Original Research Article **Allied science**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

SNR-TR GENE RANKING METHOD: A SIGNAL-TO-NOISE RATIO BASED GENE SELECTION ALGORITHM USING TRACE RATIO FOR GENE EXPRESSION DATA

SHRUTI MISHRA*[1] AND DEBAHUTI MISHRA[1]

[1]Department of Computer Science & Engineering, ITER, S'O'A University, Bhubaneswar, Odisha

ABSTRACT

In this paper, we propose a hybridization of two popular gene selection methods, i.e. Trace Ratio (TR) algorithm and Signal-to-Noise Ratio (SNR) method. The proposed model defines two basic phases: in the first phase, the gene would be ranked using TR algorithm where the scoring weight matrix is re-defined using the SNR scoring technique (known as SNR-TR gene ranking method); and the second phase describes the validation of the top ranked genes using two variants of Neural Network (NN) classifier called resilient propagation and back propagation. It was observed that the SNR-TR gene ranking method found a subset of informative genes from the huge data of available genes and it is acknowledged by the maximum accuracy with less number of iterations obtained as compared to the existing TR algorithm. Further, our model have been experimentally analyzed using five benchmark datasets like Colon, Leukemia, Medulloblastoma, Lymphoma and Prostate Cancer. *K* number of top ranked genes were extracted using SNR-TR algorithm where *K* is defined as 50, 100 and 150. Though the technique is validated using different types of classifier, it can still be validated by using various performance index metrics like Balanced Classification Rate, Balanced Error Rate, Stability index etc.

KEYWORDS: **Gene Regulatory Network, Gene Selection, Trace Ratio Algorithm, Signal-to-Noise Ratio, Classification**

SHRUTI MISHRA
Department of Computer Science & Engineering, ITER, S'O'A University, Bhubaneswar, Odisha

INTRODUCTION

Gene Regulatory Network (GRN) plays a major role in calculating the biological process and molecular organization of existing beings.[1-3] Only the major fear that always exists is, modelling of the networks. Several computational approaches have been sprung up for accessing the gene expression and finding their regulator network and components.[4] Though in the current scenario, developing the GRN model is a major challenge in the biological research.[5] GRN models provide an insight view about the observational facts of the data or the gene, their interaction patterns between the data and ultimately factors affecting the interactions. In other words, the model allows us to deliver an overall dynamic behaviour of the network. In order to develop a GRN model, a reverse engineering path can be followed where instead of creating the network from the whole genome sequence or gene expression data, we can establish a small subset of informative genes from which the regulatory interaction pattern can be generated.[6] The formation of these small subset of genes is known as gene selection.[7,-10] Gene selection from microarray data is a highly statistically significant problem that have persisted since long time. One of the major issues is the number of samples that are usually less as compared to the thousands of genes whose expression levels are measured.[11-12] We know that from thousands of genes available in the gene expression data, all genes are not relevant and disease causing. Hence, it is every bit important to extract those relevant and diseases causing genes from the vast set of irrelevant data available.[13] One of the methods for obtaining the above is through selection or ranking. Gene selection or feature selection can be defined using three fundamental methods or categories: filter method, wrapper method and embedded method.[14,-6] Filter method can be obtained using two operations such as ranking and subset selection based on the ranking, i.e. it assigns scores to each feature.[17] Wrapper methods use the search problem where different combinations of the features are prepared, extracted and compared.[18] Embedded methods select the feature that contributes more in the accuracy of any model which is created.[19] Apart from these techniques, there are various computational techniques and methods for gene selection. Venet *et al.* proposed a signal-to-noise (SNR) for microarray data called SNAGE or signal-to-noise applied to gene experiments.[20] The above method is based on the gene -gene correlation and is a fruitful method for measuring the data quality. Model *et al*. established how phenotypic classes can be predicted by amalgamating feature selection methods and discriminant analysis for methylation pattern based discrimination between acute lymphobatic leukemia and acute myeloid leukemia.[21] Pechenizkiy *et al.* used the principal component analysis for dimensionality reduction after partitioning large datasets with *k-d*-tree.[22] Cavill *et al.* projected a GA/*k*-NN based move for concurrent feature and sample selection from metabolic profiling data.[23] Hengpraprohm *et al.* proposed a method for weighing the feature value by SNR score.[24] They compared their outcomes with different feature selections and classifiers and it was noted that it achieved good effect in terms of accuracy of the classifier. Goh *et al.* proposed a SNR method that is hybridized by Pearson's correlation coefficient according to their discrimination power towards the classes.[25] A hybridization technique using Independent Component Analysis (ICA) and SNR was proposed by Aziz *et al.* for feature selection or extraction.[26] They expressed that their proposed combined method provided better results with naïve Bayesian classifier as compared to the existing methods.A. Kourid used parallel *k*-means on MapReduce for clustering features and then they applied iterative MapReduce that uses SNR as means for ranking clusters.[27] This method provided a better performance in terms of accuracy when applied to large data sets. Maulik *et al.* proposed a prediction technique by combining fuzzy preference based rough set method for feature selection with semi-supervised SVMs.[28] Cawley *et al.* proposed a straight forward Bayesian approach which gets rid of the regularization parameter fully, by integrating it out systematically using an uninformative Jeffrey's prior.[29] Piao *et al.* projected an Ensemble Correlation-Based Gene Selection algorithm based on symmetrical indecision and Support Vector Machine.[30] Nie *et al.* proposed an optimized subset-level score and algorithm to proficiently discover the global optimal feature subset such that the subset-level score is maximized.[31]Shruti *et al.* proposed an SVM-BT-RFE technique based on the bayesian t-test and recursive feature elimnation technique using SVM.[32]The result obtained outperformed the existing SVM-RFE technique.In our study, the scoring method for the existing TR algorithm have been re-designed using the SNR scoring technique.[33] The TR algorithm's fundamental base lies in the Fisher's or Laplacian score that we substituted with a new scoring scheme for computation of the weight matrices within-the-class and between-the-class. The stated proposed method was then measured with five benchmark datasets i.e. Colon, Leukemia, Medulloblastoma, Lymphoma and Prostate Cancer.[34-38] The dataset is quite large enough in terms of number of genes, but have minimal number of sample size. When compared with the existing TR algorithm then it was found that SNR-TR algorithm ousted in terms of classification accuracy and number of iterations.[39] It was shown that the two classifiers (Resilient and back propagation) that was considered provided excellent results in SNR-TR gene ranking than TR algorithm, with less number of iterations. The result was many a times compared with different *K* value like 50, 100 and 150 where, *K* is the number of genes selected initially and ultimately. The rest of the paper is divided as follows: the first section states the materials and method that have been used in our work, i.e. dataset used, methods and techniques followed like TR algorithm and SNR. The next section shows the experimental evaluation where the detailed process of the work is depicted in a stepwise manner. Following this section, is the result of the proposed technique with respect to the original algorithm. And lastly, are the summary and the conclusion of the work with some relevant future directions.

