International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Classification of Gene Expression Data using Efficient Feature Selection Technique and Resampling Method

Rabindra Kumar Singh, M. Sivabalakrishnan

Abstract: ***Microarray technology has been developed as one of the powerful tools that have attracted many researchers to analyze gene expression level for a given organism. It has been observed that gene expression data have very large (in terms of thousands) of features and less number of samples (in terms of hundreds). This characteristic makes difficult to do an analysis of gene expression data. Hence efficient feature selection technique must be applied before we go for any kind of analysis. Feature selection plays a vital role in the classification of gene expression data. There are several feature selection techniques have been induced in this field. But Support Vector Machine with Recursive Feature Elimination (SVM-RFE) has been proven as the promising feature selection methods among others. SVM-RFE ranks the genes (features) by training the SVM classification model and with the combination of RFE method key genes are selected. Huge time consumption is the main issue of SVM-RFE. We introduced an efficient implementation of linier SVM to overcome this problem and improved the RFE with variable step size. Then, combined method was used for selecting informative genes. Effective resampling method is proposed to preprocess the datasets. This is used to make the distribution of samples balanced, which gives more reliable classification results. In this paper, we have also studied the applicability of common classifiers. Detailed experiments are conducted on four commonly used microarray gene expression datasets. The results show that the proposed method comparable classification performance.***

Index Terms: feature selection, classification, gene expression data, Microarray

I. INTRODUCTION

DNA microarray is a benchmarked tool which has attracted the researchers to analyze gene expression level in the organism. Microarray data can imitate physiological status and gene activities of the organism at the level of transcriptome. Cancer is known to the deadly disease in the world, but if it is identified in an early stage, then it can be controlled by medical science. Microarray gene expression dataset typically comprises high dimensions with small sample size and noise [14]. The characteristics of gene expression data have remained almost the same over the decades. Out of these characteristics high dimensionality, small sample sizes along with class imbalance are challenging issues to be taken care [15]. Feature selection is recognition of genes, which are strongly associated to specific diseases. Usually, by measuring the classification performance, feature selection quality is evaluated. Therefore, in gene recognition task classification plays a major role. Basically, classification task is nothing but a disease diagnosis in gene expression data. Large number of feature with small sample size of training dataset is a curse for a classification task, and classification model generalization ability can be faulty [16]. Taking into consideration the properties of gene expression data like high dimensionality along with small sample size, reduction of the dimensions is necessary before we go for classification. Normally, feature selection is one of the best options for microarray data dimensions reduction. Hence a suitable classifier with efficient feature selection becomes essential for disease diagnosis or gene recognition for microarray data. On the other side, a badly class imbalance can easily lead to misleading classification results [17]. Therefore an effective resampling technique is also required in order to solve this problem. Feature selection techniques have received considerable attention of many researchers in the last decades [1]. Many feature selection techniques have been proposed to extract disease-effected genes [3–5]. LS bound measure was proposed in order to address several redundant genes [2]. Many statistical methods ($\chi 2$, etc) and some classical classifiers (SVM, etc) were used in feature selection [6]. These feature selection approaches can be categorized in three: 1. Filter, 2. Wrapper and 3. Embedded methods [6, 7]. Filter method used to select feature subset using suitable evaluation rule on the basis of statistical techniques [8]. The wrapper method depends upon the classifier performance to assess the significance of feature subsets [9], and the embedded [10] combines the benefit of wrapper and filter techniques. Features (genes) are selected with the help of pre-determined classifier [11, 12]. The filter approaches are suitable for massive data processing because it is of the classifier and the computational complexities of these methods are relatively low [13]. SVM(Support Vector Machine) gives a competitive performance in classification, because of this property and inherent feature selection capability, SVM attracted many researchers interest for a long time. A novel feature selection technique proposed by Guyouet. al. called SVM-RFE, they used the capabilities of Support Vector Machine(SVM) and one feature was deleted recursively which happens to be least important in the ranked list till number of the remaining features meets requirements [24].

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Rabindra Kumar Singh*, School of Computing Science and Engineering, VIT Chennai Campus, India.
Dr. Sivabalakrishnan M., School of Computing Science and Engineering, VIT Chennai Campus, India.

In subsequent studies this method was quickly considered as a benchmark among the feature selection algorithm. But SVM-RFE does not take into consideration the probable hidden correlation among features in the process of feature selection [25]. This can be considered as one of the drawbacks of SVM-RFE. A hybrid method that combined the mRMRwith SVM-RFE was proposed by Mundraet. al. to select more important genes, to solve this problem [26]. Another variant of SVM-RFE was suggested by Yoon S et al. which work on the basis of mutual information [27]. Another disadvantage of SVM-RFE is extremely time-consuming in feature selection process. Two-stage SVM-RFE method was proposed by Tang et al. [28] to accelerate the process of feature selection. Recursive Feature Elimination (RFE) was improvised by Ding Y et. al.[29], in which during every iteration number of feature to delete keeps changing. Here, 1 / (j+1) of the left over features was deleted in the jth iteration. So we can say that microarray dataset having 2500 genes, 12500 genes to be removed, and then 6250 genes will be removed in first and second iteration respectively and so on. This method is treated as "too rude" for feature elimination. Even though, this procedure was greatly speeded, but this kind of procedure can definitely affect the quality of feature selection process. Yin J et. al [30] also proposed to improve RFE. These techniques accomplish better performance and reduce time utilization somewhat. The main aim of this paper is to handle the feature selection problem in order to achieve feature selection quality. First, with the help of variable step size we recommend a new version of RFE. The meaning of step size is number of features to be eliminated in the iteration process. Correctly we can say that when selection of number of feature gets reduced step size also reduced. When it reaches to a certain point in the latter, the former remains unchanged with one. We also introduced linear SVM implementation which is an efficient one as a replacement of SVM and combined with improved RFE in order to further speed up the feature selection process. The experiments show that we achieve promising classification accuracy.

