International Journal of Engineering & Technology, 7 (3.27) (2018) 62-69

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Gene Selection Approaches for Classifying Disease Relevant Data Sample

J. Briso Becky Bell[1*], S. Maria Celestin Vigila[2]

[1]Research Scholar, CSE Department, Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India.
[2]Associate Professor, IT Department, Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India.
E-mail:celesleon@yahoo.com
**Corresponding author E-mail:brisobell@gmail.com*

Abstract

In the latest field of gene expression profiling, the identification of most highly expressed genes with respect to diseases is been in focus lately, As to study the disease types and classify normal from disease syndrome samples. This paper portrays four gene selection approaches such as Pearson correlation, Signal to Noise Correlation, Feature Assessment by Sliding threshold and Feature Assessment by Information Retrieval for retrieving highly relevant genes oriented to a specific disease. This experiment uses various disease dataset for operating on the typical gene selection methods and to select top ten most relevant genes and thus selected genes are learned on using classifiers such as Support Vector Machine, K-Nearest Neighbour and Naïve Bayes to classify the specific disease oriented classes distinctively. Here we also compare the performance of our classifier with the previous papers techniques using classification Accuracy.

Keywords: *Microarrays, gene-expression, genomics, wrapper, dimensionality reduction.*

1. Introduction

DNA microarrays [1] are recent technology for gene expression profiling. In this technology the level of expression of thousands of genes, or even entire genome, can be estimated for a sample of cells Within a hybridization. Usually the microarray data are images, which are converted into gene expression matrices, in this matrix the columns consists of various genes belonging to a specific tissues and rows consists of various samples belonging to mutated as well as normal disease condition, and intersection cells of this matrix holds the expression level of particular gene in a sample.

In recent time gene expression profiles is the most ideally suited system for disease diagnostics[2]. The number of genes in a gene expression profile is always much larger in number present in a range of 2,000-54,675than the number of samples ranging only 24-100.This low sample high dimensional data is very hard when it is analysed manually. So there is a need for automatically analysing the microarray data which matters the most on identifying disease genes from expression matrix.

A diagnostic system designed to get the relevant set of expressed genes from large gene collections has very high computational cost. It also provides lower classification accuracy and slower learning process due to the extreme dimensionality in gene count. Recently researches [3] have assessed that a very small number of genes are typically enough for accurately diagnosing most of the disease cases. Rather more, by using a minimal subset of genes, so one can get an opportunity to further examine the nature of the diseases and the genetic functions responsible for it. Thus gene selection plays a exquisite role in selecting vital genes from gene expression data.

Feature selection and feature extraction are the two different approaches for handling gene selection. In feature selection, it selects a subset of genes from the set of available genes which saves the computation cost and the selected genes retain their original physical interpretation. Feature extraction nonlinearly or linearly transforms the original gene sets into reduced one and thus transformed genes generated by feature extraction are very hard to interpret and do not have a clear physical meaning. Therefore feature selection is preferred more than other methods for a gene expression profile based diagnostic system.

Further feature selection methods are classified into three categories depending on the way they associate classification models with the feature selection methods. In filter approach [4] the feature selection is performed independently of the learning algorithm. The wrapper method [5] uses the learning algorithm in the feature selection process. The last method is termed as embedded technique, [6] it uses searches for optimality in subset of feature and is built as a classifier construction and it can be seen as a search in the combined subsets of feature space. Filter methods are very popular to high dimensional data because of the high computational efficiency, and seems to be an appropriate method in selecting informative genes from high dimensional low sample gene expressions.

2. Literature Survey

In the field of Omics, genomics [1] has gained popularity and microarray technology is widely used to measure the expression in thousands of genes simultaneously for studying the nature of certain diseases, treatments, and development of vaccines. There are various microarray methods that are used to realize this idea of sample, a membrane or glass slide is "arrayed" or spotted with oligonucleotides or fragments of DNA that represents specific gene coding region. The Purified RNA is labelled as by radioactive or in a fluorescent manner and hybridized to the glass slide /membrane. Then the raw data is obtained by washing, using auto radiographic imaging or laser scanning. Thus the microarray imaged data, are transformed into *n*m* expression matrices as shown in the Table I.

Each row in the Table I, represent a sample that consists of *m* genes from one experiment. Each sample belongs to certain class Normal (N)/Disease (D)in some cases or Type-1(T_1)/ Type-2(T_2) in other cases. In each data set the researchers repeated the same experiment on *n* different samples, each line in this data set representing the samples. The numbers in each arrayed cell characterize the level of expression of particular gene in a sample. A typical gene expression experiment produces expression level up to 54,613genes for about 55 samples in pancreatic cancer dataset.

Table I: Gene Expression Data Matrix

S	G_1	G_2	…	G_{m-1}	G_m	Class
S_1	96.42	21.43	…	71.59	40.71	N/D
S_2	38.42	29.19	…	37.06	31.15	N/D
S_3	98.6	43.12	…	54.7	12.4	N/D
⋮	⋮	⋮	…	⋮	⋮	⋮
S_{n-1}	54.25	67.52	…	16.46	37.68	N/D
S_n	21.72	38.05	...	12.42	26.41	N/D

As the microarray data is highly dimensional in nature, gene selection has gained major interests for such type of research. In an experiment, from thousands of genes only a least amount of genes show strong relevance in targeted phenotypes. By research some have exposed that a limited number of genes are enough for diagnosing accurately for most diseases and the number of genes varies highly between different diseases. So the prediction accuracy is increased and computation speed is reduced via gene selection methods.