MATERIALS AND METHODS

(i) Datasets Used

Five different benchmark datasets have been used i.e. Colon cancer, Leukemia, Medulloblastoma, Lymphoma and Prostate cancer. Colon cancer or colorectal adenomas and normal mucosas from 31 patients were downloaded from the Princeton database, where this dataset consists of 62 samples and 2000 genes.[34] The Leukemia dataset is said to consist 10,056 genes with 48 samples of both ALL and AML (24 ALL and 24 AML each).[35] Medulloblastoma dataset [26] have 5893 genes with 34 samples of 25 C and 9 D samples while, Lymphoma dataset [27] have 7070 genes having 77 samples of 58 DLBCL and 19 FL samples (Affymetrix HuGeneFL array), and the prostate cancer dataset [28] have 12,533 genes with 102 samples of 50 normal and 52 tumor samples (Affymetrix Human Genome U95Av2 Array platform).[36-38]

(ii) Signal-to-Noise Ratio

One of the most widely practiced technique is signal-to-noise (SNR).[33] It is rather popular because of its ease and comfort of utilizing it. The quality of the biological data can be assessed as the amount of biological signal to the amount of noise. The best part in the SNR technique is, it not only depends on the noise factor but, also on the amount of signal. SNR is also expressed as a statistical criterion for determining the effectiveness of the feature in identifying a class out of another division. In other words, it identifies the pattern with a minimal difference in mean expression between two groups used and a minimal variation of expression within each group. The description of the score can be stated as in eq. (1):

$$SNR = \left|\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}\right| \qquad (1)$$

where, μ_1 and μ_2 denote the mean of the expression level for the samples in class 1 and class 2 respectively. σ_1 and σ_2 denote the standard deviation for the samples in each class. SNR helps to select the features that have high SNR i.e. maximal SNR (or minimal error).

(iii) Trace Ratio Algorithm

The trace ratio problem is an optimization obstacle concerned with many dimensionality reduction algorithms. Conventionally, the result is approximated via generalized eigen value decomposition due to the difficulty of the original problem. TR is a graph based gene or feature selection algorithm that uses the two scores (Fisher and Laplacian score) as the evaluation criterion measure.[39,40,41] Consider two undirected graphs G_w and G_b for within-class and between-class relations that are constructed using Fisher score, where the equivalent adjacency matrices being W_w and W_b. For a dataset *X*, where both the instances x_i and x_j belong to the same class, the within-class relationship will be higher. Thus, the feature subset selection should minimize (eq. (2)),

$$\Sigma_{ij}\|l_i - l_j\|^2 (M_w)_{ij} \quad (2)$$

for the same class, otherwise maximize. Between-class relationship, both for x_i and x_j will be higher when they belong to different classes. So, the selected gene or feature subset should maximize (eq.(3)),

$$\Sigma_{ij}\|l_i - l_j\|^2 (M_b)_{ij} \quad (3)$$

for the different classes, otherwise minimize. Here, l_i is the instance of class for x_i. In order to find the weight matrices M_w and M_b, fisher score or laplacian score is used based on whether it is supervised or unsupervised feature selection. The weight matrices for fisher score can be classified as given below in eq. (4) and eq. (5):

$$(M_w)_{ij} = \begin{cases} \frac{1}{num_{l_i}}, if\ l_i = l_j \\ 0, \quad if\ l_i \neq l_j \end{cases} \qquad (4)$$

$$(M_b)_{ij} = \begin{cases} \frac{1}{num} - \frac{1}{num_{l_i}}, if\ l_i = l_j \\ \frac{1}{num}, \quad if\ l_i \neq l_j \end{cases} \qquad (5)$$

