IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2019, accepted June 2, 2019, date of publication June 12, 2019, date of current version June 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922432

Feature Selection for Binary Classification Within Functional Genomics Experiments via Interquartile Range and Clustering

ZARDAD KHAN[1,2], MUHAMMAD NAEEM[1], UMAIR KHALIL[1], DOST MUHAMMAD KHAN[1], SAEED ALDAHMANI[3], AND MUHAMMAD HAMRAZ[1]

[1]Department of Statistics, Abdul Wali Khan University, Mardan 23200, Pakistan
[2]Department of Mathematical Sciences, University of Essex, UK
[3]Department of Statistics, College of Business and Economics, United Arab Emirates University, UAE

Corresponding authors: Zardad Khan (zkhan@essex.ac.uk) and Muhammad Hamraz (mhamraz@awkum.edu.pk)

ABSTRACT Datasets produced in modern research, such as biomedical science, pose a number of challenges for machine learning techniques used in binary classification due to high dimensionality. Feature selection is one of the most important statistical techniques used for dimensionality reduction of the datasets. Therefore, techniques are needed to find an optimal number of features to obtain more desirable learning performance. In the machine learning context, gene selection is treated as a feature selection problem, the objective of which is to find a small subset of the most discriminative features for the target class. In this paper, a gene selection method is proposed that identifies the most discriminative genes in two stages. Genes that unambiguously assign the maximum number of samples to their respective classes using a greedy approach are selected in the first stage. The remaining genes are divided into a certain number of clusters. From each cluster, the most informative genes are selected via the lasso method and combined with genes selected in the first stage. The performance of the proposed method is assessed through comparison with other state-of-the-art feature selection methods using gene expression datasets. This is done by applying two classifiers i.e., random forest and support vector machine, on datasets with selected genes and training samples and calculating their classification accuracy, sensitivity, and Brier score on samples in the testing part. Boxplots based on the results and correlation matrices of the selected genes are thenceforth constructed. The results show that the proposed method outperforms the other methods.

INDEX TERMS Clustering, classification, feature selection, high dimensional data, microarray gene expression data.

I. INTRODUCTION

A fundamental challenge for machine learning approaches is to learn and analyse data produced via high-throughput data generating technologies such as microarray gene expression data [1], [2]. Microarray data consist of a small number of observations with tens of thousands of genes. Conventional machine learning approaches pose a number of problems to learn from such datasets. This phenomenon is also known as the curse of dimensionality. The problem of high-dimensionality is due to the existence of many redundant and irrelevant features having no contribution in classifying observations to their respective classes. A lot of work has been done to illustrate the importance of feature selection and dimensionality reduction in the literature [3]–[5]. The main theme of feature selection is to select a subset of features/genes that mainly regulate the response variable [2], [6]–[8].

Since many of the features in microarray data are highly collinear, also known as redundant, consequently machine learning methods lose generalization power and interpretability when applied on unseen data [7], [9]–[11]. Thus feature selection methods, when carefully used, not only solve the above mentioned problems, they also save a lot of computational resources. There are three main types of feature selection techniques, i.e. filter, wrapper and embedded methods.

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

A. WRAPPER METHODS

In these methods, a predictive model, run on a data partitioned into training and testing sets, is used to evaluate possible gene subsets. Each of the genes subsets is used with training data in order to train the model which is then tested using the test data. Model accuracies are calculated for each subset of genes that are used as scores for the respective subsets. Gene subset with the highest score is chosen as the final set to run the model. As each gene subset requires a new model to be fitted, wrapper methods require considerable computational resources. Examples of such methods can be found in [12]–[14].

B. EMBEDDED METHODS

In these methods feature selection is part of model construction. The model favours those features that mainly regulate the target class. Classification and regression tree is one of the examples of embedded methods [15]–[17].

C. FILTER METHODS

Filter methods assess the relevance of features by computing the relevant score for each gene. Genes with high-relevance scores are considered for the purpose of classification via different classifiers. Filter methods select features without taking the classification algorithm into consideration. Methods based on filter approach can easily deal with big data as they are computationally simple and fast. Examples of filter methods are correlation coefficient score, chi-squared test statistic and information gain [18], [19].

This paper proposes a filtering approach, GClust, for genes selection based on the combination of a greedy approach and clustering. The greedy approach selects those genes that unambiguously assign maximum number of samples to their correct classes. The remaining genes are then divided into a certain number of clusters based on their similarity and the top ranked genes are selected via the lasso method. This ensures the removal of genes that carry the same information as others, also called redundant genes. The final subset of genes is obtained by combining the genes selected through the greedy approach and clustering. The rest of the paper is organised as follows. A brief review of related work done on feature selection techniques is given in Section II. Description of the proposed method is given in Section III. Description of the gene expression datasets used and experiments based on the proposed method in comparison with other state-of-the-art methods in Section IV. The paper ends with a discussion in Section V based on the work done in this paper.

II. RELATED WORK

Feature selection is a tedious job in microarray data analysis. The basic task is to determine an optimal gene set that is helpful in classifying samples to their correct target classes. Feature selection is important in that there are usually very few genes that regulate the response while some genes are linear combinations of the other genes. As the method proposed in this paper is based on filtering approach, a brief review of these methods is given as follows. Apiletti *et al.* [20] proposed a filtering based feature selection technique in which at first stage they identify outliers in gene expression data for each gene. Authors in [21] proposed a gene selection score called relative simplicity (RS) by evaluating gene pairs according to integrating vertical comparison with horizontal comparison. By this way they built an RS-based direct classifier based on a set of informative genes for binary classification using a paired vote strategy. Although, high generalization power for the RS method is shown on the datasets they considered, however, their method might still suffer from the problem of redundancy in the selected genes due to pairwise evaluation. Filtering approach assesses the importance of genes in discriminating the sample observation given a target class by setting a threshold [22] or by fitting a statistical model [23], [24] to microarray gene expression data.

