International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 105-109
Published Online June 2019 in IJEAST (http://www.ijeast.com)

MULTI FILTER ENSEMBLE METHOD FOR CANCER PROGNOSIS AND DIAGNOSIS

Bibhuprsad Sahu
Department of CSE
Gandhi institute For Technology,
Bhubaneswar, Odisha, India

***Abstract*— The nature of the gene expression profiles are high dimension, very small sample size, continuous types so it is really a challenged task to achieve good classification accuracy from the tumor samples. The main aim of feature selection is to find out most relevant features, which may increase the computational speed and accuracy. We thus proposed a multi filter ensemble based hybrid gene selection method. Here we have used four filter methods such as Information Gain (IG), Gain ratio (GR), Relief, Correlation to filter the irreverent and redundant genes. By the help of computationally efficient filters candidate features are selected .The top N genes with highest rank of individual subset are integrated to produce a new dataset. Then SVM attribute evaluator is applied for attribute evaluation purpose. Finally LIBSVM classifier is used to detect the best feature subset. This experimental result proves the proposed method is quite efficient then other gene selection methods and it provide a high accuracy under some characteristics genes.**

Keywords— **IG, GR, CORRELATION, ReliefF, LIBSVM**

I. INTRODUCTION

Due to availability of large scale genes expression profiles, for disease diagnosis DNA chips are used to represent expression level of millions of gene in the a single sample for single experiment .To achieve a great significant level of diagnosis for cancer treatment, molecular level accurate classification is needed. The size of tumor gene expression data is high dimensional and small size in nature[1-4].As the gene expression profiles consists of high redundancy and noise data ,so choosing a better filter approach to minimize the computational time with good classification accuracy[5-6].Each and every sample consists million of genes but from these few genes are really responsible for sample classification .To train an effective classification model one should have a clear idea how to filter the group of related genes form the huge dataset, That is known as coarse of dimensionality problem[7].

While implementing subset feature selection it selects the all attributes of the gene expression profile data [8] and from this it identify the strong attributes to identify the disease. Gene selection methods are generally two types such as filter and wrapper [9, 10]. For gene selection the filter methods uses feature ranking approaches. The filter methods are of two different types such as univariate and multivariate. The Uni-variate methods are based on single criterion principle nothing to do with the feature selection process such as SNR [11], TS [12], FT [13], PC [14].By the help of filter methods we have consider the top ranked genes as biomarker genes. In this research we have used Information Gain (IG) [15], Gain ratio (GR), ReliefF [16], Correlation as filter methods.

Filter method for feature selection depends on the individual vector data and final subset selection is independent of classifiers. According to the basic principle of a classifier the wrapper methods detects the biomarker gene datasets. Selection of feature subset done by different machine learning algorithm [17-19],the next evaluation is done by the classifiers[20].The computational speed of filter methods are fast as compared to wrapper but the relation between genes are not established by the filter [21].To achieve good accuracy various hybrid methods has been proposed by different researchers[22-26].

In this research study, the overall proposed model is divided into 3 stages such as filter gene selection stage, combinational stage, classification stage. In filter gene selection stage is used to identify and remove redundant and irrelevant genes. IG, GR, ReliefF, Correlation are chosen as ranker algorithm .These four resulted top ranked features are combined together for further fine tuning purpose. This is designed at combinational stage. Later stage using SVM filtration and reduction of redundant genes performed. At last LIBSVM classifier is used to obtain the informative genes through classification.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 105-109
Published Online June 2019 in IJEAST (http://www.ijeast.com)

II. HYBRID FEATURE SELECTION

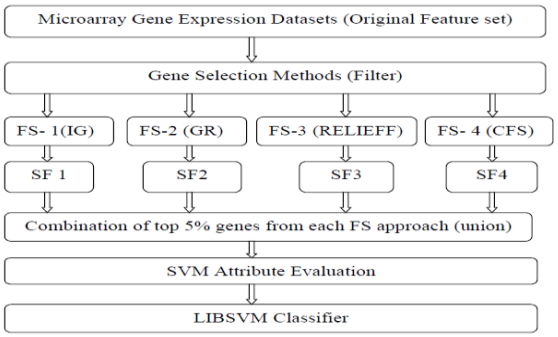

Fig.1. Work Flow of Purposed Model

The purposed model is basically divided into 4 stages such as filter genes selection stage, Combinational stage, SVM attribute selection stage, Classification stage. The discussion of individual stage is done below.