Where, l_i denotes the class label of the ith instance of x_i and num_{l_i} denotes the number of data or records belonging to class l_i. The adjacency matrix using Laplacian score can be calculated as shown in eq (8) and eq (9):

$$(M_w)_{ij} = \begin{cases} e^{\frac{\|x_i - x_j\|^2}{t}}, & if\ x_i\ and\ x_j\ are\ neighbours \\ 0, & otherwise \end{cases} \quad (6)$$

$$(M_b)_{ij} = \left\{ \frac{1}{1^T DM_w 1} DM_w 11^T DM_w \right. \quad (7)$$

Where, eq.(6) denotes the radial distance and *t* denotes any constant. In order to unite both the objectives in a single function, the ratio of the two is considered and maximized. The ratio is given by eq.(8) and eq.(9):

$$\varphi(S_p) = \frac{\sum_{ij} \|l_i - l_j\|^2 (M_b)_{ij}}{\sum_{ij} \|l_i - l_j\|^2 (M_w)_{ij}} \quad (8)$$

$$\varphi(S_p) = \frac{tr(S_p^T XLM_b X^T S_p)}{tr(S_p^T XLM_w X^T S_p)} \quad (9)$$

Where, $S_p = [s_{i_1}, s_{i_2}, \ldots . s_{i_k}]$ denotes the selection matrix, where i_1, i_2,... i_k are the first *k* elements of the transformation [1,2,...n], which is gene or feature number. s_{i_r} denotes a column matrix with all zeros excluding 1 in the r^{th} position and *tr* is the TR of the matrix. Let LM_w and LM_b are Laplacian matrices of the form given in eq.(10) and eq.(11):

$$LM_w = DM_w - M_w \quad (10)$$

$$LM_b = DM_b - M_b \quad (11)$$

Where, DM_w and DM_b are diagonal matrices given in eq.(12) and eq.(13).

$$(DM_w)_{ii} = \sum_{ij} (M_w)_{ij} \quad (12)$$

$$(DM_b)_{ii} = \sum_{ij} (M_b)_{ij} \quad (13)$$

Let $Y = XLM_b X^T$ and $Z = XLM_w X^T$. The score of the feature or gene set is calculated as per the TR criteria for a particular selection matrix S_p which is given as in eq.(14),

$$\beta = \varphi(S_p) = \frac{tr(S_p^T Y S_p)}{tr(S_p^T Z S_p)} \quad (14)$$

Score of each gene or feature f_i is computed using eq (15),

$$F(f_i) = m_i^T (Y - \beta Z) m_i \quad (15)$$

Where, m_i is the column vector with all zeros except 1 and the i^{th} position, and *F* is the selected feature or gene set. The algorithm of the trace ratio is stated below (shown in Algorithm I):

Algorithm I: Trace Ratio [39]

Step 1: Calculate adjacency matrices for within the class (M_w) and between the classes (M_b) using Fisher score as follows (eq. (4) and eq. (5)):

$$M_w = \frac{1}{num_{l_i}}, if l_i = l_j \wedge 0,\ if l_i != l_j$$

$$M_b = \frac{1}{num} - \frac{1}{num_{l_i}}, if l_i = l_j \wedge \frac{1}{num}, if l_i != l_j$$

Step 2: Calculate the diagonal matrices (DM_w and DM_b) for the above adjacency matrices as given below (as in eq.(12) and eq.(13):

$$(DM_w)_{ii} = \sum_{ij}(M_w)_{ij}$$

$$(DM_b)_{ii} = \sum_{ij}(M_b)_{ij}$$

Step 3: Calculate Laplacian matrices (LM_w and LM_b) using the eq.(10) and eq.(11).

$$LM_w = DM_w - M_w$$

$$LM_b = DM_b - M_b$$

Step 4: Construct a matrix of k features by initially selecting randomly k features from original dataset (say R_k).
Step 5: Declare an empty matrix (say N_k)to store top k features after finding scores of each feature
Step 6: Repeat steps 6 to 10 until $R_k! = N_k$
Step 7: Calculate $Y = XLM_bX^T$ and $Z = XLM_wX^T$
Step 8: Calculate Trace Ratios as $TR_y = TR(R_k^T Y R_k)$ and $TR_z = TR(R_k^T Z R_k)$
Step 9 : Calculate $\beta = \frac{TR_y}{TR_z}$
Step 10 : Calculate Score of each feature as $F(f_i) = m_i^T(Y - \beta Z)m_i$
Step 11: Select new top k features based on the score and store in N_k
Step 12: Store final k features R_k for further processing
Step 13: Stop

EXPERIMENTAL EVALUATION

In this part, we would give the basic pre-processing step that was utilized for our five datasets. This would be further accompanied by a schematic view of the suggested model. In the evaluation process, MATLAB version R2014a was used with the system requirement of 8GB RAM.

(i) Pre-processing

One of the primal stages of pre-processing is normalization. Normalization process formulates the data into an outline that will be more simplified and effectively processed for the intent of the user. In our area, the data sets were normalized using min-max normalization.[42] Min-Max normalization is a graceful technique where the technique can mostly fit the data in a pre-defined boundary with a pre-set limit. In other words, it's a way that one linearly formulates the real data values such that the minimum and the maximum of the transformed data to take certain values. The technique can be represented as shown in eq.(16):

$$x' = \frac{(x - x_{min})}{(x_{max} - x_{min})} \qquad (16)$$

where, x_{min}=minimal data value appearing and x_{max}= maximal data value appearing.

