VOL. 10, NO. 4, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

A FAST AND EFFICIENT FEATURE SELECTION ALGORITHM FOR MICROARRAY GENE EXPRESSION AND CANCER CLASSIFICATION

M. Yasodha and P. Ponmuthuramalingam
Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India
E-Mail: vmyasodha@gmail.com

ABSTRACT

Cancer research is one of the most important research areas in the medical field. The most vital process for gene is identification and classification of cancer. The significance of the every gene is to be initiated by the gene raking measurement. Gene expression summary by microarray method has been efficiently use for classification and analytical guess of cancer nodules. Gene ranking technique which use mostly is T-Score. Genes are collected from the dataset. Amount of feature selection algorithm may form mistake for their ranked gene appearance. To avoid this, proposed technique makes the improved precision by producing a feature selection algorithm in gene expression data investigation of model classifications. That the planned technique choose the gene and divides the genes into subset, from the features, gene ranks are chosen. The Lymphoma and Leukemia dataset genes are utilized. The proposed technique shows capable classification accuracy for the whole test data sets.

Keywords: DNA, Gene important ranking, T-score, feature selection, linear discriminant analysis, fuzzy neural network.

1. INTRODUCTION

The gene expression profiles that are achieved from particular microarray testing have been extensively used for cancer classification to construct a successful model. This form can distinguish normal or dissimilar cancerous states by using selected useful genes. The judgment of difficult genetic diseases like cancer has usually been completed based on the non-molecular characteristics like variety of tumor tissue, pathological individuality and clinical phase. DNA microarray techniques have disturbed huge concentration in the area of scientific and in industrial. Several examinations have been offered on the convention of microarray gene expression examination for molecular classification of cancer. An application field where these techniques are probable to make key contributions is the classification of cancers depends on clinical stage and biological behavior. Such classifications contain a vast contribution on diagnosis and management. Normally, a classifier for this reason must deal with the following troubles:

- The classifier should present an easy-to understand measure of declaration for its decision. Thus, the last diagnosis rests with the medical specialist who assess if the assurance of the classifier is extremely enough.

- The classifier should judge asymmetrical incorrect classification costs for false optimistic and false negative classifications.

This work ranked the whole set of 4,026 genes according to their t-scores (TSs) in the training data set. Then, take out the top 100 genes with the highest TSs. 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. RELATED WORKS

Nam *et al.* (2008) developed gene set investigation technique calculate differential expression patterns of gene groups as a substitute of those of personality genes. This advance particularly targets gene groups whose ingredient show subtle but coordinated expression changes, which may not be distinguished by the normal individual gene analysis. Chuang *et al.* (2011) shows gene expression profiles, which correspond to the state of a cell at a molecular level, have great possible as a medical diagnosis tool. In cancer classification, obtainable training data sets are normally of a fairly small example size compared to the number of genes involved. Feature (gene) selection can be used to productively take out those genes that straight influence classification accuracy and to reduce genes which have no influence on it.

Variable and feature selection have turn into the focus of a great deal research in region of request used for datasets with tens or hundreds of thousands of variables are obtainable. These areas include gene expression collection investigation, and combinatorial chemistry. The purpose of changeable selection is three-fold: improving the calculation presentation of the predictors, offering quicker and additional cost-effective predictors, and offering an improved accepting of the fundamental procedure that created the data by Guyon *et al.* (2003). Microarray data has been shown by Deutsch (2003) it to be efficacious in individual closely connected cell types that often appear in different forms of cancer, but is not yet realistic clinically. Gene expression profiles may present additional information than morphology and

present an alternate to morphology-based tumor categorization systems. Gene selection engage a search for gene subsets that are capable to distinguish tumor tissue from normal tissue, and might have either clear biological understanding or a few inference in the molecular machinery of the tumor genesis. Gene collection is an essential problem in gene expression-based cancer categorization. In the pattern of a discriminate rule, the amount of genes is great relative to the amount of tissue samples. Large genes can harm the presentation of the tumor classification system and enlarge the cost as well. In this report, they talk about criteria and illustrate techniques for dropping the amount of genes and choose a best (or near optimal) subset of genes from an original set of genes for tumor categorization. The realistic advantages of gene selection over additional technique of reducing the dimensionality and amount of genes are given by Xiong *et al.* (2001).

Yu *et al.* (2004) shows established gene selection method frequently choose the top-ranked genes according to their individual discriminative control without conduct the elevated degree of redundancy among the genes. Newest research shows that removing unnecessary genes among selected ones attain an improved demonstration of the individuality of the targeted phenotypes and direct to enhance classification accuracy. The proposed algorithm is asymptotically convergent. It is easy and enormously easy to execute; it neither utilize some complicated mathematical programming software nor wants any matrix procedure. It can be applied to a multiplicity of real-world troubles like classify marker genes and construction a classifier in the situation of cancer analysis using microarray data are given by Shevade *et al.* (2003). Tibshirani *et al.* (2002) devised an advance to cancer class prediction from gene expression profile, based on an improvement of the simple nearest prototype (centroid) classifier. This technique of "Nearest Shrunken Centroids" makes out subsets of genes that greatest distinguish each class. The method is general and can be utilized in various other classification problems.