3. Experimental Design

Four gene selection methods are developed to select the highly expressed gene features. The high dimensional imbalanced microarray binary class datasets are input to the system and four continuous genes selection methods are developed, Pearson Correlation Coefficient (PCC), Signal to Noise Coefficient(S2N), Feature Assessment by Information Retrieval (FAIR)and Feature Assessment by Sliding Threshold (FAST) for to select the most expressive genes in a dataset forming reduced gene size. The system developed for Expressing Significant Genes by Gene Reduction is shown in Fig.1.Then the reduced gene sets are ranked based on expressed gene value and the ranked genes sample sets are constructed and those samples are classified using classifiers such as Support Vector Machine (SVM), K-Nearest Neighbour(K-NN) and Naïve Bayes methods and classifier performance in evaluated.

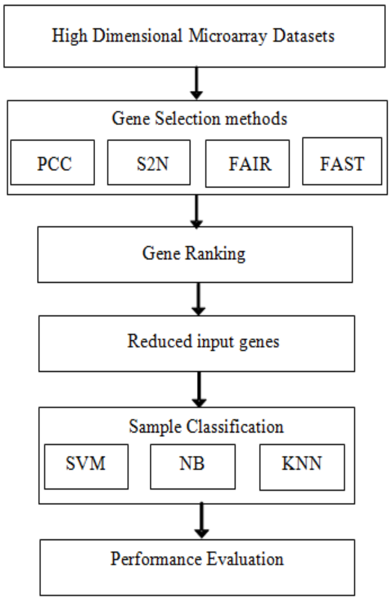


Fig. 1: Classification of disease sample using gene selection methods

4. Gene Selection

Gene Selection (GS)methods [7] are designed to operate on continuous data and they do not require any pre-processing. Here Gene expression data is applied on GS methods.

Pearson Correlation Coefficient

PCC [8] is a statistical measure that tests the quality and strength of the relationship of two variables. The magnitude of coefficients ranges from 1 to −1. Where, the values closer to 1 indicate a stronger relationship. And the direction of the relationship is given by the sign of the coefficient. When it is negative, one variable increases as the other decreases. If it is positive, then the two variables decrease or increase with each other, PCC is used to evaluate the accuracy of a gene prediction on the target independent in the context of other genes. Then the featured genes are arranged based on the correlation score. For problems where the covariance cov(Xi,Y) between a GeneXi and the Class$Y$ and the variances of the Genevar(Xi) and Classvar(Y) are known, the correlation can be directly known by using the formula given in Equation (1).

$$\text{PCC}=\frac{1}{N-1}\sum\left(\frac{X-\mu X}{\sigma X}\right)\left(\frac{Y-\mu Y}{\sigma Y}\right) \tag{1}$$

Signal to Noise Coefficient

The signal-to-noise ratio [8] is originally a concept in electrical engineering. It is the ratio of a signal's power to the power of the noise present in the signal. If a signal has a lot of noise present, it is much more difficult to isolate the signal. It compares the ratio of the difference between the class means to the sum of the standard deviations for each class. The signal-to-noise correlation coefficient (S2N) is a similar measurement to PCC. But instead of taking the correlation on genes and targets the two distinct class genes are correlated. For a given gene, if the two class means are distant from each other, there is less chance of a sample being drawn from the other class. Else if the class means are close, and there is a high chance of mislabelling, and if the standard deviations is larger or smaller it scales the distance appropriately. The formula for calculating S2N is represented in equation (2).

$$\text{S2N}=\frac{\mu_1-\mu_{-1}}{\sigma_1+\sigma_{-1}} \tag{2}$$

Feature Assessment by Sliding Threshold

In FAST, [8] classification of the samples based on multiple thresholds and gathering statistics on the performance at each boundary is done. Here one can calculate the True Positive Rate (TPR) and False Positive Rate (FPR) at multiple thresholds using (3) & (4).In order to find TPR & FPR one has to find the total number of True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*) and False Negatives (*FN*) then the Region under Curve is calculated by Area Under the Curve (AUC). This score can be used for gene ranking, while choosing the genes, having the highest AUC's which ensures the best predictive power for the dataset. The formula for the area is given in (5).

$$\mathrm{TPR}\,(1-\beta) = \frac{TP}{(TP+FN)} \qquad (3)$$

$$\mathrm{FPR}\,(1-\alpha) = \frac{FP}{(FP+TN)} \quad (4)$$

$$\mathrm{AUC} = \sum_i \left\{ (1-\beta_i \cdot \Delta\alpha) + \frac{1}{2}\left[\Delta(1-\beta)\cdot\Delta\alpha\right] \right\} \quad (5)$$

Where,

$$\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1})$$
$$\Delta\alpha = \alpha_i - \alpha_{i-1}$$

Feature Assessment by Information Retrieval

The major deviation of FAIR [8] was the use of P-R curve as our non-parametric statistic. The PRC [9] are vastly different than FAST and strongly indicate the use of one algorithm over the other. This modified approach is called Feature Assessment by Information Retrieval (FAIR) because it uses the information retrieval standard evaluation statistics of precision and recall to build the curve. This is accomplished by examining the P-R curves built by starting from each direction and taking the maximum area. For the P-R curve, here one can simply take a parallel tabulation of the precision and recall for the majority class and build the P-R curve from these values, and take the maximum area. Precision and recall can be calculated using formulas (6) & (7). Then area under the P-R curve can be calculated using formula (5).

$$\mathrm{Precision}(1-\beta) = \frac{TP}{(\mathrm{TP}+\mathrm{FP})} \qquad (6)$$

$$\mathrm{Recall}(1-\alpha) = \frac{TP}{(TP+FN)} \qquad (7)$$

5. Classification Mechanisms

Classification of Data is a supervised learning process that intakes labelled data samples and generate a classifier model for classification of new data samples in different classes. Mathematically, this is stated in (8) given a set of data.

$$\{(x_1,y_1),\dots\dots(x_n,y_n)\}\ \ h{:}\,X \rightarrow Y \ \ ; \qquad (8)$$

The objective is to produce a classifier *h* which maps an object to its classification label.

$$x \in X \qquad y \in Y$$

The classifiers used for classification task is SVM, K-NN and NB and they use various induction techniques for classification.