High dimensionality with small sample size is the key problem for the microarray data analysis, at the time case becomes more worse because of class imbalance. Large gap among the sample belonging to diverse classes is known as class imbalance. Class imbalance often gives an unpredictable classification result. Such as, if the given test dataset has sample of two classes, in such a way that the sample of X is double that of the other Y, and if all samples predicted in test dataset as X, then we may obtain the accuracy of 66.67% which much higher than 50%. Hence we can say that the class imbalance will lower the classification accuracy credibility. Therefore, various resampling approaches have been suggested by researchers to solve this issue [17-27]. Over-sampling and under-sampling are traditional methods. The working principles of these methods are either selection of samples is done from minority class randomly or deletion of samples from major randomly, and then replicate them. But this results in overfitting or loss of information [27]. Then, SMOTE (Synthetic Minority Over-sampling Technique) was proposed [17] which are well known and a popular resampling technique. This method showed effective result [18]. The characteristics of SMOTE is to synthesize the values of generated samples. Hence, this method (SMOTE) is not appropriate for microarray data in particular when the objective is gene recognition. Ensemble methods have got significant attention for their competitive result [19-22]. However, the ensemble method complexity is very high for microarray dataset because of the small sample size. In this paper, we proposed an effective resampling method, which is appropriate for microarray data. Here, we select randomly feature value instead of selecting a random sample and then new sample is constructed.

Classification is the main component for analyzing the microarray data. However, we have not worked towards building new classifiers rather we used the existing one. In this paper, we have chosen four most benchmarked microarray dataset, preprocessed with the proposed resampling method, and then select the meaningful feature with our proposed system. Finally, we performed a classification task with SVM, k-Nearest Neighbors as well as Logistic Regression [30]. Obtained results show that the accuracy of all classifiers are very different and any particular classifier may not be the best choice all the time. The remainder in this paper is structured as follows: Methods and materials are introduced– resampling method, RFE with step size, etc. in section II. In section III, we presented experiments, results are presented in section IV which includes dataset description, data preprocessing, performance evaluation, experimental results, and discussion. The conclusion is narrated in section V.

II. MATERIALS AND METHODS

A. System Architecture

System architecture is illustrated in Figure 1. In our system input data are microarray gene expression data. First data are preprocessed in order to take care of noise and inconsistency. After that resampling method is used and balanced dataset is created. Then appropriate feature selection methods are used in order to select only important features i.e. genes. Finally different classifiers are used and efficiency and effectiveness are measure.

We divided the system in three parts

i. Resampling and balancing the dataset
ii. Feature selection
iii. Classification

The detailed explanation will follow in below subsection

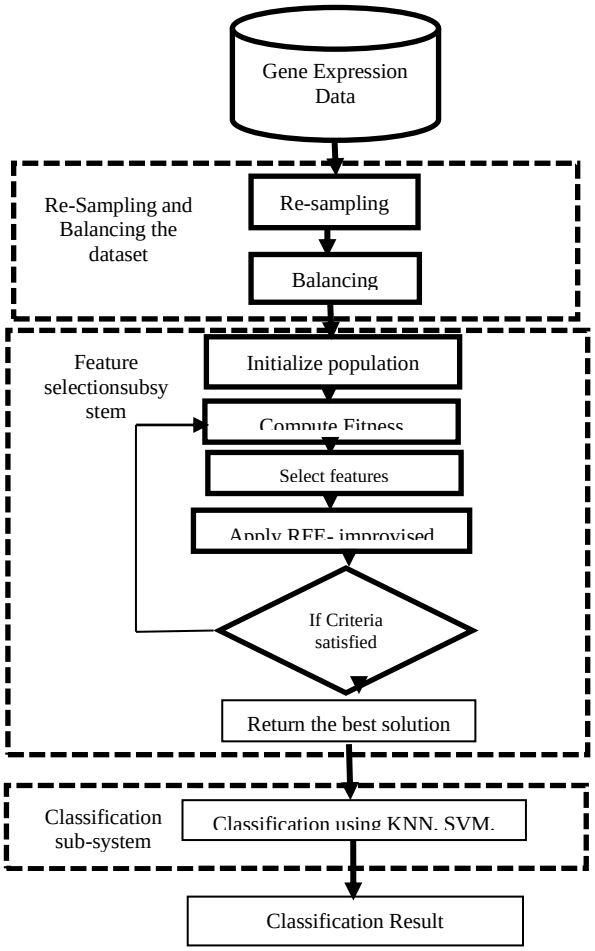


Figure 1: **System Architecture**

B. *Re-sampling based on Random Value(RSRV)*

The gene expression data are biologically precise and hence are not supposed be altered randomly. Here, the methodology presented objective is to resolve the class imbalance issue of microarray data with retaining the biological importance without model overfitting and information loss. Here we assume that the samples of the same class belong from the same distribution. Now data matrix has been constructed considering the minority class, and then chooses one value arbitrarily from every column.Now current sample is saved and in order to prepare number of equal sample of both the classes we repeat this process k times. Finally k samples are obtained, that are different from the original dataset but from the same distribution.