A. Original dataset –

For this research experimental study, we have considered five bench marked microarray datasets. These five datasets are having two class such as normal and cancerous. Datasets chosen are leukemia, prostate cancer, Colon cancer, DLBCL, lung cancer. We have used AML and ALL (Leukemia) type where as the DLBCL dataset of two sub groups i.e. germinal center B-cell like subgroup, activated B-cell subgroup. The details of gene expression dataset are mentioned below.

Dataset Name	No of Genes	Sample count	Positive Samples	Negative Samples	No of Classes
Leukemia	7129	72	25	47	2
DLBCL	4026	47	24	23	2
Colon	2000	62	40	22	2
Lung	7129	96	86	10	2
Prostate	12600	102	52	50	2

Table 1. Details of considered Gene expression Datasets.

B Filter gene selection stage-

In this model we have considered four different filter methods such as IG, GR, ReliefF and Correlation for feature selection in high dimensional data.

Information Gain (IG)

To indentify top ranked genes from pool of genes many researchers used information gain (IG) method. Help of this approach the difference between entropy and conditional entropy is of genes are gathered [32, 33].

$$g(Y,X) = H(Y) - H\left(\frac{Y}{X}\right)$$

H(Y) - Entropy of gene dataset Y (Ratio of uncertainty to predict the random variable)
H(Y/X)-Conditional Entropy (uncertainty with respect to the variable X)

$$H(Y) = -\sum p(Y)\log p(Y) \quad (2)$$

$$H(Y/X) = \sum_{x \in X} P(X) H(Y/X = \quad (3)$$

The rank order of individual gene is evaluated and from that high ranked feature genes are selected.

Gain Ratio (GR)

Information gain is suitable to identify the attributes having large no of values.C4.5 enhances the decision tree induction algorithm ID3. C4.5 is called gain ratio.
Consider A is the set contains data sample with n different classes. So the information expected for classification is evaluated as

$$R(A) = - \sum p_i \log_2 p_i \quad (4)$$

p_i Represents probability of sample c_i and calculated by D/Di. Let us consider B has Z distinct values. Let D_{ij} is the no of samples and C_i is the subset of subset D_j.

$$E(B) = \sum I(A) \ (D_{1j} + D_{2j} + \cdots + D_{nj})/D \quad _{n\ i=1} \quad (5)$$

So Gain (A) =R (A)-E (B) (6)
Gain Ratio (B) =Gain (A)/Spilt info$_A$(A)

ReliefF

Instance based learning used by ReliefF to assign a weight to individual features. The weight of each helps to identify each from its different class values. According to the nearest hit /nearest miss (same instance class /opposite instance class) the weights of the features get updated. If it differentiates between inter classes then it receives high weight but in intra it maintains a same weight value. The value for nominal features the value defines either 1 or 0. 1 indicates values are different and 0 indicates values are same. So the nominal value of a feature varies between 0 & 1.

$$\text{Weight}_{\text{Feature}} = \text{Prob}(\text{Prob}(\text{different value of features}//\text{different class}) - \text{Prob}((\text{different value c})$$

CFS

In 1999 Mark developed correlation based feature selection by focusing the relationship exists between individual attribute with respect to class. In CFS it indentifies high correlated dataset with its respected class by ignoring relation between them. By the feature selection techniques, irrelevant features are reduced and the high correlated genes can be indentified for classification.