Support Vector Machine (SVM)

It [10] is a new classification method for both non linear and linear data. It uses a nonlinear mapping for transforming the original training data into a higher dimension. With a new dimension, it searches for the decision boundary or linear optimal separating hyperplane. So with an appropriate nonlinear mapping to a high dimension, one can segregate data from two classes by a hyperplane. SVM finds this hyperplane using support vectors or essential training tuples and margins which are defined by the support vectors. Features training on this classifier can be slow in computation but shows high accuracy thus enabling their ability to model nonlinear complex decision boundaries.

Here it starts with a set of data X = $\{(x_i, c_i)\}$, where each xi is a training sample and c_i is set of associated samples for to be classified and the hyper-plane is written using equation $w^Tx + w_0 = 0$. The goal is to select the weight vector and bias that separate the data at maximum limit. If the two parallel hyper-planes is having the maximum margin then it is expressed as $w^Tx+w_0 = \pm1$. This procedure is account for each sample of the classes in c_i by seeing weather all $w^Tx_i + w_0 \geq 1$.

Naïve Bayes Classifier (NB)

It is a statistical classifier [11] which predicts based on probabilistic calculations, i.e.,It predicts class membership purely based on Bayes' Theorem. Here each training example can incrementally decrease/increase the probability that a hypothesis is correct prior knowledge and is combined with observed data. Bayesian methods provide a standard optimal decision making even when they are computationally intractable.

A probability model can be suited for using the features as conditions for the probability of a sample being drawn from a class. In a probability model, first it is required to find $p(C|F_1, \dots, F_n)$, where each F_i is the value for each feature and C is the class of the sample. This is commonly called the posterior. Once the posterior for each class is found, then it is assigned for a sample to the class with the highest posterior. And by using Bayes' rule, one can express the posterior as a ratio of the prior times the likelihood over the evidence. Formally, this is expressed as given in (9).

$$p(C \mid F_1,\dots,F_n) = \frac{p(C)p(F_1,\dots,F_n \mid C)}{p(F_1,\dots,F_n)} \qquad (9)$$

K-Nearest Neighbor Classifier (KNN)

The nearest neighbor algorithm [10] is an instance-based Lazy learning algorithm, which defer the computation for classifying a sample until a test sample is ready to be classified. It meets the criteria by storing the entire training set in memory and calculating the distance from a test sample to every training sample at classification time; the predicted class of the test sample is the class of the closest training sample.

The nearest neighbor algorithm is a specific instance of the K-Nearest Neighbor algorithm where k = 1. In this algorithm, take the test samples which are likely to be classified, and tabulate the classes for each of the k closest training samples and predict the class of the test sample as the mode of the training samples' classes. The mode is the most common element of a set. In binary classification tasks, k is normally chosen to be an odd number in order to avoid ties. Then use k <= 5 because this value is the most fair to the minority class. Nearest neighbor algorithms can use any metric to calculate the distance from a test sample to the training samples. A metric is a two-argument function d(x, y). The standard metric used in nearest neighbor algorithms is Euclidean distance which is given in (10).

6. Disease Dataset Model

The various small sample cancer datasets are downloaded from NCBI National Centre for Bio Informatics is used in the experiments. The input data taken are various microarray binary class disease sample gene expression datasets. The details of those datasets are tabulated in table II. The data sets are Leukaemia, colon cancer, prostate cancer, pancreatic cancer and Rheumatoid Arthritis versus Osteoarthritis (RAOA), Lymphoma and Rheumatoid Arthritis versus Healthy Controls (RAHC), Type 2 Diabetes (T2D), Ovary cancer, Breast cancer and Carcinoma has balanced class ratios. This this system uses the formal test cases i.e. our key requirement is to select highly relevant genes in each of the CFS Approaches.

Table II: Gene Expression Disease Datasets

Dataset	#Samples	#Genes	#N/T_1	#D/T_2
Leukaemia [12]	72	7129	47 T_1	25 T_2
ColonCancer [12]	62	2000	22 N	40 D
ProstateCancer [13]	33	12,626	09 N	24 D
Pancreatic Cancer [14]	52	54,613	16 N	36 D
RAOA [15]	31	18,432	22 T_1	09 T_2
Lymphoma [16]	45	4026	23 T_1	22 T_2
RAHC [17]	33	4000	18 T_1	15 T_2
T2D [18]	34	19,319	17 T_1	17 T_2
OvaryCancer [19]	24	54,675	12 N	12 D
BreastCancer [19]	36	13,267	18 N	18 D
Carcinoma [19]	36	7457	18 N	18 D

The Data sets used are pre-processed with identifiers, the set of genes are identified uniquely by gene ID and the samples are identified uniquely with sample ID. The class labels is an attribute indicate the sample belonging to a disease class or a normal class [20], which are mostly used by these CFS approaches. In some datasets class-wise samples are T1/T2 & N/D in others.

7. Results and Inferences

PCC, S2N, FAST and FAIR are the four continuous gene selection methods used to select the most expressive genes. In which each gene selection methods are trained on each of the binary class small sample high dimensional gene expression disease datasets. The resultant outcome of the system is a set of highly relevant most expressive top ranked genes [20] are obtained.