Algorithm 1: Re-sampling based on Random Value (RSRV) for class imbalance issue

1. **X** - Given data matrix, and k- the number of new samples;
2. While k >= 1:
   (1) For j = 1, 2, . . ., n (n denotes the column size of **X**):
      i. Randomly choose a value V from **X**j (the jth column of **X**);
      ii. Save V to the respective position of the current sample(new);
   (2) Save sample(new) to **X**;
   (3) k = k± 1;
3. Return **X.**

Here X is the given data matrix, which represents the minority class of the microarray data where in row represents samples and columns represents features (genes).

C. *Recursive feature elimination (RFE) by means of variable step size*

Here, the proposed RFE method by Guyon et.al.[24] is nothing but a backward feature elimination. Weights are assigned to features based on external estimator. Eliminating the most irrelevant feature recursively is the main objective of RFE. Therefore, first training is given to the estimator with starting set of feature and weight has been assigned. Then, weights are arranged in descending order taking the absolute value. Finally, last feature (features)iseliminated. This procedure is repeated on pruned set till the required number of features gets selected.

RFE main working principle is to find correlation between the features and remove feature or set of features which are least relevant. The main problem of RFE is it takes tremendous time in particular when the dimensions of dataset are extremely high. And hence, essentiallythe step size has to be increasedin order to reduce the number of iteration. But, some researchers have stated that increasing the step size will give the negative impact of feature selection in particular for microarray dataset [24].

Hence, we proposed RFE along with variable step size in order limit the contraryimpact on feature selection. First we took the initializestep size which happens to be a large value and then reduced to half,i.e.the number of features which has to be eliminated is brought to half compared with the actual original size. Repeat this procedure till the step size reaches to one. It very well may be additionally clarified with two angles: that is the step sizechanges from large to small and not necessarily will change in each iteration. It changes based on update condition, updated rule as well as number of features which has to be eliminated. Additional, the feature elimination process is refined progressively from roughness. Normally, microarray gene expression dataset contains large number of genes (features), and only few of them are strongly related to class labels. Hence, definitely we can eliminate relatively more number of genes (features)in the beginning which are irrelevant. In other words, eliminated gene at later stage is more significant. Hence, we can set relatively large step size in earlier stage in order to reduce the number of iterations. Step size is reduced gradually and features are selected with extra care at the later stage. Thus, the feature selection quality is ensured. The detailed procedure is illustrated in Algorithm 2.

Algorithm 2: RFE using variable step size(RFEVSS)

1. **X**- Given set of genes, **Y**- labels of sample, n_selected - number of genes to select, s_initial-initial step size
2. Get total quantity of genes from **X**, n_total

3. Temp = n_total; N = n_total; S = s_initial
4. While N >n_selected:
   (1) N = N±S;
   (2) If temp / N = 2 and S >1:
      i. Temp = N;
      ii. S = S / 2;
   (3) Train LLSVM with X and Y and get sorted weights vector **W**;
   (4) Delete features according to **W** and S, and update **X**;
5. Return **X.**

D. *Large scale linear SVM*

Many researcher have taken SVM is a good choice as feature selection and often deployed as classifier for microarray gene expression data. However, SVM is based on kernel technique (normally, liner kernel) and Lagrange dual solver [24]. We have used large scale liner SVM (LLSVM) in order to speed up the process of allocating weights [31]. LLSVM is known as pure linear classifier. This classifier is designed specifically performing classification task on large-scale dataset e.g. classification of text. Microarray dataset also have very large dimensions just like text. Hence LLSVM must be suitable for microarray data set as well.

The objective function of large scale liner SVM is defined as:

$$\min_{w} f(w) \equiv \| w \|_1 + C \sum_{i \in I(w)} b_i(w)^2 \quad (1)$$

Where,

$$b_i(w) \equiv 1 - y_i w^T x_i \quad (2)$$

$$I(w) \equiv \{i | b_i(w) > 0\} \quad (3)$$

x_i, $y_i$and w denotes the feature vector for i^{th} sample, corresponding label and the weight vector for features respectively. In this way, large scale linear SVM's loss function is adjusted, i.e. L1 is regularized. Penalty factor C > 0 shows that the weight vector (w) is sparse. As C increases, less important genes with more weights will get penalized to 0, hence w becomes sparser. Next and final decision function is of same form, like liner SVMs.

$$f(x^*) = sign(w.x^*) \quad (4)$$

x* is denoted as the unknown feature vector of the sample.