Chuang *et al.* (2008) show gene expression profiles, which characterize the condition of a cell at a molecular level, contain huge possible as a medical diagnosis tool. Compare to the amount of genes involved, obtainable training data sets normally contain a fairly little sample dimension in cancer type classification. A reliable selection method for genes is used to speed up the dispensation rate, diminish the analytical error rate, and to eliminate incomprehensibility. The robust and correct gene chosen technique are necessary to recognize differentially expressed group of genes across dissimilar samples, e.g. among cancerous and normal cells. Successful gene selection will help to categorize dissimilar cancer types, guide to an enhanced understanding of genetic signatures in cancers and get better treatment approaches are given by Zhang *et al.* (2006).

In preceding work was given by Setiono and Rudy (2000), they have obtainable an algorithm that take out classification rules from trained neural networks and talk about its application to breast cancer diagnosis. In this work, they explain how the correctness of the networks and the correctness of the rules extracted from them can be enhanced by a easy pre-processing of the data. Chu *et al.* (2004) use a t- test-based feature selection technique to select some significant genes from thousands of genes. After that, they categorize the microarray data sets with a Fuzzy Neural Network (FNN) that they planned earlier. This FNN merge significant features of original fuzzy model self-generation, restriction optimization, and rule-base simplification. They useful this FNN to three well-known gene manifestation data sets, i.e., the lymphoma data set, little round blue cell tumor (SRBCT) data, and the liver cancer data set.

Kim *et al.* (2005), offer this work to attribution methods based on the least squares formulation are planned to approximation missing values in the gene expression data, which develop local parallel structures in the data as well as least squares optimization procedure. Bolón-Canedo *et al.* (2013), numerous synthetic datasets are working for this reason, aiming at reviewing the presentation of feature selection technique in the attendance of a crescent amount or unrelated features, noise in the data, redundancy and communication among attributes, as well as a little ratio between number of samples and amount of features.

3. PROPOSED METHODOLOGY

The proposed method is containing of two steps. A rank of every gene 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.

3.1 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) T-Test

The t-score (TS) of gene *i* is defined as follows:

$$TS_i = \max_k \left\{ \left| \frac{\bar{x}_{ik} - \bar{x}_i}{m_k s_i} \right|, k = 1, 2, \ldots, K \right\} \quad (1)$$

Where

$$\bar{x}_{ik} = \sum_{j \in G_k} \frac{x_{ij}}{n_k} \quad (2)$$

$$\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n} \quad (3)$$

$$s_i^2 = \frac{1}{n-K} \sum_{k} \sum_{j \in G_k} (x_{ij} - \bar{x}_{ik})^2 \quad (4)$$

$$m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}} \quad (5)$$

There are K classes. $\max\{y_k, k = 1,2,\ldots K\}$ 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}

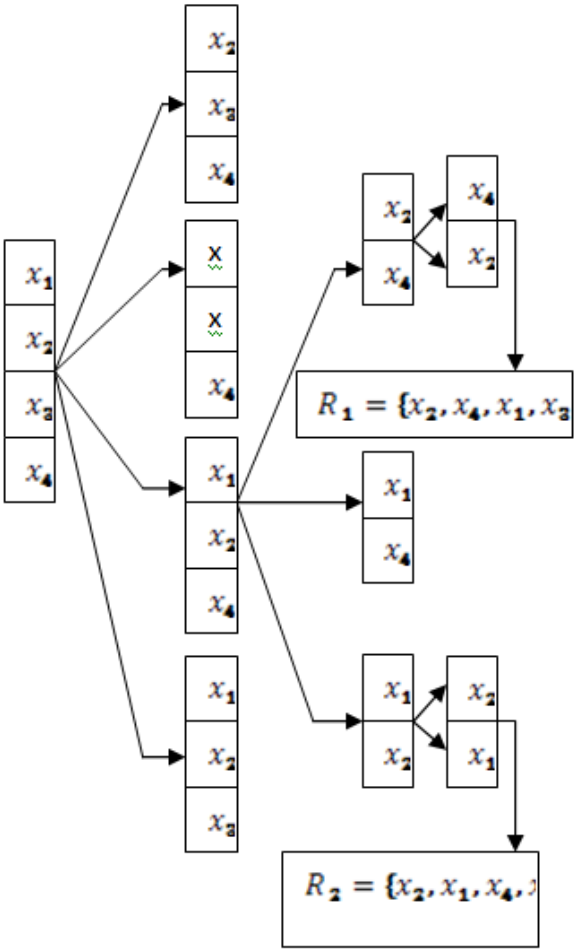

Figure-1. Successive feature selection.