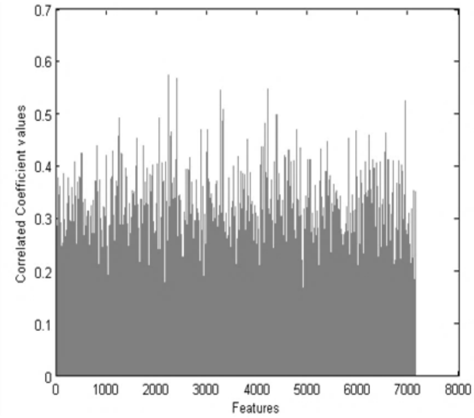


Fig. 2: PCC GS on leukaemia data

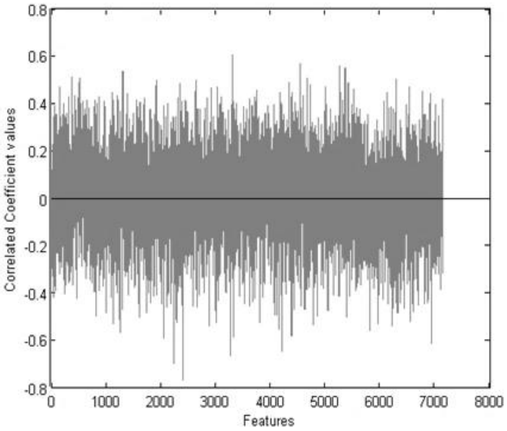


Fig. 3: S2NGS on leukaemia data

In this gene selection scheme, each gene holding their coefficient value for PCC and S2N in figure 2 & 3, and area values for FAST and FAIR in figure 4 & 5 are visualized for leukaemia data.

The figure 2, 3, 4 & 5, shows the individual gene's value for each of the approaches, the gene value is in terms of coefficient for PCC and S2N methods, where gene features are taken on X-axis and their corresponding coefficient values are taken on the Y-axis, the each genes having maximum coefficient values are taken as top gene.

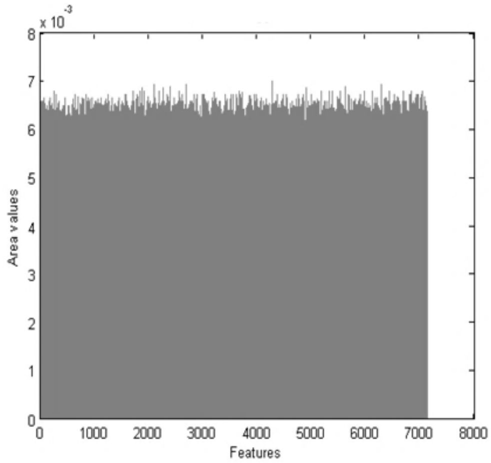


Fig. 4: FASTGS on leukaemia data

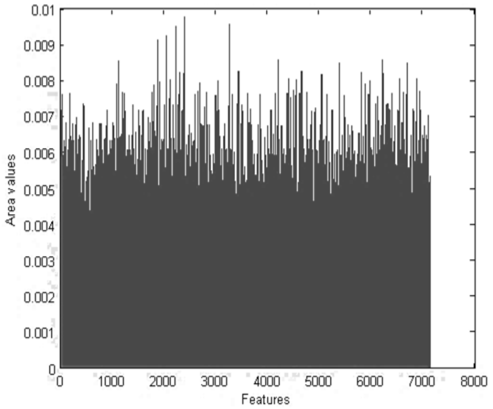


Fig. 5: FAIR GS on leukaemia data

In FAST and FAIR methods the area values are taken for each of the gene. Here the gene features are taken in X-axis the relative area values are plotted on the Y-axis for both the methods.

The top ten genes for various gene selection methods in all the disease datasets are calculated at first. By seeing the correlated coefficient value in PCC, S2N methods and the area values in FAST and FAIR methods, then ranksare assigned to each of the genes. The PCC S2N, FAST, FAIR metric's top ten genes along with Gene no, Gene ID and coefficient / area values for Leukaemia data have displayed the in table III-VI.

Table III: Most Expressive Genes Selected by PCC Metric

Gene no.	Gene ID	Coefficient value	Rank
2242	M80254_at	0.575	1
2402	M96326_rna1_at	0.568	2
4196	X17042_at	0.548	3
3258	U46751_at	0.546	4
6919	X16546_at	0.527	5
3320	U50136_rna1_at	0.509	6
4377	X62654_rna1_at	0.501	7
2056	M58603_at	0.493	8
1260	L09717_at	0.493	9
3301	U49248_at	0.486	10

Table IV: Most Expressive Genes Selected by S2N Metric

Gene no.	Gene ID	Coefficient value	Rank
3301	U49248_at	0.607	1
4535	X74262_at	0.568	2
5254	D38073_at	0.561	3
5352	M69181_at	0.553	4
1306	L13278_at	0.537	5
379	D32050_at	0.514	6
532	D63874_at	0.509	7
4661	X81372_at	0.508	8
6281	M31211_s_at	0.506	9
2641	U05237_at	0.502	10

Table V: Most Expressive Genes Selected by FAST Metric

Gene no.	Gene ID	Area value	Rank
4271	X54938_at	0.0007	1
2688	U08316_at	0.0069	2
2111	M62762_at	0.0069	3
6285	U05681_s_at	0.0069	4
2402	M96326_rna1_at	0.0069	5
4903	X99140_at	0.0069	6
2267	M81933_at	0.0069	7
1882	M27891_at	0.0069	8
5618	S79862_s_at	0.0069	9
7037	HT3061_f_at	0.0068	10