Cyclic coordinate descent method was applied by Yuan et. al. introduced to solve formula(1) [31]. Cyclic coordinate descent method updates one variable at a time to generate $w^{k,j} \in R^n$, j = 1,…, n+1 from the current solution w^k. Here J is feature and k is iteration. So $w^{k,1} = w^k$, $w^{k,n+1} = w^{k+1}$, and

$$w^{k,j} = [w_1^{k+1}, \dots, w_{j-1}^{k+1}, w_j^k, \dots, w_n^k] for\ j = 2, \dots, n \quad (5)$$

The following one-variable optimization problem is solved, to update $w^{k,j}$ to $w^{k,j+1}$:

$$\min_z g_j(z) = |w_j + z| + L_j'(0; w)z + \frac{1}{2} L_j''(0; w)z^2 + constant \quad (6)$$

Where,

$$e_j = [0, \dots, 0, 1, 0, \dots, 0]^T \in R^n, \quad (7)$$

$$L_j(z; w) \equiv C \sum_{i \in I(w+ze_j)} b_i(w + ze_j)^2, \quad (8)$$

And

$$L_j'(0; w) = -2C \sum_{i \in I(w)} y_i x_i b_i(w), \quad (9)$$

$$L_j''(0; w) = max\left(2C \sum_{i \in I(w)} x_{i,j}^2, 10^{-12}\right), \quad (10)$$

Equation (6) is an approximate method because $L_j(z; w)$ is not a double differentiable. Change value is Z for the variable j. And if z* is the solution of equation (6), at that timejth element is updated by:

$$w_j^{k,j+1} = w_j^{w,j} + z^* \quad (11)$$

If all variables have been updated then we say one iteration is completed. The result tends be stable after m iteration. The framework of the method is shown in Algorithm 3.

Algorithm 3: Cyclic coordinate descent method for Large scale liner SVM

1. Given $\mathbf{w}^1$;
2. For k = 1, 2, 3, . . . m:
   (1) $w^{k,1} = w^1$;
   (2) For j = 1, 2, . . ., n:
      i. By solving sub-problem (6) Z* is obtained;
      ii. $\mathbf{w}^{k,j+1} = \mathbf{w}^{k,j} + z^*\mathbf{e}_j$;
3. Return $\mathbf{w}^{m+1}$

III. EXPERIMENTS

This section we emphasis on the experimental verification of the proposed methods. We have chosen four most frequently used benchmarked datasets. Dataset descriptions are presented in section A. Experimental settings are described in section B including data preprocessing; parameter estimation described in section C,and in section D measures of performance evaluation are outlined.

All experiments were conducted with following system specification:

Hardware:

Processor:	Intel, Corei5-7200 CPU- 2.50GHz 2.70 GHz
Memory(RAM):	8 GB
System Type:	64-bit

Software:

Operating System:	Fedora 30, 64 bit
Language:	Python 3.7

A. *Datasets*

We have selected four benchmarked gene expression microarray datasets, in order to conduct extensive experiments.

All of them are widely used by many researchers in this field [28, 41] and available online. Leukemia (ALL AML) and Colon datasets are available in [42]. Ovarian and Breast datasets are available in [43]. All these selected datasets are binary. Class imbalance problems are common in all of them. The details are shown in Table 1. Sample-to-dimension ratio is denoted by SDR, i.e., (No of class 1 + No of class 2) / No of Features. Imbalance ratio is denoted by IR, i.e. (No of class 2 / No of class 1).

The proposed RSRV algorithm in section II (B) is applied, for solving the class imbalance issue to the four datasets stated above. New samples are acquired from class 1 in such way that number of class 1 is equal to number of class 2. So, IR fits to 1.0 for all the datasets and SDR changes subsequently.

Table 1. Raw dataset characteristics.

Dataset	No of Class 1	No of Class 2	No of Features	SDR	IR	Reference
Colon	22	40	2000	3.1%	1.82	[33]
Leukemia	25	47	7129	1.01%	1.88	[34]
Ovarian	91	162	15154	1.67%	1.78	[35]
Breast	46	51	24481	0.40%	1.11	[36]

B. *Data Preprocessing*

All the datasets have been standardized (raw datasets and balanced datasets), empirically with unit variance and zero mean. Hence, the adverse effects produced through different genes having a large gap between the gene expression values can be eliminated. Here, mRMRtechnique is used, which is based on mutual information. Therefore, discretization of dataset is necessary. We used the measure, which has been outlined in [12] as give below:

$$\overline{x} = \begin{cases} +2, if\ x > \mu + \sigma/2 \\ -2, if\ x < \mu - \sigma/2 \\ 0, \quad otherwise \end{cases}$$

Where the mean value is denoted by μ and standard variance by σ. $\overline{x} = +2$ is over expression $\overline{x} = -2$ is under expression and $\overline{x} = 0$ means normal expression.

Hence, we get continuous and discrete (two versions) of datasets, which are standardized.

C. *Estimation of Parameter*

Penalty factor C is the key parameter for large scale linear SVM, SVM and LR. The feature selection result is depends on the value of C. Penalty factor C also affects the complexity of Classification model for SVM, Large scale linier SVM and LR. The depth is set to be five for LR, so important parameter is set to be the basic trees number. Here, N denotes the basic trees number, theoretically larger N gives better performance. Algorithm execution time increases linearly, when N go beyond a certain limit for some dataset. Hence, N should be assigned neighbor's number is referred by K to be determined for kNN, too small to specific valuesfor different datasets. The nearest or too big will not be a good choice. So, it requires patient tuning. Hence, parameter estimated with corresponding model, when we deploy these models as feature selector of classifiers. However, another parameter step size input (S) is need to be determined before applying RFEVSS and large scale linear SVM(represented as LLSVM-RFEVSS and SVM-RFEVSS respectively). We utilize grid search and 5-fold cross validation, for specifying these parameters. Details are shown in table 2.