(ii) Implementation and Performance Analysis

The algorithm of the proposed SNR-TR gene ranking algorithm is shown as follows in algorithm II:

Algorithm II: Modified Trace Ratio Algorithm (SNR-TR Gene Ranking Algorithm)

Step 1: Calculate adjacency matrices for with-in the class (M_w) and between the classes (M_b) using Signal-to-Noise Ratio as follows:

$$(M_w)_{ij} = \left|\frac{\mu_i - \mu_j}{\sigma_i + \sigma_j}\right|, if\, l_i = l_j \wedge 0, Otherwise$$

$$(M_b)_{ij} = \left|\frac{\mu_i - \mu_j}{\sigma_i + \sigma_j}\right|, if\, l_i \neq l_j \wedge 0, Otherwise$$

where, i and j are sample numbers.

Step 2: Calculate the diagonal matrices (DM_w and DM_b) for the above adjacency matrices as given below (as in eq.(12) and eq.(13):

$$(DM_w)_{ii} = \sum_{ij}(M_w)_{ij}$$

$$(DM_b)_{ii} = \sum_{ij}(M_b)_{ij}$$

Step 3: Calculate Laplacian matrices (LM_w and LM_b) using the eq.(10) and eq.(11).

$$LM_w = DM_w - M_w$$

$$LM_b = DM_b - M_b$$

Step 4: Construct a matrix of k features by initially selecting randomly k features from original dataset (say R_k).

Step 5: Declare an empty matrix (say N_k) to store top k features after finding scores of each feature

Step 6: Repeat steps 6 to 10 until $R_k! = N_k$

Step 7: Calculate $Y = XLM_bX^T$ and $Z = XLM_wX^T$

Step 8: Calculate Trace Ratios as $TR_y = TR(R_k^T Y R_k)$ and $TR_z = TR(R_k^T Z R_k)$

Step 9 : Calculate $\beta = \frac{TR_y}{TR_z}$

Step 10 : Calculate Score of each feature as $F(f_i) = m_i^T(Y - \beta Z)m_i$

Step 11: Select new top k features based on the score and store in N_k

Step 12: Store final k features R_k for further processing

Step 13: Stop

Figure 1, depicts the overall model of the proposed SNR-TR gene ranking algorithm. After normalizing the dataset using the min-max normalization, the refined dataset would be taken. The SNR-TR gene ranking phase involves certain sub-phase within it were using all score of SNR the weight matrix between the class and within the class is redefined. *K* number of genes are passed to the SNR-TR algorithm in order to obtain the rank of all the genes with respect to that *K*. The range of *K* in our domain varies significantly as 50, 100 and 150. Based on this metric, the new ranked list is generated with less number of iterations. Subsequently, this set of rank list is handed to the variants of neural network classifier for detecting the accuracy and simultaneously validating the rank list. In other words, the classifiers act as a validation metric for the obtained rank list. Ultimately, we found that instead of changing the base algorithm if we are able to suitably change the scoring pattern of the TR algorithm, then a huge difference in the performance can be obtained. In the end, it was noted that the accuracy of the system improved a good deal as compared to the existing TR algorithm and the number of iterations for convergence of the algorithm is rather less. For most of the datasets, the accuracy obtained was 100%.

Schematic View of the proposed model

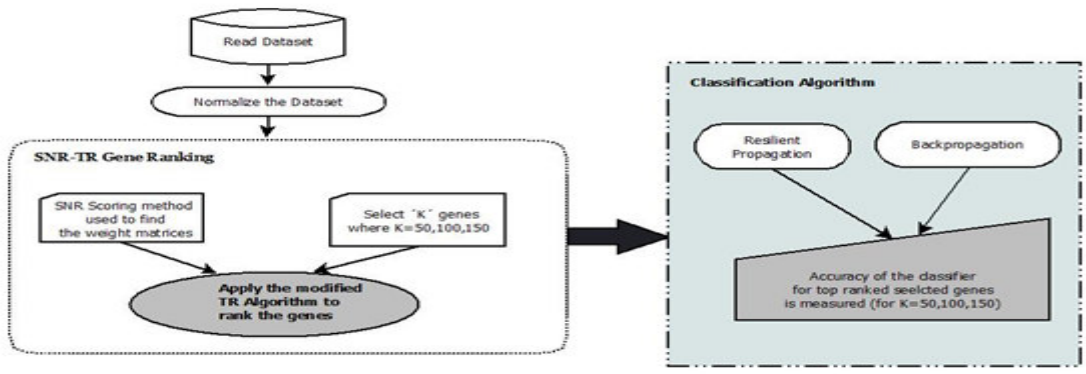


Figure 1
Schematic View of the proposed model

RESULTS AND DISCUSSION

Table 1, Table 2, Table 3, depicts the performance of the original TR algorithm and the proposed SNR-TR algorithm for *K*=50, 100 and 150 genes. Two classifiers have been habituated and it was noticed that in most of the dataset, the accuracy is 100% in less number of iterations in comparison with the original TR algorithm. The datasets used for our domain is basically two class dataset but it can still work the same in multi-class datasets.