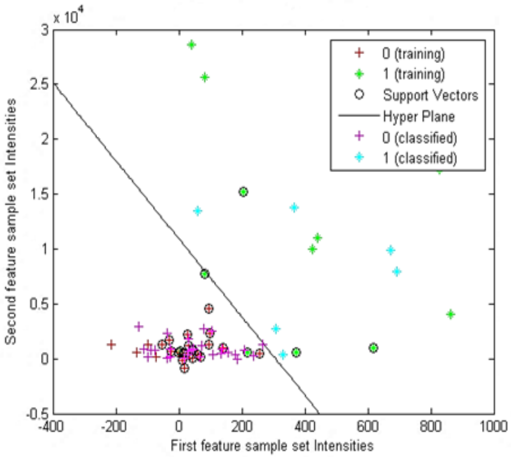
Table VI: Most Expressive Genes Selected by FAIR Metric

Gene no.	Gene ID	Area value	Rank
2402	M96326_rna1_at	0.0098	1
3258	U46751_at	0.0096	2
2242	M80254_at	0.0095	3
2056	M58603_at	0.0093	4
1883	M28209_at	0.0091	5
6215	M19508_xpt3_s_at	0.0086	6
4196	X17042_at	0.0086	7
2238	M77810_at	0.0086	8
1133	J04990_at	0.0085	9
6677	X58431_rna2_s_at	0.0085	10

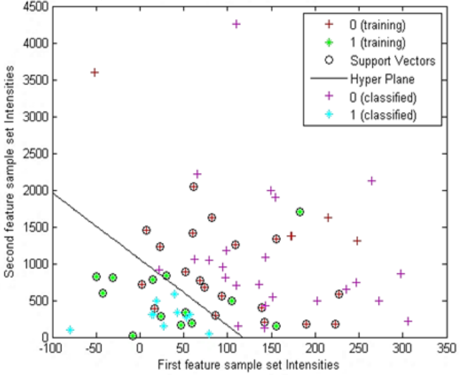
Table VII: Training & Testing Sample Size In Classification

Datasets	Classifier Sets	Sample	N/T1 Sample	D/T2 Sample
Leukaemia	Train	37	24	13
	Test	35	23	12
Colon Cancer	Train	31	20	11
	Test	31	20	11
Prostate Cancer	Train	17	12	5
	Test	16	12	4
Pancreatic Cancer	Train	26	18	8
	Test	26	18	8
RAOH	Train	16	11	5
	Test	15	11	4
Lymphoma	Train	23	12	11
	Test	22	11	11
RAHC	Train	17	9	8
	Test	16	9	7
T2D	Train	18	9	9
	Test	16	8	8
Ovary Cancer	Train	12	6	6
	Test	12	6	6
Breast Cancer	Train	18	9	9
	Test	18	9	9
Carcinoma	Train	18	9	9
	Test	18	9	9

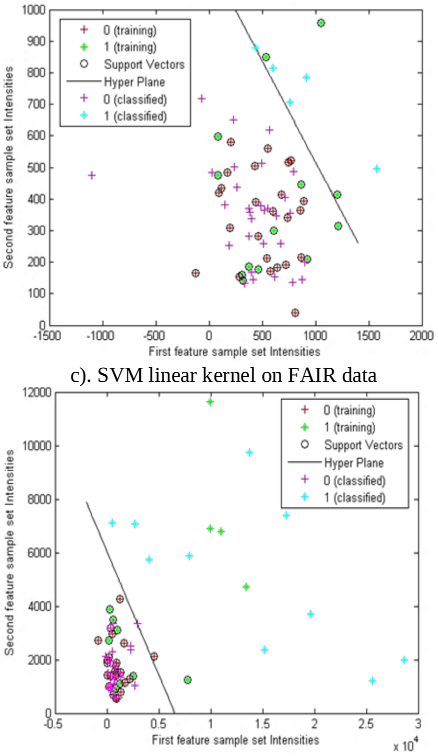
During classification task the factors which affect the classification accuracy on imbalanced test cases is the ratio of test to train sample data setup and the number of key genes selected during genes selection process. The train and test samples are distributed in 50:50 ratio based on number of samples and the class of the samples. Here, the first half is taken as training set and the next half is taken as test setas shown in table VII. And we used cross validation method for training and testing sets to obtain the maximum accuracy.