Table 2. Parameters of feature selectors / classifiers for balanced datasets.

		Parameter	Leukemia	Ovarian	Breast	Colon
Feature Selectors	LLSVM	C	0.1	0.3	0.3	0.9
	SVM	C	0.1	0.5	0.1	0.1
	RF	N	100	100	400	200
Step size	LLSVM	S	600	1000	800	100
	SVM	S	400	1000	800	200
Classifiers	SVM	C	9	3	0.09	7
	kNN	K	1	7	7	6
	LR	C	19	7	3	9

In this paper balanced datasets were used to conduct the experiments, but even though we have tuned C and S for validation of the performance of RSRV procedure (Algorithm 1) for SVM-RFEVSS on raw datasets. The details are presented in Table 3.

Table 3. Parameters for SVM-RFEVSS on raw datasets.

Feature selectors	Parameters	Leukemia	Ovarian	Breast	Colon
SVM	C	0.3	0.5	0.1	0.5
RFEVSS	S	100	1000	1000	60

We can see in Tables 2 and 3 that the starting step size is quite different for different datasets. The step size becomes larger, when datasets have more genes (Breast and Ovarian). On other hand, when datasets have fewer genes (Leukemia and Colon), the starting step size becomes smaller. This confirms exactly, which is depicted in section II (C) i.e. importance of gene assumption and the basis for improvising RFE.

D. *Measures for performance evaluation*

Here, we have chosen three commonly used measures in this paper for performance evaluation measure: MCC, ACC and AUC. All these measures are commonly used in evaluation of classification. MCC and ACC and are defined as:

$$\text{MCC} = \frac{(\text{TP X TN}) - (\text{TP X FN})}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(TN + FN)}}$$

$$ACC = \frac{TN + TP}{TP + TN + FP + FN}$$

Where,

MCC – Matthew's correlation coefficient
ACC – Accuracy
AUC–Area under ROC curve
TP – True Positive
FP – False Positive
TN – True Negative
FN – False Negative

Here, most common evaluation measure is ACC, but using it alone is not enough. Best choice often considered MCC because, MCC gives good performance evaluation dataset suffer with class imbalance. MCC is basically a coefficient of correlation of the observed and predicted value. It has the values between -1 and +1. The coefficient value -1 refers the worst prediction whereas +1 refers the perfect prediction. False Positive Rate (TPR) and True Positive Rate (FPR) both are taken in account for AUC, which are defined as below:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

AUC can be considered as probability values which classify correctly one sample, the bigger is better.

IV. RESULTS

Now in this part of paper, three sets of comparative analysis and evaluation of model are performed. We verify our proposed RSRV, RFEVSS and LLSVM algorithms in the comparative analysis. Then, we evaluate three commonly used classifiers and discuss the suitability of them for microarray data.

A. *Comparative experiments of balanced datasets with RSRV and raw datasets:* Here, we have used RSRV to balance the raw data and experiments were conducted with SVM-RFEVSS on four balanced and raw datasets for selection of genes. We have chosen SVM-RFEVSS as a feature selector, since SVM-RFE takes more time than SVM-RFEVSS for achieving the same objective. Linear SVM (with C=1) has been used as a classifier, and all the dataset has been executed 128 times in order to select 1 to 128 genes.

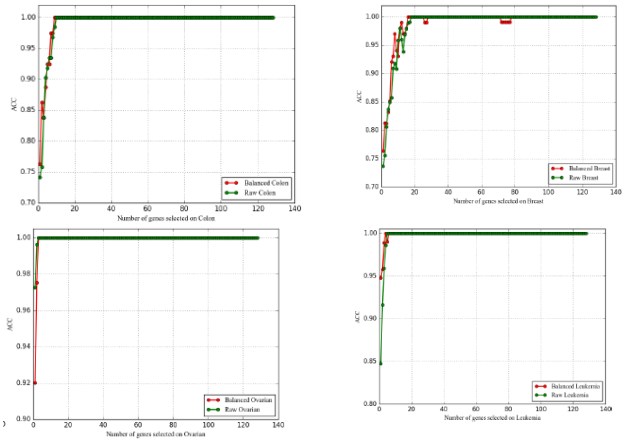


Figure 2: The comparison of ACC acquired on four raw and balanced datasets.

Performance comparison of three (ACC, MCC and AUC) evaluation measures onbalanced and raw datasets is shown in Figures: 2-4. We can observe that the balanced Leukemia gives better performance on all measures. The balanced Colon and Breast perform better on MCC and ACC while on AUC it is comparable. We also observed that the performance of balanced ovarian is unsatisfactory, but this happens if less number of genes are selected. Results obtained on the balanced dataset are better than the raw dataset, if the number of genes increases. This shows that for solving the class imbalance issue of microarray dataset, RSRV is good choice.

