Gene Selection using a Hybrid RFE Along with LASSO for Cancer Classification

M. J. Abinash, V. Vasudevan

Abstract*: Gene expression profiling using microarray technology has done with the chip based phenomena. For studying gene expression data are more helpful in knowing various diseases and more useful in finding diseases. Recently in the bioinformatics field, cancer prediction using gene expression data had made the assuring area. Samples having the gene attributes will not surely give the efficient amount of classification. Overcoming these contribution, a strong method is required for selecting the relevant gene features for building the classification model effectively. Basically least absolute shrinkage and selection operator (LASSO) and Recursive feature elimination (RFE) are automatic gene feature selection methods used for classification. Here in our proposed work, we use these two methods as a hybrid one for selecting the features and later it applied into the Support vector machine (SVM) for easy classification. It made best when compared to the existing techniques by their performance measures, were regulated on six publically available cancer datasets. Just out it gives the good awareness in the selection of features.

Keywords* : *LASSO, Gene selection, RFE, SVM, Cancer Classification.

I. INTRODUCTION

Recently, the microarray technology is used for measuring the abundant amount of genes arranged in a single chip. It helpful in measuring the genes with expression values. These gene expression levels indicate the varying number of gene Ribo Nucleic Acid (RNA). For analyzing these genes expression is normally difficult. It causes redundancy. Gene expression measure is the transfer of DNA into RNA protein molecule, while using the microarray technology causes the dimensionality issue for measuring the genes [1]. For some classification problems using gene expression profiling it is hard to use the traditional methods straightly. It leads to the dimensionality issue. So we need to use the feature selection it reduces the redundancy and the dimensionality. Towards it reduces the redundancy level of the gene expression data [2, 3]. The statistical learning methods such as SVM was first charted by Vapnik et.al, in 1960 mainly for classifying the data. SVM classifies the large datasets which will be in the linear or non-linear, by constructing the hyperplane. By separating the data in two subsets among these data closer to the hyperplane are called as support vectors. And draw a margin between the hyperplane where the good separation of data is attained by greatest distance named as functional margin for non-linear kernel methods are also applicable [4]. Cancer classification is made using the genetic programming by the gene expression data [5]. L.Wang et al. and Statnikov A., et al. [6, 7] proposed the classification, using support vector machines done in two phases i.e., gene selection and classification. An experiment has been handled on human acute leukemia's for cancer classification with the help of gene expression microarrays to finding the class discovery acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) for predicting and finding the classes [8]. Using the three feature selection methods filter, wrapper and embedded are used in classifying the predictive accuracy with the help of the significant features [9].By using support vector machine, the tissue samples of cancer diseases are validated[10].RFE-SVM is used for choosing the suited features and later the given datasets are changed to two subsets for further classification. And then the result validation is carried out to get the goodness of the features. Finally, RFE-SVM yields the better results [11]. Two stages of SVM-RFE method are used for feature selection because to avoid inconsistency. The noisy, irrelevant genes are eliminated in the first stage by the pre-filtering process. And the elimination of the single gene is carried in each step is occurred in the next stage. Later classification is done by the linear SVM and this method is competed along with the correlation method [12]. The functional margin that maximizes the hyper-plane for classification using support vector machine [13]. Tibshirani et al. proposed the method for regularization and the selection of variables for linear regression called LASSO [14]. Gene selection and class prediction by SLR and SLRB, among the high dimensional of genes most of them are most of them are noisy, irrelevant. Later the deficiency of LASSO causes consistency [15]. An automatic gene selection method is carried with Dantzig selector and LASSO technique, this method is more supportive in choosing the significant genes with the linear regression method [16]. Here in our work we have to suggest a newfangled method for selecting features as hybrid recursive feature elimination (RFE) with least absolute shrinkage and selection operator (LASSO) and later classification is done using SVM. This feature selection method yields the better one compare to the other available methods. The surplus of the work is prepared with gene selection method, recommended work, results and discussions append with conclusion and future work.