Table 1
Performance Result of Original TR algorithm and SNR-TR Gene Ranking Algorithm for K=50 genes

Dataset	Resilient Propagation				Back Propagation			
	Original TR Algorithm		SNR-TR Gene Ranking Algorithm		Original TR Algorithm		SNR-TR Gene Ranking Algorithm	
	Acc	Itr	Acc	Itr	Acc	Itr	Acc	Itr
Colon	98.36	122	98.38	160	79.03	2024	98.38	544
Leukemia	91.6	12	95.8	34	91.66	14	100	42
Medulloblastoma	99.10	44	99.01	41	76.47	725	100	79
Lymphoma	98.7	2335	100	3242	97.40	4901	98.04	3982
Prostate Cancer	99.01	8197	99.01	1731	50.98	4987	59.11	4775

Table 2
Performance Result of Original TR algorithm and SNR-TR Gene Ranking Algorithm for K=100 genes

Dataset	Resilient Propagation				Back Propagation			
	Original TR Algorithm		SNR-TR Gene Ranking Algorithm		Original TR Algorithm		SNR-TR Gene Ranking Algorithm	
	Acc	Itr	Acc	Itr	Acc	Itr	Acc	Itr
Colon	98.3	108	100	95	97.24	1042	100	215
Leukemia	100	16	100	12	98.45	25	100	27
Medulloblastoma	99.01	41	100	49	98.24	3988	100	41
Lymphoma	98.70	1452	100	1307	99	4254	99.25	2167
Prostate Cancer	99.01	4496	99.01	1344	95.25	2471	97.11	2315

Table 3
Performance Result of Original TR algorithm and SNR-TR Gene Ranking Algorithm for K=150 genes

Dataset	Resilient Propagation				Back Propagation			
	Original TR Algorithm		SNR-TR Gene Ranking Algorithm		Original TR Algorithm		SNR-TR Gene Ranking Algorithm	
	Acc	Itr	Acc	Itr	Acc	Itr	Acc	Itr
Colon	99.01	123	100	104	97.58	1234	98.38	383
Leukemia	91.6	12	93.75	15	93.57	38	100	42
Medulloblastoma	99.01	56	100	38	98.65	4078	100	38
Lymphoma	98.7	2190	100	1515	93.47	4378	95.15	4289
Prostate Cancer	99.01	4108	99.01	1026	88.25	2854	93.51	2711

Table 4, shows the average accuracy obtained from 5 runs for *K*=200 genes. Considerably, almost similar results were obtained as the previous table results where the SNR-TR algorithm outperformed the TR algorithm with more accuracy in less number of iterations.

Table 4
Average performance result of Original TR algorithm and SNR-TR Gene Ranking Algorithm for 5 runs with K=200 genes

Dataset	Resilient Propagation				Back Propagation			
	Original TR Algorithm		SNR-TR Gene Ranking Algorithm		Original TR Algorithm		SNR-TR Gene Ranking Algorithm	
	Acc	Itr	Acc	Itr	Acc	Itr	Acc	Itr
Colon	98.22	142	100	102	94.10	405	97.44	278
Leukemia	99.01	25	99.58	17	99.01	52	100	39
Medulloblastoma	98.14	56	99.41	37	93.22	79	98.82	55
Lymphoma	97.56	1254	98.7	988	93.11	3658	95.39	3179
Prostate Cancer	99.01	1087	99.80	940	88.22	3567	94.11	2542

Figure 2- 6, shows the accuracy graph for the original TR algorithm and the SNR-TR method for *K*=50, 100 and 150 of Resilient propagation. It can be clearly seen that the proposed method provides a good accuracy measurement as compared to the other one.

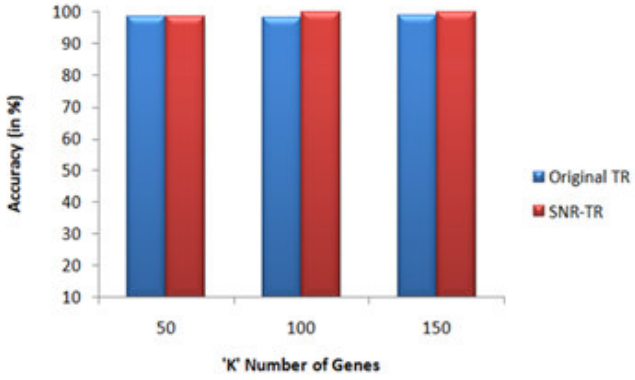


Figure 2
Resilient propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Colon dataset for K=50, 100, 150

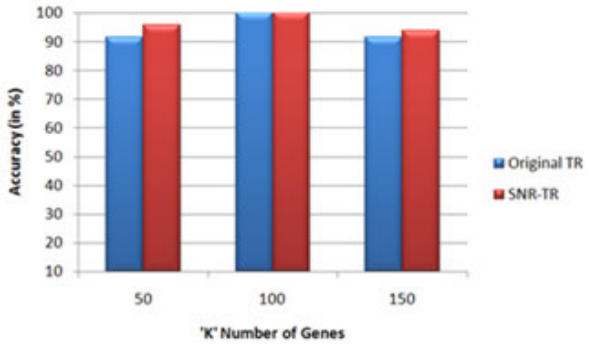


Figure 3
Resilient propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Leukemia dataset for K=50, 100, 150

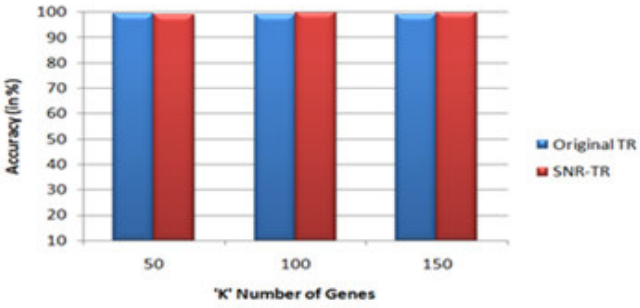


Figure 4
Resilient propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Medulloblastoma dataset for K=50, 100, 150