a) SVM linear kernel on PCC data



b). SVM linear kernel on S2N data



c). SVM linear kernel on FAIR data

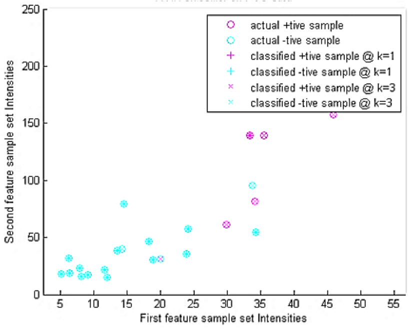
d). SVM linear kernel on FAIR data

Fig. 6: SVM Showing Classification Result on Gene Selection Methods

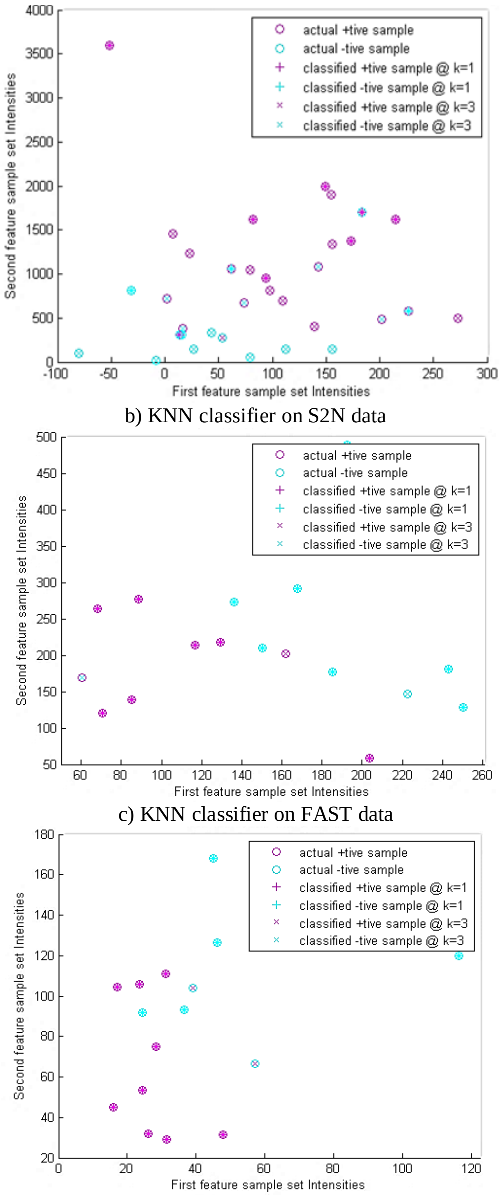
While training and testing the Leukemia data's top 10 gene feature score using SVM classifier, the obtained results are as shown in figure 6. In a gene feature space containing intensities of first feature to the intensities of next feature for the same sample set are plotted in X-axis and Y-axis respectively, during training support vectors and an optimized hyper-plane is created and classification of test samples are done based on the predicted support vectors on either side of hyper-plane as positive and negative.

More over the computed classification using K-NN classifier in the same data and the results obtained are shown in figure 7.

In a feature space containing intensities of first feature to the intensities of next feature for the same sample set are plotted in X-axis and Y-axis respectively, Here the samples are classified twice, by taking the k value 1 and 3 at each classification as the K-NN performs well at odd ranges, so the neighborhood distance is 2 in consensus approach by Euclidian method. At first iteration the k value is taken 1 and it produces certain range of unclassified samples and those samples are classified using next k value i.e. 3 thus producing refined accuracy. While classification, the second iterative clustering method can induce more accuracy by correcting the state of false classification done by previous iteration.



a) KNN classifier on PCC data



b) KNN classifier on S2N data

c) KNN classifier on FAST data

d) KNN classifieron FAIR data

Fig. 7: KNN Showing Classification Result on Gene Selection Methods

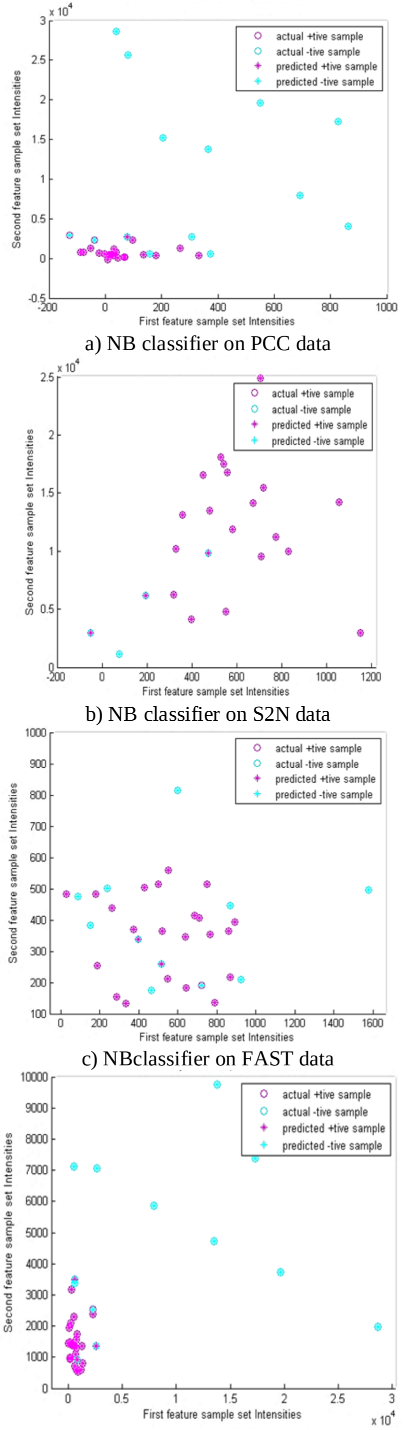
Naïve Bayes' algorithm classifier also used in classification of Leukemia data and the results obtained are shown in figure 8. In a feature space containing intensities of first feature to the intensities of next feature for the same sample set are plotted in X-axis and Y-axis respectively, here conditional probability is used for classification task, the illustration shows the actual positive and negative samples to the correctly and wrongly predicted positive and negative samples.

The performance evaluation can be done by Receiver Operating Characteristics (ROC). The performance efficiency of gene selection methods in each of the classifier is visualized using ROC graph. The ROC curve measure the overall goodness of a classifier across all possible discrimination thresholds between the two classes.

Classifiers give not only a classification for a sample, but also a quantity representing how confident the algorithm is of these results. Using the confidence values for the samples, we can calculate statistics using the discrimination threshold between each pair of samples. The ROC calculates the true positive and false positive rates as given in equations (3) & (4). The curve plotted using these points are given by equation(5).

The ROC graph [21] plotted for the gene selection methods are illustrated on figure 9. The Leukemia data is used for classification in all three classifier graphs.