Revised Manuscript Received on December 16, 2019.
***** Correspondence Author
Abinash*, Department of Information Technology, Kalasalingam Academy of Research and Education, Srivilliputhur, India. Email: mj.abinash@gmail.com
Vasudevan, Department of Information Technology, Kalasalingam Academy of Research and Education, Srivilliputhur, India. Email: vasudevan.klu@yahoo.co.in

II. FEATURE SELECTION METHODS

A. RFE

RFE is comes under the wrapper based feature selection method [13], by the sequential backward elimination the waited features are selected, among all the features it removes a single feature. The feature with good ranks are selected and it removes the unimportant feature. For every step the weight vector of coefficients are helps to calculate the feature score in every step of the linear SVM. In SVM – RFE the objective function k will enhance in changing the removal of insignificant genes.

$$k = \|w\|^2 / 2$$

B. LASSO

Tibshirani et al. [14] proposed the regression method called LASSO, it performs selection of variables and the regularization of variables for linear regression. It correlates to ordinary least squares (OLS) with residual sum of squares (RSS) that reduces the sum where the constant value is less while compared with the absolute values. And as an additional constraint, it is slightly similar to ridge regression with the sum of squared values as coefficients to the larger constraint. So the simple modification occurred during the LASSO in the variable selection. It set the coefficients values to zero during the variable selection, it automatically punishes the extra features shrunk to zero entirely.

$$Lasso = \min\left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Where β_j denotes full least square estimates

x_{ij} denotes predictor variables

y_i denotes output response

β_0 is usually '0' & not necessary, will be neglected

III. PROPOSED WORK

A. Hybrid RFE Based LASSO

In association with the combination of SVM-RFE and LASSO is the new approach. Better performance gives by SVM-RFE on classification but it gives the poor performance in the repetitions of class labels. In concord with the LASSO it evades the redundancy level. The combination of SVM-RFE with LASSO will attain their good in performance. The figure 1 displays the proposed work carried in this scenario. Using RFE the significant features are selected by calculated with the threshold value and the irrelevant features get eliminated and later the selected features get regularized and the coefficients value are set to zero automatically during the selection of good features by LASSO. Among them, that 70% features are carried for training during classification and the remaining 30% are used for testing under the evaluation process in classification.

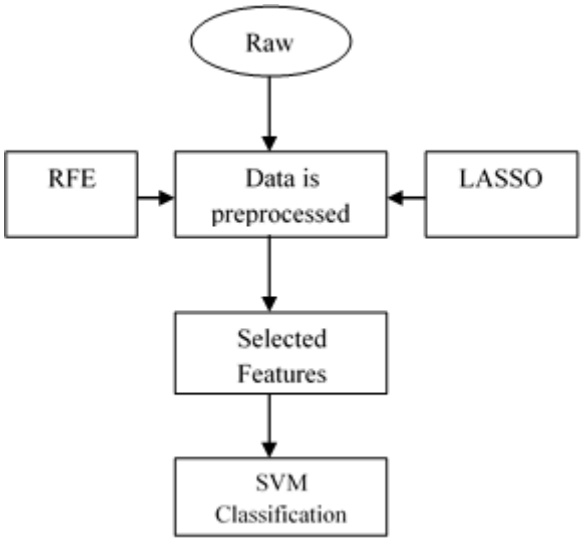

Fig. 1. Proposed methodology for HRFEBL

B. Procedure for Feature Selection

The procedure for feature selection method comes as,

Input: Consider a Scored Set K= [], Begin with feature member T= [1….f]

Repeat with all features get scored in subset

Variables get added from scored set K

Output: Minimize Set

Select with all features which get scored

Compare with the threshold value

Eliminate the irrelevant features which is less than threshold value

And select the significant features

Features with the smallest score get identified

By using the coefficient value

It gets regularize

Calculate the value with the fewer coefficients

It automatically punishes the extra features

To zero

And eliminate it

Repeat Steps

Until it gets good features

Update

Minimize set

C. SVM Classification

After selection of features some set is used for training and some is used for testing. Support vector machine (SVM) is a classification technique [4, 13]. This technique is mainly used to perform both classification and regression tasks. It divides the data into two subsets using hyperplane. And the data points that closer to the hyperplane are denoted as support vectors. And draw a margin on the support vectors and that nearer to the hyperplane and later calculates the distance between the two support vectors, where it comes under. And checks for linearity and if the non- state occurs means, moves to the kernel function. By placing more number of planes on the data is leads to the better classification. The plane attains the largest in their measurement for the closest data is called as the functional margin. This margin is fussy means the error rate is less in the classifier.