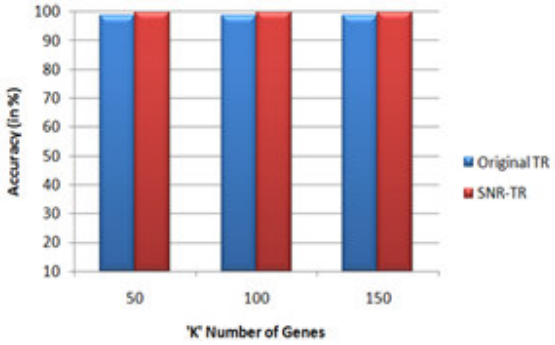


Figure 5
Resilient propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Lymphoma dataset for K=50, 100, 150

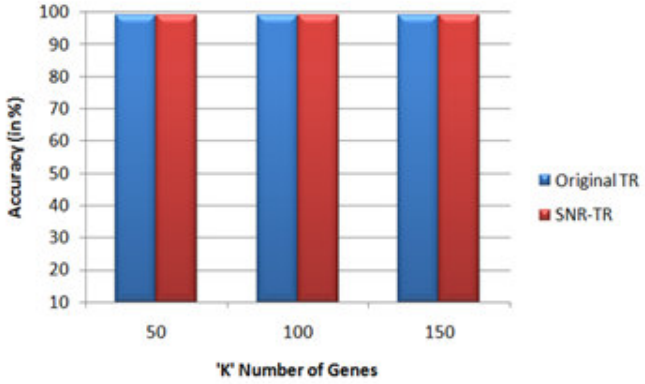
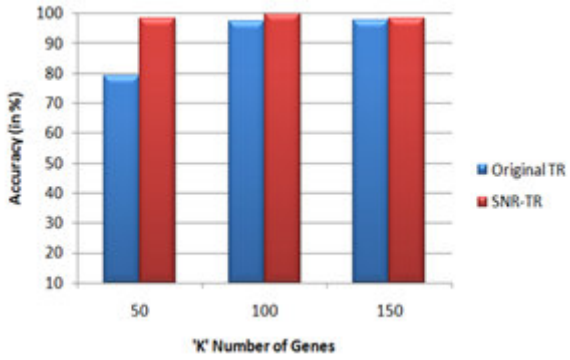


Figure 6
Resilient propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Prostate Cancer dataset for K=50, 100, 150

Similarly, Figure 7-11 also shows the accuracy measurement for K=50, 100 and 150 of back propagation where again the SNR-TR algorithm outperform the original TR algorithm.



Back propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Colon dataset for K=50, 100, 150

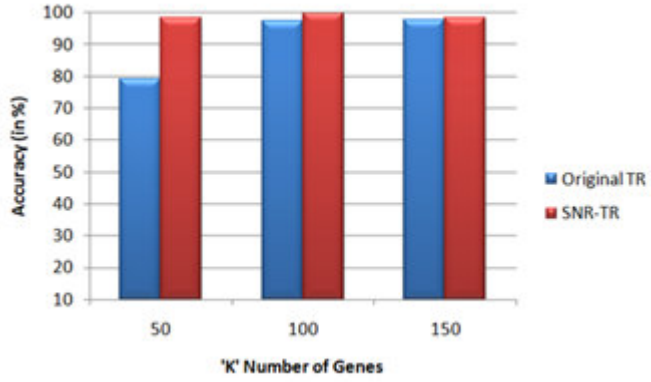


Figure 7
Back propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Colon dataset for K=50, 100, 150

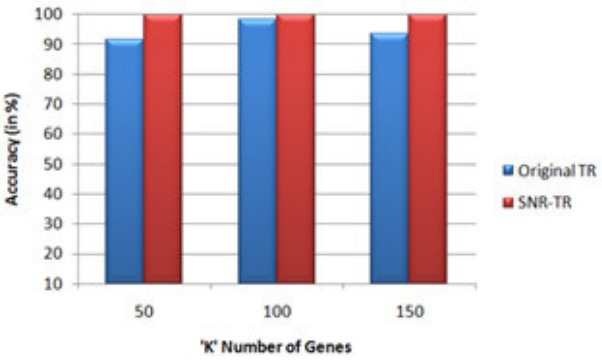


Figure 8
Back propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Leukemia dataset for K=50, 100, 150

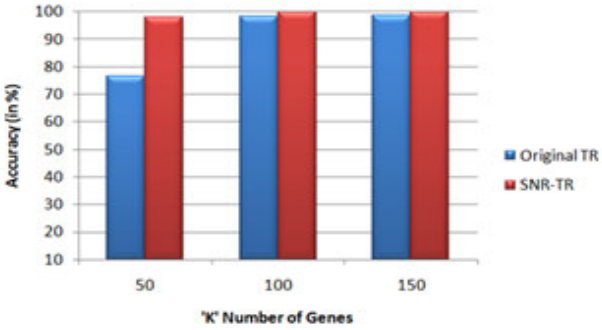


Figure 9
Back propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Medulloblastoma dataset for K=50, 100, 150