C Combinational Model

When the filter gene selection method is completed, high ranked feature subset are selected by IG, GR, ReliefF, Correlation feature selection techniques. The selected features

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 105-109
Published Online June 2019 in IJEAST (http://www.ijeast.com)

are taken as the top ranked ones as compared to other features. Not only considering all selected features for attribute evaluation and classification, may increase the computational cost not may it provide better classification accuracy. So to avoid redundant we again apply filter approach to enhance the classification accuracy.

D Attribute selection model and classification model

In the attribute selection model we have implemented SVM attribute selection evaluator. Addition to IG, GR, ReliefF, Correlation is combined with SVM were implemented to select the biomarker genes. In classification model the further classification of genes are done using LIBSVM classifier.

III. ALGORITHM OF THE PURPOSED MODEL

Input:
DS [FS1, FS2, FS3, FS4…..]: n>=4
Ns <- Total no of feature in the subset.
Algorithm flow:

1. nFS<- ns // initialization of n feature
2. DS_{cv}<- 10 CV (DS) // 10 cross validations
3. R_{IG}=eval (IG, DS_{cv}) // Apply IG 10 cross validations
4. R_{GR}= eval (GR, DS_{cv}) // Apply GR 10 cross validations
5. R_{RL}= eval (RL, DS_{cv}) // Apply Relief 10 cross validations
6. R_{CFS}= eval (CFS, DS_{cv}) // Apply CFS 10 cross validations
7. R_{mean}=(R_{IG}+ R_{GR}+ R_{RL}+ R_{CFS})/4 // Average feature positions
8. $R_{R\ mean}$=eval (R_{mean} ,R DS_{cv}) // R_{mean} ranking
9. Ranked Feature List (RFL)= [R_{IG} ,R_{GR} ,R_{RL} ,R_{CFS} , R_{mean}) // sorting Ascending order
10. For I =1 to RFL
11. For J=1 to trunc(RFL_I /ns)
12. Do
13. $S_{top\ ranked\ genes}$= extract (nF/ RFL_I)
14. $S_{top\ ranked\ genes}$=10 CV ($S_{top\ ranked\ genes}$)
15. $C_{(I, j, SVM)}$=eval (SVM,S_{cv})
16. nFS <- NF+ns // feature update
17. nFS
18. end of for Loop
19. C (aux)= C (aux)+max(C)
20. End of second for
21. C(biomarker genes)=max (C (aux)) // biomarker gene with high statistical

This proposed model is supported by combination of 4 filter selection methods and machine learning classifiers (LIBSVM) for building a classification scheme that provide high performance. In this proposed model we have used IG, GR, ReliefF and Correlation for preliminary gene selection .For this we have used different attribute selection Evaluation and ranker selection tools available in Weka. The attribute evaluations evaluates clinical genes relevant to outcome based according to the filter methods .Then ranker ranks each genes on the basis of evaluation outcomes.

To indentify gens with high classification value we have considered filter and wrapper methods. The top ranked gene pool was created by considering top 5 percent ranked genes .With addition to filters ,SVM was employed as feature selection algorithm. At the end of the informative genes detected after SVM served as a data input for LIBSVM classifier to achieve classification accuracy. As the gene expression datasets are small in nature, for achieving high accuracy 10 cross validation was utilized. The above process was explained in the algorithm section.

IV. EXPERIMENT AND RESULT

Dataset Name	Proposed Method	IGSVM	GRSVM	ReliefFSVM	CORSVM
Leukemia	99.52	98.61	94.44	97.22	97.22
DLBCL	100	100	97.87	95.74	100
Colon	91.26	90.32	85.48	87.1	87.1
Lung	100	100	98.96	98.96	98.96
Prostate	97.32	96.08	93.14	91.18	93.14

Table 2.Accuracy comparison

For the experimental analysis we have considered total 150 top most ranked genes from four filters by thinking that small no of genes are enough to produce high accuracy and the same is used as input for the SVM. After filter we understood that the top most genes are identified by different filter methods are almost similar, so top 5 genes are considered for classification purpose.

