriminant Analysis (LDA) is a well-known technique for feature selection-based dimensionality reduction.

## 3. EXPERIMENTAL RESULTS

The innovative of the work reason to discover the ranking gene with correct cancer classifications for this LDA classification is chosen, it is an adequately good classifiers. The planned methodology was useful to the openly obtainable cancer datasets namely Lymphoma and Leukemia cancer dataset and the experimented using MATLAB.

***(i) Lymphoma dataset***

Lymphoma data set holds 42 samples resultant from diffuse large B-cell lymphoma (DLBCL) and 9 samples from Follicular Lymphoma (FL) later than 11 samples from Chronic Lymphocytic Leukaemia (CLL). The whole data set hold 4026 genes. In this data set, a little part of data is absent.

***(ii) Leukemia dataset***

The leukemia data set holds expression levels of 7129 genes taken over 72 samples. Labels point out that which of two variants of leukemia is present in the sample. This dataset is of the similar type as the colon cancer dataset and can consequently be used for the similar kind of experiments.

**Table 1 Dataset used in the Experiment**

| Dataset | Class | Number of Gene | Training samples | Test samples |
|---|---|---|---|---|
| Lymphoma | 3 | 4026 | 43 | 19 |
| Leukemia | 2 | 7129 | 40 | 19 |

 ISSN: 2321-8169
2480 – 2484

**Table 2 Accuracy and Execution Time for proposed method**

| Dataset | Methods | Number of Selected Genes | Accuracy (%) | Execution Time (Seconds) |
|---|---|---|---|---|
| Lymphoma | Existing Feature Selection with LDA | 250 | 89 | 10 |
| | Proposed Feature Selection with LDA | 20 | 97 | 8 |
| Leukemia | Existing Feature Selection with LDA | 150 | 83 | 14 |
| | Proposed Feature Selection with LDA | 25 | 95 | 11 |

Table 2 shows the accuracy and execution time for feature selection in gene expression data. Figure 3 shows the comparison of accuracy for the proposed method of feature selection with LDA classification and the Existing method of feature selection with LDA, from the table obviously noticed that the planned technique provides improved results by their correctness in percentage.

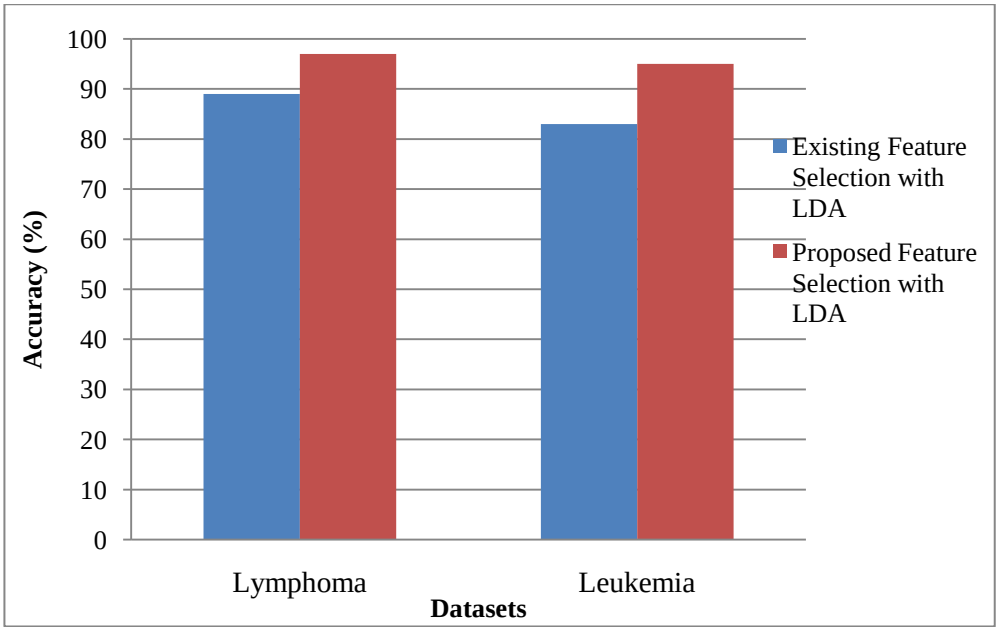


**Figure 3 The accuracy for Proposed feature selection methods**

Figure 4 shows the comparison of execution time in seconds for the existing feature selection with LDA classification and the proposed feature selection with LDA used by the Lymphoma dataset and Leukemia dataset. By the similarities clearly noticed that the proposed feature selection with LDA classification construct the improved outcome in the reduced time.