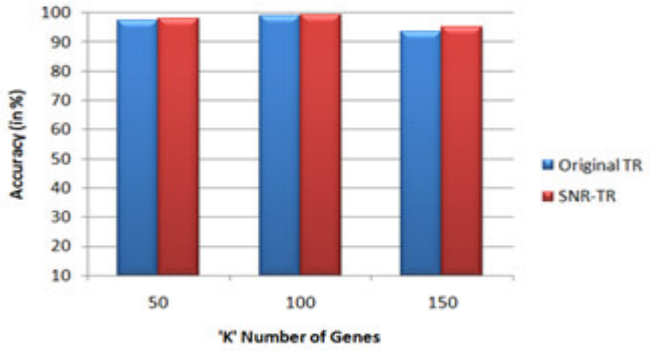


Figure 10
Back propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Lymphoma dataset for K=50, 100, 150

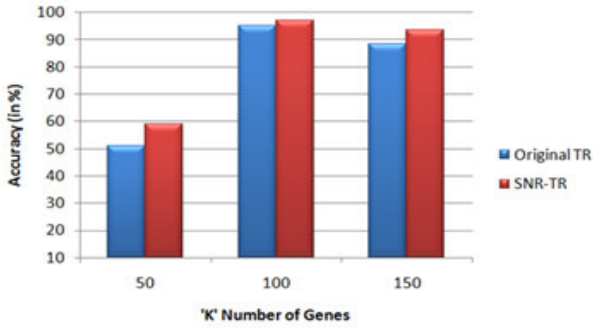


Figure 11
Back propagation accuracy measurement of Original TR algorithm Vs SNR-TR Gene Ranking Algorithm of Prostate Cancer dataset for K=50, 100, 150

SUMMARY

This paper can be summarized as follows

- For the original TR algorithm and SNR algorithm, the dataset has been normalized using the min-max normalization.
- From the original TR algorithm, by selecting the *K* number of genes (50, 100 and 150) the rank of the genes was calculated and was drawn to the classifier for obtaining the accuracy.
- Now, we more or less modified the existing TR algorithm by using the powerful statistical measure called SNR.
- This measure was used to act as a new scoring or ranking criterion for the TR algorithm. Initially, Fisher's score or Laplacian score was used as a scoring criterion that defines the weight matrix between the class and within the class.
- SNR-TR algorithm ultimately generates the new rank list according to the specified K value and this list was subsequently extended to the resilient and back propagation neural network classifier, where the accuracy obtained for most of the dataset was 100%.
- In the end, the proposed method was validated and compared with the existing TR algorithm where it was found that the proposed method outperformed the latter.

CONCLUSION

A new gene selection method was proposed using the statistical SNR method and TR algorithm. The algorithm was used with five different datasets. The main objective of the technique was to generate an efficient rank list from a randomly selected set of genes. The rank list that was obtained was further validated using variants of neural network classifiers and the accuracy of the dataset were found to be mostly 100% with few numbers of iterations. We have also shown the average accuracy of the classifier for *K*=200 genes for 5 runs. For rank generation, different *K* values for considered and it was concluded that by choosing a small set of *K* we are able to acquire a better ranking pattern for the genes.

CONFLICT OF INTEREST

I on behalf of all authors would like to state that the no case of animal study conduction has been done in accordance of the relevant ethical committee.

REFERENCES

1. Babu MM, Teichmann SA. Evolution of transcription factors and the gene regulatory network in Escherichiacoli. Nucleic Acids Research. 2013; 31: 1234-1244.
2. Liu GX, Feng W. Reconstruction of Gene Regulatory network Based on Two-Stage Bayesian Network Structure learning Algorithm. Journal of Bionic Engineering. 2009; 6: 86-92.
3. Mishra S, Mishra D. Methodologies for Modelling Gene Regulatory Networks. Encyclopedia of Information Science and Technology. 2014: 426-436.
4. Li P, Zhang C. Comparison of Probabilistic Boolean Network and Dynamic Bayesian network Approaches for Inferring Gene Regulatory Networks. BMC Bioinformatics. 2007; 8: 1-10.
5. Zhao W, Serpedin E. Inferring Gene Regulatory Networks from Time series Data Using the Minimum Description Length Principle. Journal of Bioinformatics. 2006; 22: 2129-2135.
6. Chai LE, Lob SK, Low ST, Mohammad MS, Deris S, Zakaria Z. A Review on the Computational Approaches for Gene Regulatory Network. Computers in Biology and Medicine. 2014; 48: 55-65.
7. Lee SH. Minimum Gene Selection using BSWFM. Indian Journal of Science and Technology. 2015; 8(26): 1-6.
8. Das K, Ray J, Mishra D. Gene Selection using Information Theory and Statistical Approach. Indian Journal of Science and Technology. 2015; 8(8): 695-701.
9. Panigrahi L, Das K, Mishra D. Missing Imputation using Hybrid Higher Order Neural Classifier. Indian Journal of Science and Technology. 2014; 7(12): 2007-2014 .
10. Tyagi V, Mishra A. A survey on different feature selection methods for microarray data analysis. International Journal of Computer Applications. 2013; 67 (16): 36-40.
11. Alshamlan HM, Badr GH, Alohali YA. The performance of bio-inspired evolutionary gene selection methods for cancer classification using microarray dataset", International Journal of Bioscience. Biochemistry and Bioinformatics. 2014; 4 (3): 166-170.
12. Lee CP, Leu Y. A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing. 2011; 11 (1): 208-213.
13. Karlebach G and Shamir R. Modeling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008; 9: 770-780.
14. Satapathy SK, Jagadev AK, Dehuri S. An Empirical Analysis of Training Algorithms of Neural Networks: A Case Study of EEG Signal Classification Using JAVA Framework. Advances in Intelligent Systems and Computing. 2014; 309: 151-160.
15. Satapathy SK, Dehuri S, Jagadev AK. An Empirical Analysis of Different Machine Learning Techniques for Classification of EEG Signal to detect Epileptic Seizure. International Journal of Applied Engineering and Research. 2016; 11(1): 120-129.
16. Maldonado S, Weber R, Famili F. Feature selection for high dimensional class-imbalanced datasets using support vector machines. Information Science. 2014; 286: 228-246.
17. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, Schaetzen V de, Dugue R, Bersini H, Nowe A. A Survey on filter techniques