Here the table 1 shows the dataset description.

Table- I: Dataset Description

Datasets	Number of samples	Number of Genes	Class
Leukemia[8]	72	7129	ALL
			AML
Colon[17]	62	2000	Tumor
			Normal
Lung Cancer[18]	181	12533	MPM
			ADCA
Breast[19]	97	24481	Tumor
			Normal
Prostate[20]	136	12600	Tumor
			Normal
Liver[21]	156	1648	Non-Tumor
			HCC's

IV. RESULTS AND DISCUSSION

Here, we are using six publically available datasets for summarization mentioned in table1.For our approach, the performance metrics is checked using precision, recall, classification accuracy. This evaluation measure, gives the better outcomes in finding the results. Here, precision denotes the positive predictor value. Recall is called as sensitivity in binary classification. Kappa statistics is the performance measure is used to check the interrelationship between the two variables whether it having good agreement or less agreement. And the classification accuracy is measured by the absolute number of the checked features divided by the total number of features. Below equations (1, 2, 3) as,

$$\text{Re}call = TP/(TP+FN)$$

$$\text{Pr}ecision = TP/(TP+FP)$$

$$Accuracy = \frac{Total_{observed}(prediction)}{Total_{entired}(prediction)}$$

Table- II: Comparative performance measure for RFE and LASSO

Algorithm	Precision	Recall	Kappa statistics	Feature Selection accuracy
RFE and BBF	0.95	0.93	0.95	92.55
SFS and LASSO	0.96	0.95	0.96	94.45
RFE and LASSO	0.97	0.96	0.97	97.61

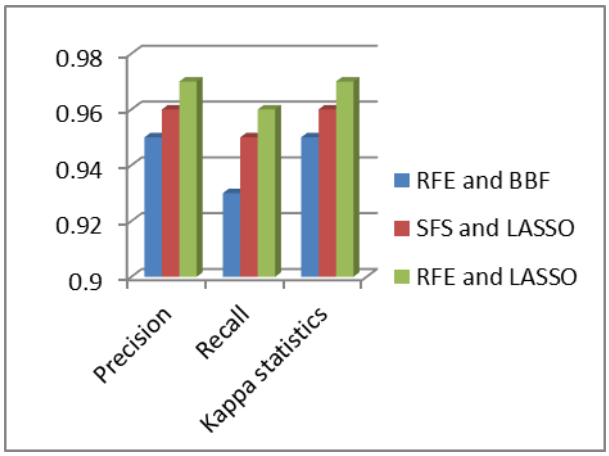

Fig. 2. Graphical representation for RFE and LASSO

The proposed work is analyzed using R programming language. It is a statistical programming language. Recently, much statistical analysis based on machine learning, deep learning techniques are evaluated using this language. The classification accuracy along with conventional SVM using feature selection method SVM-RFE and LASSO have attained the better results as 97.61%. Likewise, 70% of selected features are used for training with the classifier and remaining 30% is used for testing the classifier model. The results obtained are observed with feature selection methods SVM-RFE and BBF, SFS and LASSO using the conventional SVM. The table 2 denotes the comparative feature selection accuracy for RFE and LASSO. And fig 2. shows the graphical representation for that.