Here, FPR is taken along X-axis and TPR along Y-axis. On SVM, FAST performs well by classifying samples at good accuracy followed by S2N, FAIR and PCC. On K-NN, PCC performs better and followed by FAIR, FAST and S2N. In Naïve Bayes classifier the S2N performs exceptional.



a) NB classifier on PCC data

b) NB classifier on S2N data

c) NBclassifier on FAST data

d) NB classifier on FAIR data

Fig. 8: NB Showing Classification Result on Gene Selection Methods

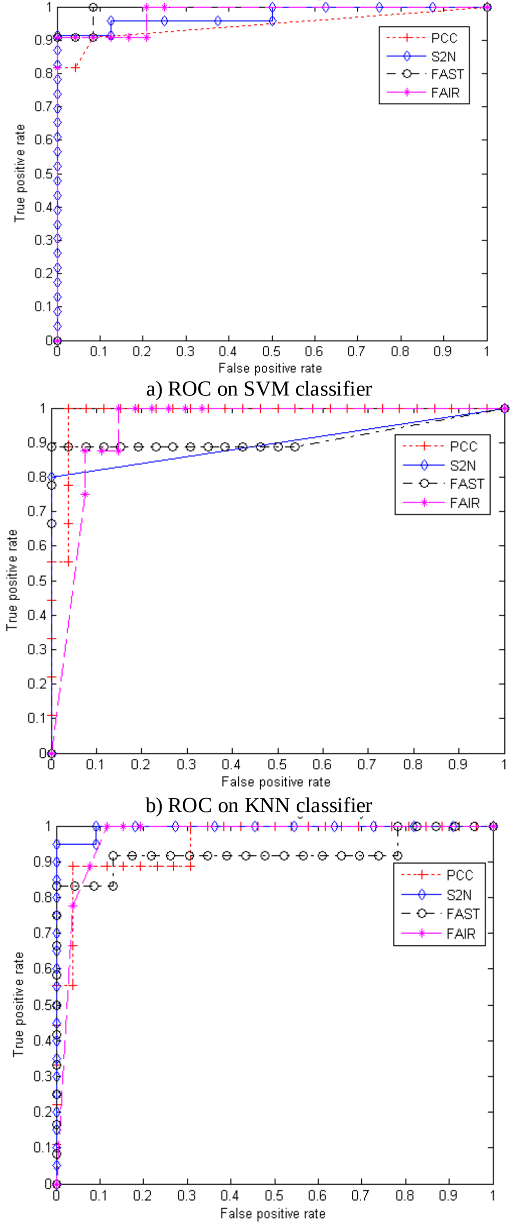
There are a lot of classifiers commonly used in machine learning, and classifiers perform differently with the exact same feature set. Thus, to measure the quality of gene selection methods, it is not sufficient to simply select one classifier. So evaluate the feature set on different classifiers with different biases to truly measure the strength of a gene selection method. According to the classification scores of the classifiers a confusion matrix is plotted as in table VIII.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (11)$$

Table VIII: Confusion Matrix

*	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

The designed classifier is then evaluated for Predictive accuracy which refers to the ability of the model to correctly predict the class label of new or previously unseen data and the classifier's accuracy is calculated using formula given in (11).



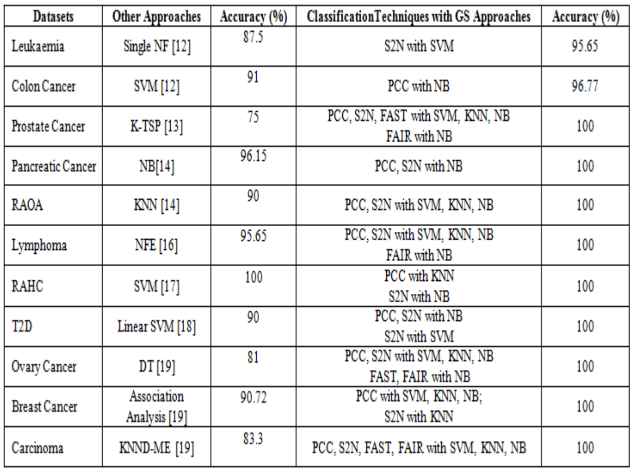
a) ROC on SVM classifier

b) ROC on KNN classifier

c) ROC on NB classifier

Fig. 9: ROC Curve Showing Classifier Performance on Leukaemia Data

Table IX: Performance Comparison with other Classification Schemes



Datasets	Other Approaches	Accuracy (%)	ClassificationTechniques with GS Approaches	Accuracy (%)
Leukaemia	Single NF [12]	87.5	S2N with SVM	95.65
Colon Cancer	SVM [12]	91	PCC with NB	96.77
Prostate Cancer	K-TSP [13]	75	PCC, S2N, FAST with SVM, KNN, NB FAIR with NB	100
Pancreatic Cancer	NB[14]	96.15	PCC, S2N with NB	100
RAOA	KNN [14]	90	PCC, S2N with SVM, KNN, NB	100
Lymphoma	NFE [16]	95.65	PCC, S2N with SVM, KNN, NB FAIR with NB	100
RAHC	SVM [17]	100	PCC with KNN S2N with NB	100
T2D	Linear SVM [18]	90	PCC, S2N with NB S2N with SVM	100
Ovary Cancer	DT [19]	81	PCC, S2N with SVM, KNN, NB FAST, FAIR with NB	100
Breast Cancer	Association Analysis [19]	90.72	PCC with SVM, KNN, NB; S2N with KNN	100
Carcinoma	KNND-ME [19]	83.3	PCC, S2N, FAST, FAIR with SVM, KNN, NB	100