for feature selection in Gene Expression Microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2012; 9(4): 1106-1119.
18. Li P, Zhang C. Comparison of Probabilistic Boolean Network and Dynamic Bayesian network Approaches for Inferring Gene Regulatory Networks, BMC Bioinformatics. 2007; 8: 1-10.
19. Suganthi J, Malathi V. Fuzzy based Feature Selection Scheme through Transductive SVM Technique for Cancer Pattern Classification and Prediction. Indian Journal of Science and Technology. 2016; 9(16): 1-7.
20. Venet D, Detours V, Bersini, H. A Measure of the Signal-to-Noise Ratio of Microarray Samples and Studies Using Gene Correlations. PLOS One. 2012; 7(2): 1-9.
21. Model F, Adorjan P, Olek A, Piepenbrock C. Feature Selection for DNA methylation based cancer classifiication. Bioinformatics. 2001; 1(17): 157-164.
22. Pechenizkiy M, Puuronen S, Tsymbal A. The impact of sample reduction on PCA-based feature extraction for supervised learning. Proc. of the 21st ACM Symposium on Applied Computing. 2006: 553-558.
23. Cavill R, Keun H, Holmes E, Lindon J, Nicholson J, Ebbels T. Genetic algorithms for simultaneous variable and sample selection in metabonomics. Bioinformatics. 2009; 25(1): 112-118.
24. Hengpraprohm S, Chongstitvatana P. Feature selection by Weighted-SNR for Cancer Microarray Data Classification. International Journal of Innovative Computing, Information and Control. 2009; 5(12): 4627-4635.
25. Goh L, Song Q, Kasabov N. A Novel Feature Selection Method to Improve Classification of Gene Expression Data. 2nd Asia-Pacific Bioinformatics Conference (APBC2004). 2004; 29: 161-166.
26. Aziz R, Verma CK, Srivastava N. A Weighted-SNR Feature Selection from Independent Component Subspace for NB Classification of Microarray Data. International Journal of Advanced Biotechnology and Research (IJBR). 2015; 6(2): 245-255.
27. Kourid A. Iterative MapReduce for feature selection. International Journal of Engineering Research & Technology (IJERT). 2014; 3(7): 1788-1793.
28. Maulik U, Chakraborty D. Fuzzy Preference Based Feature Selection and Semi-supervised SVM for Cancer Classification. IEEE Transactions On Nano-bioscience. 2014; 13(2): 152-160.
29. Cawley GC and Talbot NLC. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics. 2006; 22(19): 2348-2355.
30. Piao Y, Piao M, Park K and Ryu KH. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics. 2012; 28(24): 3306–3315.
31. Nie F, Xiang S, Jia Y, Zhang C and Yan S. Trace Ratio Criterion for Feature Selection. Proc. of the Twenty-Third AAAI Conference on Artificial Intelligence. 2008; 22(8): 671-676.
32. Mishra S, Mishra D. SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm. Karbala International Journal of Modern Science. 2015; 1(2): 86-96.
33. Plapous C, Marro C, Scalart P. Noise Reduction Using Reliable A Posteriori Signal-To-Noise Ratio Features. 14th European Signal Processing Conference (EUSIPCO 2006). 2006: 1-5.
34. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. of National Academy of Sciences of the United States of America. 1999; 96(12): 6745-6750.
35. Leukemia
Set, http://www.github.com/Leukemia.gct.
36. Broad institute, http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.
37. Shipp MA, Ross KN, Jackson DG, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine. 2002; 8: 68-74.
38. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, DÁmico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behaviour. Cancer Cell. 2002; 1: 203-209.
39. Wang H, Yan S, Xu D, Tang X, Huang T. Trace Ratio vs. Ratio Trace for Dimensionality Reduction. Proc. of 2007 IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1-8.
40. Beniwal S, Kumar D. Gene Selection using Bacterial Foraging Optimization. International Journal of Pharma and Bio Sciences. 2015; 6(2): 663-667.
41. Mary MJ, Deecaraman, M, Vijayalaskshmi, M, Umashankar V. A Systemic Review of Differential Regulation of Genes in Polycystic Ovarian Syndrome Disease. International Journal of Pharma and Bio Sciences. 2015; 6(2): 893-900.
42. Suarez-Alvarez MM, Pham DT, Prostov MY, Prostov YI. Statistical approach to normalization of feature vectors and clustering of mixed datasets. Proc. Royal Society. 2012; 468 (2145): 21-30.