Table- III: Comparative feature selection accuracy for RFE and LASSO

Algorithm	Feature Selection accuracy
RFE and BBF	88.55
SFS and LASSO	95.45
RFE and LASSO	97.61

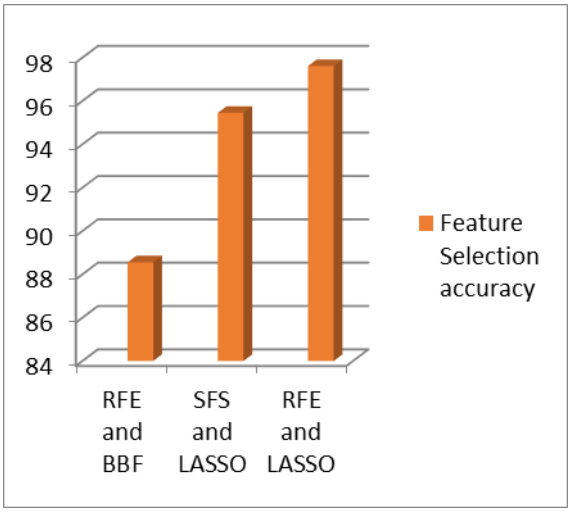

Fig. 3. Graphical representation for RFE and LASSO feature selection method

Table- IV: Classification model performance with different datasets

Classifier Model	Data Sets					
	Leukemia	Colon	Lung Cancer	Breast	Prostate	Liver
ANN	62	51	65	72	75	63
KNN	74	63	85	87	62	83
SVM	92	86	94	95	81	92

Compared with the existing methods RFE and BBF, SFS and LASSO table 3 shows the feature selection accuracy results of the RFE and LASSO have attained 97.61%. Fig 3. denotes the graphical representation for that. Along with that, the training and testing done with the classifier model such as ANN, KNN and SVM among these SVM will attained the better results for six datasets as laid out in table 4. With their graph representation in figure 4.

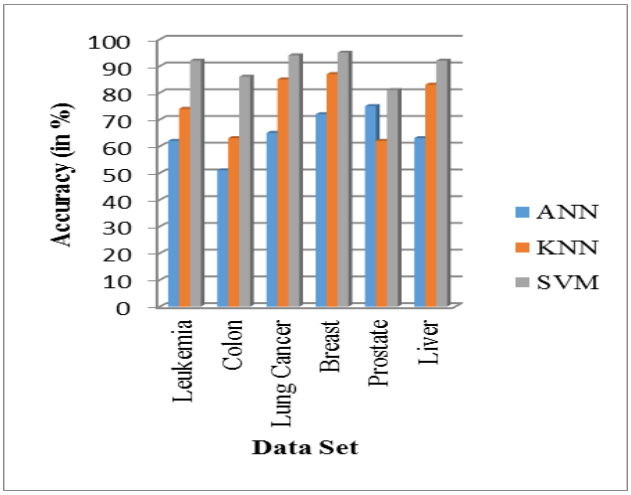

Fig. 4. Graphical representation for different datasets with classifier model

V. CONCLUSION

Cancer based classification using gene expression data is an assuring research recently. Here in our paper, a compound approach using the SVM-RFE and LASSO is used for variable selection. After that, this method combined with SVM for classification. This is the novel approach had applied in the six datasets. And later it checks for the precision, recall and classification accuracy. This work results is compared with other classifier models such as ANN, KNN. By using this approach, it reduces the redundancy level. In future, this approach is compared with the other techniques such as deep learning, reinforcement learning using large dimensional tools such as Hadoop, Spark.

ACKNOWLEDGMENT

The authors would like to thank the Department of Information Technology of Kalasalingam Academy of Research and Education, Tamilnadu, India for permitting to use the computational facilities available in open source research laboratory.