Table 2 represents the classification accuracy of different four hybrid methods individually (IGSVM, GRSVM, reliefFSVM, CorrelationSVM. From this table one can easily find out that comparison to all four filter methods IGSVM performed better. For some instance the IGSVM and Correlations achieved 100% for DLBCL dataset. Since various tools using default settings to achieve high accuracy by altering setting and selection is not possible. Meo et al. work out with prostate cancer with accuracy of 95.10% by using the randomized test.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 105-109
Published Online June 2019 in IJEAST (http://www.ijeast.com)

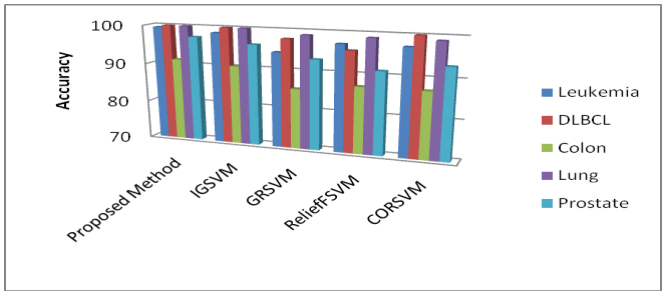

Fig 2**.** Performance comparison of different methods Vs proposed method.

Gene selection methods are used to determine biomarker genes from the microarray dataset. This study done with five various cancerous and noncancerous datasets are considered presented in the Table 1. The biomarker genes selected by our proposed model performed better as compared to individual, the detail of selective genes are mentioned in Table 3.

Datasets Name	Selected Gene Description	Probe set Details
Colon	F765	M76378_at
	F1423	-
Lung	F2968	M61906_at
	F4530	U45973_at
Prostrate	F6185	37639_at
	F7067	40436_g_at

V. CONCLUSION

In this research study we have proposed a hybrid method with combination of four filter methods to select the biomarker gene by following a wrapper and classification method. After considering the best genes from all individual filters we can easily remove the irrelevant features from the microarray gene expression dataset which is high dimension in nature. A wrapper method is applied to eliminate the redundant natured genes out of 150 genes from the four filters. Finally after using LIBSVM classification tool we achieved 2 biomarker genes from three datasets such as Colon Cancer, Lung Cancer and Prostate Cancer. We demonstrated better performance with our proposed model with comparison with individual filter-wrapper classification approach. In this study we concluded that few biomarker genes are enough to achieve the classification accuracy from the microarray high dimension datasets. Even if we have detected some biomarker genes with high accuracy still we have to go though the functionality of the gene for accurate biomarker identification. This experiment is workout with small dimension datasets but it should need to be validating with large database.

VI. REFERENCE

[1] Rakkeitwinai, S., Lursinsap, C., Aporntewan, C., & Mutirangura, A. (2015). New feature selection for gene expression classification based on degree of class overlap in principal dimensions. Computers in biology and medicine, 64, 292-298.

[2] Zhou, W., & Dickerson, J. A. (2014). A novel class dependent feature selection method for cancer biomarker discovery. Computers in biology and medicine, 47, 66-75.

[3] Zhang, X., Song, Q., Wang, G., Zhang, K., He, L., & Jia, X. (2015). A dissimilarity-based imbalance data classification algorithm. Applied Intelligence, 42(3), 544-565.

[4] Wu, R., & Lu, L. (2017, August). Multi-platform microarray integration research. In AIP Conference Proceedings (Vol. 1864, No. 1, p. 020102). AIP Publishing.

[5] Kabir, M. M., Shahjahan, M., & Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. Neurocomputing, 74(17), 2914-2928.

[6] Pugalendhi, G., Vijayakumar, A., & Kim, K. J. (2016). A new data-driven method for microarray data classification. International Journal of Data Mining and Bioinformatics, 15(2), 101-124.