The details of the simulation carried out on ten gene expression datasets on accuracy using other approaches and the efficient proposed approach are compared. Here the various datasets operate on various gene selection methods to the classifier approaches and the best suited gene selection-classifier approach are found for the disease sample. Thus the compared results are shown as tabulation IX. [22]

In leukemia dataset the S2N method perform well with SVM classifier, while in colon and pancreatic cancer PCC–NB gene selection classifier approach performed well. When taking prostate cancer dataset, all three classifiers did well on PCC, S2N & FAST. While taking lymphoma and RAOH datasets PCC and S2N did well in all three classifier. In RAHC data PCC performed better in KNN and S2N did well in NB. PCC, S2N did well with NB classifier in Type 2 Diabetes dataset. While considering the ovary cancer data all three classifiers done well with PCC and S2N methods. In breast cancer data, PCC had done perfect with all three classifiers. All the gene selection methods has given an one hundred percent accuracy with respect to all classifiers in carcinoma dataset.

8. Conclusion

The goal to find the highly expressed genes by effective gene selection methods is developed. Thus the evaluation technique helps users to learn appropriate genomic datasets and has retrieved highly relevant disease causing genes. By this system one can also easily compare the observation of various approaches using any dataset. The gene selection methods selects highly optimized genes and the efficient sample classifiers classify samples with high accuracy and provide highly reliable machine learning tasks done.

References

[1] Fang OH, Mustapha N & Nasir Sulaiman MD, "Integrating Biological Information for Feature Selection in Microarray Data Classification", *IEEE Computer Society, IEEE Conference on Computer Engineering and Applications*, Vol.2, (2010), pp.330-334.

[2] Osareh A & Shadgar B, "Microarray Data Analysis for Cancer Classification", *IEEE Conference on Computer Engineering and Applications*, (2010), pp.125-132.

[3] Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L & Brown P, "Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns", *Genome biology*, Vol.1, No.2, (2000).

[4] Saeys Y, Inza I & Larranaga P, "Review feature selection technique bioinformatics", *Bioinformatics*, Vol.23, No.19, (2007), pp.2507-2517.

[5] Maji P & Pal SK, "Fuzzy Rough sets for information measures and selection of relevant genes from microarray data", *IEEE Transaction on Systems, Man, and Cybernetics*, Vol.40, No.3, (2010), pp.741-752.

[6] Jose CHH, B´eatrice D & Jin KH, "A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data", *Springer*, (2007), pp.90-101.

[7] Wasikowski M & Chen X, "Combating the small class imbalance problem using feature selection", *IEEE Trans. Knowledge and Data Engineering*, Vol.22, No.10, (2010), pp.1388-1400.

[8] Davis J & Goadrich M, "The Relationship between Precision-Recall and ROC Curves", *23rd Int'l Conf. Machine Learning*, (2006), pp.30-38.

[9] Chen X & Wasikowski, "FAST: A ROC-Based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems", *Proc. ACM SIGKDD*, (2008), pp.124-133.

[10] Ganeshkumar P, Aruldoss T, Devaraj D & Renukadevi M, "Design of fuzzy Expert system for microarray data classification using a novel Genetic Swarm Algorithm", *Expert Systems with Applications*, Vol.39, (2012), pp.1811-1821.

[11] Maji P, "Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data", *IEEE Transaction on Systems, Man and Cybernetics*, (2010), pp.1-10.

[12] Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Dowing J, Caligiuri M, Bloomfield C & Lander E, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, Vol.286, (1999), pp.531-537.

[13] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D & Levine A.J, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc.Nat. Acad. Sci. U.S.A.*, Vol.96, No.12, (1999), pp.6745-6750.

[14] Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J & Moskaluk CA, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate Cancer", *Cancer Research*, Vol.61, (2001), pp.5974–5978.

[15] Hayward J, Alvarez SA, Ruiz C, Sullivan M, Tseng J & Whalen G, "Machine learning of clinical performance in pancreatic cancer database", *Artificial Intelligence in Medicine*, Vol.49, No.3, (2010), pp.187-193.

[16] Kraan TCTM, Gaalen VFA, Kasperkovitz PV, Verbeet NL, Smeets TJM, Kraan MC, Fero M, Tak PP, Huizinga TWJ, Pieterman E, Breedveld FC, Breedveld AA, Alizadech AA & Verweij CL, "Rheumatoid arthritis is a heterogenous disease: Evidence for differences in activation of STAT-1 pathway between rheumatoid tissues", *Arthritis Rheum.*, Vol.48, No.8, (2003), pp.2312-2145.

[17] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A & Powell JI, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, Vol.403, No.6769, (2000), pp.503-511.

[18] Teixeira VH, Olaso R, Martin-Magniette ML, Lasbleiz S, Jacq L, Oliveira CR & Petit-Teixeira E, "Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients", *PloS one*, Vol.4, No.8, (2009), pp.e6803.

[19] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J and Houstis, N, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes", *Nature genetics*, Vol.34, No.3, (2003), pp.267-273.

[20] National Centre for Biotechnology Information (NCBI), U.S. National Library of Medicine, Available Online at http://www.ncbi.nlm.nih.gov, 2009.

[21] Hayward J, Alvarez SA, Ruiz C, Sullivan M, Tseng J & Whalen G, "Knowledge discovery in clinical performance of cancer patients", IEEE International conference on Bio-Informatics and Bio-Medicine, Vol.49, No.3, (2010), pp.187-193.

[22] Villalobos Antúnez, JV (2017). Karl R. Popper, Heráclito y la invención del logos. Un contexto para la Filosofía de las Ciencias Sociales. Opción Vol. 33, Núm. 84. 5-11