REFERENCES

1. Y. L, and J. Han, Cancer classification using gene expression data. *Information Systems*, vol. *28*(4), 2003, pp.243-268.
2. M.Al-Rajab,J. Lu and Q.Xu, Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. *Computer methods and programs in biomedicine*, vol. *146*, 2017, pp. 11-24.
3. S.M.Ayyad,A.I.Saleh and L.M.Labib, Classification Techniques in Gene Expression Microarray Data. *International Journal of Computer Science and Mobile Computing*, vol. *7*(11), 2018, pp. 52-56.
4. V.Vapnik and V.Vapnik, Statistical learning theory Wiley. *New York*, 1998, pp. 156-160.
5. K.Yamunadevi, and R.Nagaraj, An optimized classification of human cancer disease for gene expression data. *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. *4*(2), 2018, pp. 8-15.
6. L. Wang, J. Zhu and H.Zou, Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, vol. *24*(3), 2008, pp. 412-419.
7. A.Statnikov, L. Wang and C.F.Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, vol. *9*(1), 2008, p.319.
8. T.R.Golub,D.K.Slonim,P.Tamayo,C. Huard, M.Gaasenbeek,J.P.Mesirov, H.Coller, M.L.Loh, J.R. Downing, M.A.Caligiuri and C.D. Bloomfield, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, vol. *286*(5439), 1999, pp. 531-537.
9. A.I.Saleh,A.H. Rabi and K.M. Abo-Al-Ez, A data mining based load forecasting strategy for smart electrical grids. *Advanced Engineering Informatics*, vol. *30*(3), 2016, pp. 422-448.
10. T.S.Furey,N.Cristianini, N. Duffy,D.W.Bednarski, M.Schummer and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, vol. *16*(10), 2000, pp. 906-914.
11. M.L.Samb,F.Camara,S.Ndiaye,Y.Slimani and M.A.Esseghir, A novel RFE-SVM-based feature selection approach for classification. *International Journal of Advanced Science and Technology*, vol. *43*(1), 2012, pp. 27-36.
12. Y. Tang,Y.Q. Zhang and Z. Huang, Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. *4*(3), 2007, pp. 365-381.
13. I.Guyon,J. Weston,S. Barnhill and V.Vapnik, Gene selection for cancer classification using support vector machines. *Machine learning*, vol. *46*(1-3), 2002, pp. 389-422.
14. R.Tibshirani, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. *58*(1), 1996, pp. 267-288.
15. Z.Y.Algamal and M.H. Lee, Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, vol. *42*(23), 2015, pp. 9326-9332.
16. S.Zheng and W. Liu, An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification. *Computers in biology and medicine*, vol. *41*(11), 2011, pp. 1033-1040.
17. U.Alon, N.Barkai,D.A.Notterman,K. Gish,S. Ybarra,D. Mack and A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, vol. *96*(12), 1999, pp. 6745-6750.
18. D.A.Wigle,I.Jurisica,N.Radulovich,M.Pintilie, J.Rossant,N. Liu,C. Lu,J.Woodgett,I.Seiden, M. Johnston and S.Keshavjee, Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, vol. *62*(11), 2002, pp. 3005-3008.
19. LJVan't Veer,H Dai,MJ Van de Vijver,YD He, AAHart,M Mao,HLPeterse,K van der Kooy,KK van der,MJMarton,ATWitteveen,GJ Schreiber, RMKerkhoven,C Roberts,PSLinsley,RBernards, SHFriend. Gene expression profiling predicts clinical outcome of breast cancer. Nature. Vol. 415, 2002; pp. 530-536.
20. D. Singh, P.G.Febbo,K. Ross,D.G. Jackson, J.Manola,C. Ladd,P. Tamayo,A.A.Renshaw, A.V. D'Amico,J.P. Richie and E.S. Lander, Gene expression correlates of clinical prostate cancerbehavior. *Cancer cell*, vol. *1*(2), 2002, pp. 203-209.
21. Received FromGene expression patterns in liver liver cancers. http://genome-www.stanford.edu/hcc/

AUTHORS PROFILE

Abinash he is a full-time research scholar in the department of information technology of Kalasalingam Academy of Research and Education Krishnankoil Srivilliputhur India. His areas of interest are big data analytics, bio informatics and block chain technology. He has published 5 papers in international journals and conferences.

Vasudevan he is working as senior professor in the department of information technology of Kalasalingam Academy of Research and Education Krishnankoil Srivilliputhur India for past 27 years and his areas of interest are big data analytics, cloud computing, network security and block chain technology. He has published 170 above papers in international journals and conferences. He has produced twenty above PhD scholars and currently guiding five PhD scholars. He is a life time member in ISTE and IAENG. He received Dr. APJ Abdul Kalam Award for life time contribution in teaching on 2016.