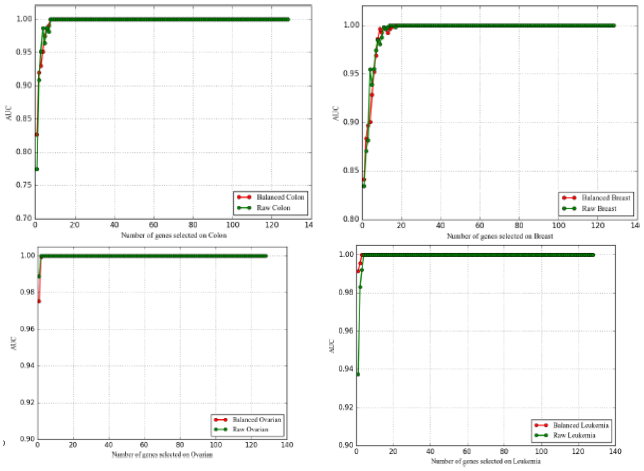


Figure 3: The comparison of AUC acquiredon four raw and balanced datasets.

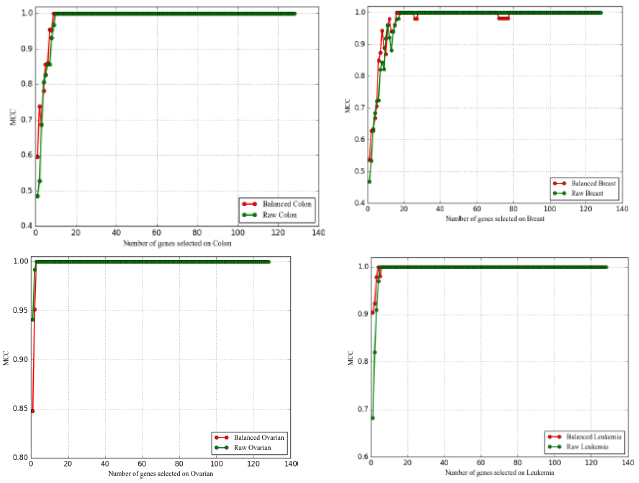


Figure 4: The comparison of MCC acquired on four raw and balanced datasets.

B. *Comparative experiments of RFEVSS and RFE*.

The effectiveness of RFEVSS is validated in this section. Here, as a basic feature selector linear SVM is used, then combined with RFE and RFEVSS individually for conducting the experiments with four balanced datasets. There are two set of experiments were conducted with the same conditions without RFE's step size. It is fixed to 1 in one case, and in other it has to be determined by the initial input value along with number of features that is to be eliminated. Linear SVM with C=1 has been chosen as the classifier.

Table 4. The comparison of performance between SVM-RFE and SVM-RFEVSS.

	SVM-RFE			SVM-VSSRFE		
	ACC	AUC	MCC	ACC	AUC	MCC

Breast	**0.8619**	**0.9457**	**0.7294**	0.8328	0.9013	0.6695
Colon	**0.9385**	**0.9885**	**0.8764**	0.8885	0.9526	0.7830
Ovarian	0.99	**1.0**	0.9809	**1.0**	**1.0**	**1.0**
leukemia	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Table 4 shows the evaluation results of SVM-RFE and SVM-RFEVSS. The best performances are highlighted in bold.

C. *Time consumption comparison between SVM-RFEVSS and LLSVM-RFEVSS*

Table 5. Time consumption (s) comparison between SVM-RFEVSS and LLSVM-RFEVSS.

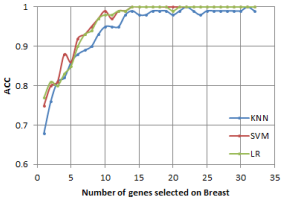
	SVM-RFEVSS	LLSVM-RFEVSS
Breast	4551.82	**970.06**
Colon	**96.87**	100.03
Ovarian	2256.14	**650.55**
Leukemia	741.60	**97.05**

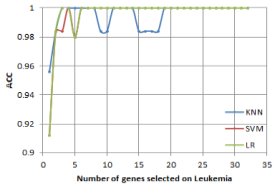
The time consumed by LLSVM-RFEVSS and SVM-RFEVSS shown in Table 5, from this we are able to observe that the time consumption of LLSVM-RFEVSS is reduced by great extent (bold faces shows the best performances), particularly if the dimensions are large (e.g. Breast).

D. *Comparison of three common classifiers performance*

In this section of the paper, we validated three frequently used classifiers; including SVM (linear SVM), LR (L2 regularized Logistic Regression) and kNN(k-Nearest Neighbors). LLSVM-RFEVSS is deployed as a feature selector on the balanced dataset to select 1 to 32 genes. Then these genes are evaluated with well-tuned classifiers. LLSVM-RFEVSS has been also utilized as a feature selector for four balanced datasets. Then we used LR as the classifier for conducting experiments that aims to acquire the training as well as testing scores thus the classification model's generalization capability can be evaluated.

Classification performance is depicted in Figures: 5-7, which show the effects of classifiers. We can see that the results acquired on the same data by different classifiers are very different. In most of the cases SVM and LR beat the evaluation measure on all datasets, the reason is simple (i.e. microarray dataset is linearly separable therefore SVM and LR are better suited). We are also able to see that the graph of LR is smoother than SVM in most of the cases; hence LR performance is more stable.





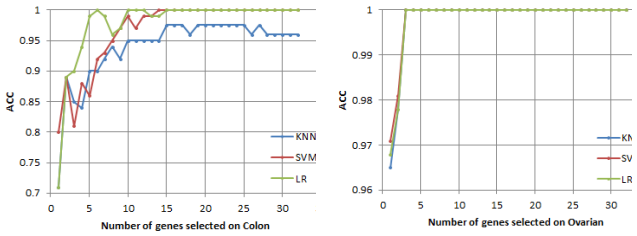


Figure 5: The comparison of ACC acquired by three classifiers.