[7] Marafino, B. J., Boscardin, W. J., & Dudley, R. A. (2015). Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. Journal of biomedical informatics, 54, 114-120.

[8] You, W., Yang, Z., Yuan, M., & Ji, G. (2014). Totalpls: local dimension reduction for multicategory microarray data. IEEE Transactions on Human-Machine Systems, 44(1), 125-138.

[9] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

[10] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

[11] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[12] Speed, T. (Ed.). (2003). Statistical analysis of gene expression microarray data. CRC Press.

[13] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[14] Leung, Y. Y., Chang, C. Q., Hung, Y. S., & Fung, P. C. W. (2006, August). Gene selection for brain cancer classification. In Engineering in Medicine and Biology

Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE(pp. 5846-5849). IEEE.

[15] Wang, Y., Makedon, F. S., Ford, J. C., & Pearlman, J. (2004). HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics, 21(8), 1530-1537.

[16] Latkowski, T., & Osowski, S. (2015). Data mining for feature selection in gene expression autism data. Expert Systems with Applications, 42(2), 864-872.

[17] Liu, J., & Zhou, H. B. (2003, November). Tumor classification based on gene microarray data and hybrid learning method. In Machine Learning and Cybernetics, 2003 International Conference on (Vol. 4, pp. 2275-2280). IEEE.

[18] Shreem, S. S., Abdullah, S., Nazri, M. Z. A., & Alzaqebah, M. (2012). Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection. J. Theor. Appl. Inf. Technol, 46(2), 1034-1039..

[19] Sahu, B., Mohanty, S. N., & Rout, S. K. (2019). A Hybrid Approach for Breast Cancer Classification and Diagnosis.

[20] Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial intelligence in medicine, 31(2), 91-103.

[21] Sahu, B. (2018). A Combo Feature Selection Method (Filter+ Wrapper) for Microarray Gene Classification. International Journal of Pure and Applied Mathematics, 118(16), 389-401.

[22] Sahu, B., Mohanty, S. N., & Rout, S. K. (2019). A Hybrid Approach for Breast Cancer Classification and Diagnosis.

[23] El Akadi, A., Amine, A., El Ouardighi, A., & Aboutajdine, D. (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems, 26(3), 487-500.

[24] Samee, N. M. A., Solouma, N. H., & Kadah, Y. M. (2012). Detection of biomarkers for hepatocellular carcinoma using a hybrid univariate gene selection methods. Theoretical Biology and Medical Modelling, 9(1), 34.

[25] Sharbaf, F. V., Mosafer, S., & Moattar, M. H. (2016). A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics, 107(6), 231-238.

[26] SAHU, Bibhuprasad. Multi-Tier Hybrid Feature Selection by Combining Filter and Wrapper for Subset Feature Selection in Cancer Classification. Indian Journal of Science and Technology, [S.l.], feb. 2019. ISSN 0974 - 5645..

[27] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[28] Vural, H., & Subasi, A. (2015). Data-mining techniques to classify microarray gene expression data using gene selection by SVD and information gain. Model Artificial Intel, 6, 171-182.

[29] Li, J., Wang, Y., Cao, Y., & Xu, C. (2016). Weighted doubly regularized support vector machine and its application to microarray classification with noise. Neurocomputing, 173, 595-605.

[30] Chan, W. H., Mohamad, M. S., Deris, S., Corchado, J. M., Omatu, S., Ibrahim, Z., & Kasim, S. (2016). An improved gSVM-SCADL2 with firefly algorithm for identification of informative genes and pathways. International Journal of Bioinformatics Research and Applications, 12(1), 72-93.

[31] Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., ... & Rao, S. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85(1), 16-23.

[32] Mao, Z., Cai, W., & Shao, X. (2013). Selecting significant genes by randomization test for cancer classification using gene expression data. Journal of biomedical informatics, 46(4), 594-601.