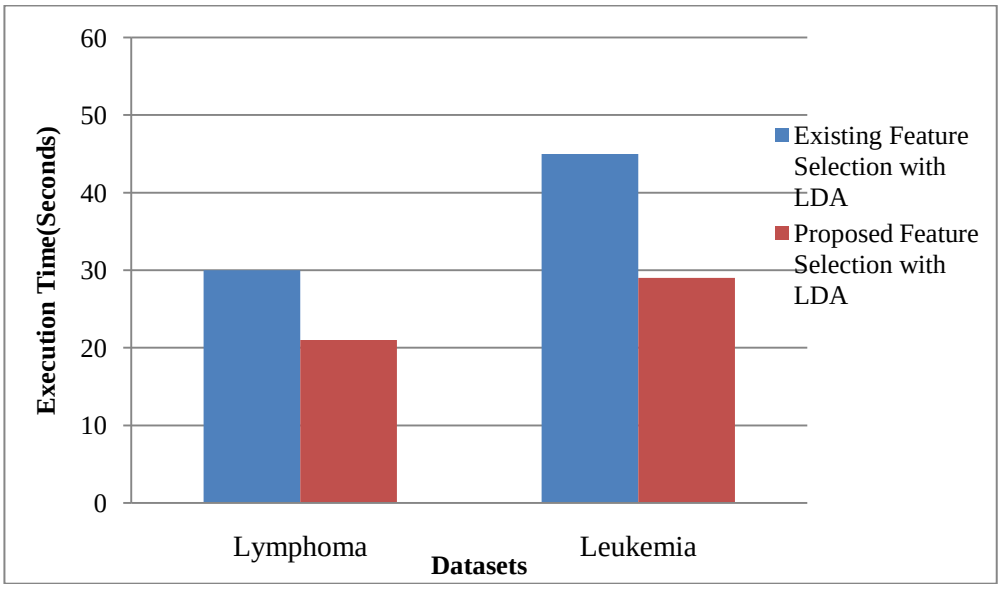


Fig4 shows the execution time for both Existing and Proposed feature selection methods

## 4. CONCLUSION

Cancer is one of the significant characteristic in the biomedicine field. Exact calculation of numerous tumor kinds has higher value in offering improved treatment and toxicity decrease on the patients. In the history, cancer classification is usually depends on morphological and clinical analysis. These preceding cancer classification methods are confirmed to have numerous drawbacks in their analytical potential. To overcome those disadvantages in cancer classification, capable technique in agreement with the global gene expression assessment have been evolved. This gene data must be preprocessed for classification with good correctness using the classifier. The gene ranking technique is utilized to preserve that task. This effort uses enhancement score for ranking the gene. Then the classifier is educated with that data. Finally, the classification of gene for identifying the cancer is carried out. This proposed technique is used to choose the top genes.

## REFERENCES

[1] Chuang, Li-Yeh, Cheng-Huei Yang, Kuo-Chuan Wu, and Cheng-Hong Yang (2011), "A hybrid feature selection method for DNA microarray data." Computers in biology and medicine 41, no. 4: 228-237.

[2] Nam, Dougu, and Seon-Young Kim(2008), "Gene-set approach for expression pattern analysis." Briefings in bioinformatics 9, no. 3: 189-197.

[3] Guyon, Isabelle, and André Elisseeff (2003), "An introduction to variable and feature selection." The Journal of Machine Learning Research 3: 1157-1182.

[4] Deutsch, J. M. "Evolutionary algorithms for finding optimal gene sets in microarray prediction." Bioinformatics 19, no. 1 (2003): 45-52.

[5] Xiong, Momiao, Wuju Li, Jinying Zhao, Li Jin, and Eric Boerwinkle (2001), "Feature (gene) selection in gene expression-based tumor classification." Molecular Genetics and Metabolism 73, no. 3: 239-247.

[6] Yu, Lei, and Huan Liu (2004), "Redundancy based feature selection for microarray data." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 737-742. ACM.

[7] Shevade, Shirish Krishnaj, and S. Sathiya Keerthi (2003), "A simple and efficient algorithm for gene selection using sparse logistic regression." Bioinformatics19, no. 17: 2246-2253.

[8] Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu (2002), "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proceedings of the National Academy of Sciences 99, no. 10: 6567-6572.

[9] Chuang, Li-Yeh, Hsueh-Wei Chang, Chung-Jui Tu, and Cheng-Hong Yang. "Improved binary PSO for feature selection using gene expression data."*Computational Biology and Chemistry* 32, no. 1 (2008): 29-38.

[10] Zhang, Hao Helen, Jeongyoun Ahn, Xiaodong Lin, and Cheolwoo Park. "Gene selection using support vector machines with non-convex penalty."*Bioinformatics* 22, no. 1 (2006): 88-95.

[11] Setiono, Rudy. "Generating concise and accurate classification rules for breast cancer diagnosis." *Artificial Intelligence in medicine* 18, no. 3 (2000): 205-219.

[12] Chu, Feng, Wei Xie, and Lipo Wang. "Gene selection and cancer classification using a fuzzy neural network." In *Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the*, vol. 2, pp. 555-559. IEEE, 2004.

[13] Kim, Hyunsoo, Gene H. Golub, and Haesun Park. "Missing value estimation for DNA microarray gene expression data: local least squares imputation."*Bioinformatics* 21, no. 2 (2005): 187-198.

[14] Bolón-Canedo, Verónica, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. "A review of feature selection methods on synthetic data."*Knowledge and information systems* 34, no. 3 (2013): 483-519.

[15] Sharma, Alok, Seiya Imoto, and Satoru Miyano. "A top-r feature selection algorithm for microarray gene expression data." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9, no. 3 (2012): 754-764.