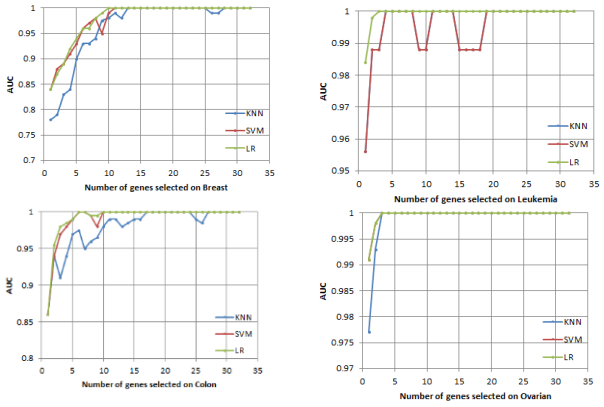


Figure 6: **The comparison of AUC acquired by three classifiers.**

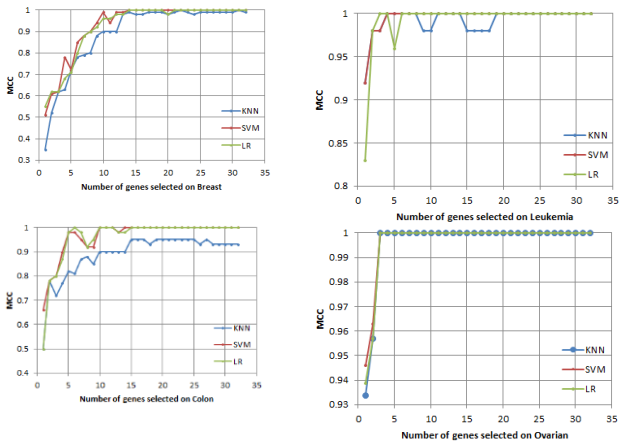


Figure 7: **The comparison of MCC acquired by three classifiers.**

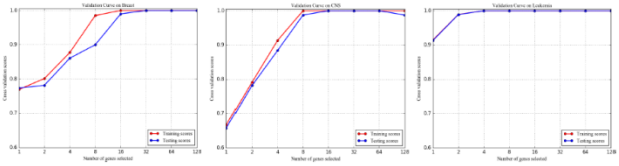


Figure 8: **The classification model evaluation**.

The evaluation results of the classification model are shown in figure: 8 with selection of 1, 2, 4, 8, 16, 32, 64, 128 genes respectively in the form of 2^n where n=0, 1, 2….As we observed, if more genes are selected then testing scores are almost the same as the training scores. This means the classification model has good generalization.

V. CONCLUSION

Diseases like cancer still considered as the greatest threat for human life in the world. The development of microarray dataset and the suitable statistical methods have given new dimensions to predict of such disease. Classification and feature selection are known to be the core technologies for microarray gene expression data analysis. High dimensionality, small sample size and class imbalance of the microarray dataset are the main issues for researchers. Out of them, researchers have rarely attempted class imbalance problem. We used simple yet effective resampling technique called RSRV for preprocessing the data. We have used RFEVSS method as an efficient feature selection technique. We have also induced linear SVM efficient implementation known as LLSVM. Then we combined with RFEVSS, LLSVM-RFEVSS grow into effective and efficient feature selector. At last we conducted a comparative study on the different classifier's effect with the classification results. In this it has been observed that few times LR can be the better option for classifying microarray data.