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.

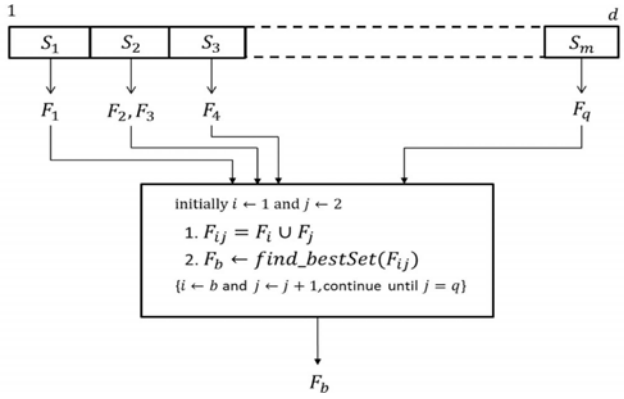

Figure-2. Block reduction.

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 top-*r* ranked feature set, F_j and its corresponding classification accuracy, α_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 (α_b) using F_b and F_b^*; i.e., $[F_b, \alpha_b] \leftarrow$ find_set &alpha (F_b,F_b*)

Repeat Steps 9-10 until α_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.

4. EXPERIMENT AND ANALYSIS

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 in use over 72 models. Labels point out that which of two variant of leukemia is present in the model. 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	No. of Gene	Training samples	Test samples
Lymphoma	3	4026	44	21
Leukemia	2	7129	40	19

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	79	10
	Proposed feature selection with LDA	20	87	8
Leukemia	Existing feature selection with LDA	150	73	14
	Proposed feature selection with LDA	25	85	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.

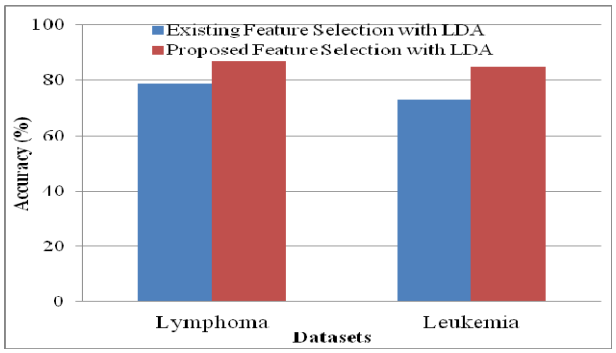

Figure-3. The accuracy for proposed feature selection methods.

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.

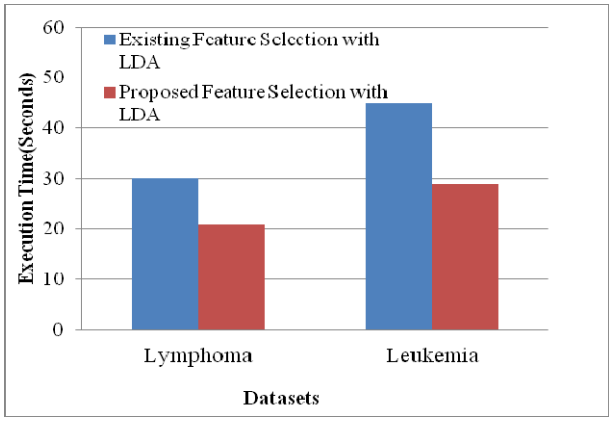

Figure-4. The execution time for proposed feature selection methods.

5. CONCLUSIONS

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 have to be preprocessed for categorization with excellent correctness using the classifier. The gene ranking method is utilized to preserve that task. This effort uses enhancement achieve for ranking the gene. Then the classifier is educated with that data. Lastly, 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(4): 228-237.

[2] Nam Dougu. and Seon-Young Kim 2008. Gene-set approach for expression pattern analysis. Briefings in bioinformatics. 9(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. 2003. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics. 19(1): 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(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. Bioinformatics. 19(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(10): 6567-6572.

[9] Chuang Li-Yeh., Hsueh-Wei Chang., Chung-Jui Tu., and Cheng-Hong Yang. 2008. Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry. 32(1): 29-38.

[10] Zhang, Hao Helen, Jeongyoun Ahn, Xiaodong Lin, and Cheolwoo Park. 2006. Gene selection using support vector machines with non-convex penalty. Bioinformatics. 22(1): 88-95.

[11] Setiono Rudy. 2000. Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in medicine. 18(3): 205-219.

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

[13] Kim Hyunsoo., Gene H. Golub. and Haesun Park. 2005. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 21(2): 187-198.

[14] Bolón-Canedo Verónica., Noelia Sánchez-Maroño. and Amparo Alonso-Betanzos. 2013. A review of feature selection methods on synthetic data. Knowledge and information systems. 34(3): 483-519.

[15] Sharma Alok., Seiya Imoto. and Satoru Miyano. 2012. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 9(3): 754-764.