REFERENCES

1. Shenghui Liu, Chunrui Xu, Yusen Zhang, Jiaguo Liu, Bin Yu, Xiaoping Liu and Matthias Dehmer. "Feature selection of gene expression data for Cancer classification using double RBFkernels" , BMC Bioinformatics (2018), 2-14
2. Zhou, Xin, and K. Z. Mao. "LS bound based gene selection for DNA microarray data." Bioinformatics 21.8 (2004): 1559-1564.
3. Kira, Kenji, and Larry A. Rendell. "A practical approach to feature selection." Machine Learning Proceedings 1992. Morgan Kaufmann, 1992. 249-256.
4. Chater, Nick, and Mike Oaksford. "Information gain and decision-theoretic approaches to data selection: Response to Klauer (1999)." (1999): 223.
5. Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3.02 (2005): 185-205.
6. Saeys, Yvan, IñakiInza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." bioinformatics23.19 (2007): 2507-2517.
7. Blum, Avrim L., and Pat Langley. "Selection of relevant features and examples in machine learning." Artificial intelligence 97.1-2 (1997): 245-271.
8. Jacobs IJ, Skates SJ, Macdonald N. Screening for ovarian cancer: a pilot randomised controlled trial. Lancet. 1999;353(9160):1207–10.
9. Xiong, Momiao, Xiangzhong Fang, and Jinying Zhao. "Biomarker identification by feature wrappers." Genome Research 11.11 (2001): 1878-1887.
10. Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.
11. Kim, Dae-Won, Kwang H. Lee, and Doheon Lee. "Detecting clusters of different geometrical shapes in microarray gene expression data." Bioinformatics 21.9 (2005): 1927-1934.
12. Duval, Béatrice, and Jin-Kao Hao. "Advances in metaheuristics for gene selection and classification of microarray data." Briefings in bioinformatics 11.1 (2009): 127-141.
13. Brenner, Sydney, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nature biotechnology 18.6 (2000): 630.
14. Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." Advances in bioinformatics 2015 (2015).
15. Bolón-Canedo, Verónica, et al. "A review of microarray datasets and applied feature selection methods." Information Sciences 282 (2014): 111-135.
16. Elkhani, Naeimeh, and Ravie ChandrenMuniyandi. "Review of the Effect of Feature Selection for Microarray Data on the Classification Accuracy for Cancer Data Sets." International Journal of Soft Computing 11.5 (2016): 334-342.
17. Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research16 (2002): 321-357.
18. Zhu, Bing, Bart Baesens, and Seppe KLM vandenBroucke. "An empirical comparison of techniques for the class imbalance problem in churn prediction." Information sciences408 (2017): 84-99.
19. Galar, Mikel, et al. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.4 (2011): 463-484.
20. Qian, Yun, et al. "A resampling ensemble algorithm for classification of imbalance problems." Neurocomputing 143 (2014): 57-67.
21. Ye, Yunming, et al. "Stratified sampling for feature subspace selection in random forests for high dimensional data." Pattern Recognition 46.3 (2013): 769-787.
22. Galar, Mikel, et al. "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling." Pattern Recognition 46.12 (2013): 3460-3471.
23. López, Victoria, et al. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." Information sciences 250 (2013): 113-141.
24. Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.
25. Mundra, Piyushkumar A., and Jagath C. Rajapakse. "SVM-RFE with MRMR filter for gene selection." IEEE transactions on nanobioscience 9.1 (2009): 31-37.
26. Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on Pattern Analysis & Machine Intelligence 8 (2005): 1226-1238.
27. Yoon, Sejong, and Saejoon Kim. "Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms." Pattern Recognition Letters 30.16 (2009): 1489-1495.
28. Tang, Yuchun, Yan-Qing Zhang, and Zhen Huang. "Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis." IEEE/ACM Transactions on Computational Biology and Bioinformatics 4.3 (2007): 365-381.
29. Ding, Yuanyuan, and Dawn Wilkins. "Improving the performance of SVM-RFE to select genes in microarray data." BMC bioinformatics. Vol. 7. No. 2. BioMed Central, 2006.
30. Yu, Hsiang-Fu, Fang-Lan Huang, and Chih-Jen Lin. "Dual coordinate descent methods for logistic regression and maximum entropy models." Machine Learning 85.1-2 (2011): 41-75.
31. Yuan, Guo-Xun, et al. "A comparison of optimization methods and software for large-scale l1-regularized linear classification." Journal of Machine Learning Research 11.Nov (2010): 3183-3234.
32. Zhu, Zexuan, Yew-Soon Ong, and Manoranjan Dash. "Markov blanket-embedded genetic algorithm for gene selection." Pattern Recognition 40.11 (2007): 3236-3248.
33. Alon, Uri, et al. "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 96.12 (1999): 6745-6750.
34. Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.
35. Petricoin III, Emanuel F., et al. "Use of proteomic patterns in serum to identify ovarian cancer." The lancet 359.9306 (2002): 572-577.
36. Van't Veer, Laura J., et al. "Gene expression profiling predicts clinical outcome of breast cancer." nature 415.6871 (2002): 530.
37. Li, Jundong, et al. "Feature selection: A data perspective." ACM Computing Surveys (CSUR) 50.6 (2018): 94.
38. Kononenko, Igor. "Estimating attributes: analysis and extensions of RELIEF." European conference on machine learning. Springer, Berlin, Heidelberg, 1994.
39. Hawkins, Douglas M. "The problem of overfitting." Journal of chemical information and computer sciences 44.1 (2004): 1-12.
40. Zhu, Pengfei, et al. "Subspace clustering guided unsupervised feature selection." Pattern Recognition 66 (2017): 364-374.
41. Xu, Qian, et al. "Robust Multi-label Feature Selection with Missing Labels." Chinese Conference on Pattern Recognition. Springer, Singapore, 2016.
42. http://featureselection.asu.edu/datasets.php.
43. http://csse.szu.edu.cn/staff/zhuzx/Datasets.html .

AUTHORS PROFILE

Prof. Rabindra Kumar Singh Working as Assistant Professor (Selection Grade) in School of Computing Science and Engineering at VIT Chennai Campus since 2013. He has 20 years of Teaching Experience and 4 years of Industry experience. He has completed M.E. in Computer Science and Engineering from Anna UniversityChennai. He is interested in teaching the subject like Operating System, Computer Network, Data Mining, Software Project Management, Distributed Computing, Python Programming, R Tools, etc. His area of research is Machine Learning, Data Mining, and Bioinformatics.

Dr. M. Sivabalakrishnanworking as Associate Professor in School of Computing Science and Engineering at VIT Chennai Campus since 2013. He has 20 + years of Teaching Experience. He has completed M.E. in Computer Science and Engineering from Anna University Chennai in 2004. He has completed his Ph. D in 2012. from Anna University Chennai. He has published more than 25 papers in International and National journals. His area of Interest is Image processing, Data Mining, Machine Learning, etc.