Filter methods based on density approach need to be robust to reduce the influence of values far from the high concentration core. Authors in [12] exploited this notion and used an expression range to construct a gene mask. Authors in [25] used minimum feature subset by utilizing the set covering approach. The authors in [20] applied the same technique for minimum gene subset and the greedy approach rather than the set covering approach. Authors in [8] proposed to reduce the dimensionality and outliers problem in gene selection by considering the greedy search approach together with proportional overlapping analysis for binary class problems. They first initialize a gene mask where the core intervals for the genes are based on the interquartile range to select a minimum subset of genes that unambiguously assign maximum number of samples to their respective classes. The proportion of overlapped samples between classes and relative dominant class (RDC) for each gene is measured for gene ranking. Genes that achieve the highest ranks are selected in the final set together with the minimum subset of genes selected via the greedy approach. Both of these methods given in [8] and [20] suffer from the problem of multicollinearity in the selected genes as demonstrated in the experiments and results section of the current paper. A measure known as 'PUL' was proposed in [26] by identifying differentially expressed genes based on retrieval information called the PUL-score. Authors in [23] used principle component analysis technique to discard genes having large variations corresponding to the components. Removal of non-informative genes was considered in [24] by using factor analysis technique instead of principle component analysis. Kulkarni *et al.* [27] proposed a method called "Recursive Cluster Elimination" (RCE) which reduces the dimension of data sets in the study of cancer classification. The authors in [27] combine statistical tests (Analysis of Variance and Principle Component Analysis) with the approach of Recursive Cluster Elimination (RCE) which reduces the gene expression data into a small number of gene subsets. This algorithm at each step removes irrelevant and redundant features. Liu *et al.* [28] modified kernel-based clustering method for gene selection (KBCGS) by proposing the use of double radial basis function (RBF) (as kernel) based

on clustering (DKBCGS). DKBCGS has shown satisfactory global classification performance for gene selection. This method, however, inherits issues related to the kernel function in terms of extrapolation.

III. METHODS

In a microarray data, there are p number of genes and n number of samples. The data matrix is in the form $R^{n\times p}$. This makes a gene expression matrix $X = [x_{ij}]$, such that $X \in R^{n\times p}$, where x_{ij} is the gene expression value for ith tissue sample ($i = 1, \ldots, n$) and jth gene ($j = 1, \ldots, p$). Each sample in the observed data corresponds to a target class label y_i showing the phenotype of the sample being studied. Thus $Y \in R^n$ is the vector of class labels where each $y_i \in \{0, 1\}$. The number of observations in such datasets is generally smaller than the number of features. This form is called small n, large p i.e. $n < p$ problem. Figure 1 shows a general layout of a gene expression data. Genes in the figure are shown in columns whereas observations are shown in the rows. Gene expression values for a gene are given in the cells of the table against each observation.

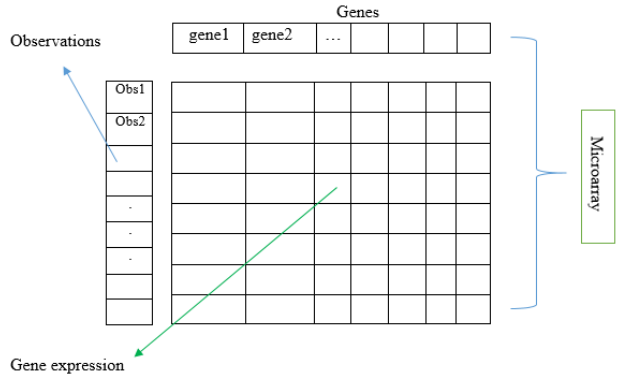

FIGURE 1. Gene expression data.

The proposed method in this paper initially identifies a minimum set of genes with the help of the greedy search approach given in [20]. This set of genes unambiguously assigns maximum number of training samples to their correct classes. Greedy search approach takes the following steps.

1) Gene mask for each gene is computed by calculating interquartile range (IQR) for observations in each class $c \in \{0, 1\}$ (binary class problem). The gene mask for each gene takes a value of either 0 or 1 where bit i of gene j is set to 1 if its expression value e_{ij} does not fall in the overlapped region of both classes, otherwise it is set to 0. By this way a matrix $M = [m_{ij}]_{p\times n}$ is constructed where

$$m_{ij} = \begin{cases} 1, & \text{if } e_{ij} \notin I_{j,1} \cap I_{j,2} \\ 0, & \text{otherwise,} \end{cases}$$

such that $I_{j,1} = [a_{j,1}, b_{j,1}]$ and $I_{j,2} = [a_{j,2}, b_{j,2}]$, where $a_{j,1} = Q_1^{(j,1)} - 1.5\ IQR^{(j,1)}$, $b_{j,1} = Q_1^{(j,1)} + 1.5\ IQR^{(j,1)}$, $a_{j,2} = Q_1^{(j,2)} - 1.5\ IQR^{(j,2)}$ and $b_{j,2} = Q_1^{(j,2)} + 1.5\ IQR^{(j,2)}$, respectively. In these expressions, $Q_1^{(j,c)}$, $Q_2^{(j,c)}$ and $IQR^{(j,c)}$ denote the first empirical quartile, second empirical quartile and interquartile range, respectively, of gene j for class c.

2) Using matrix M in Step 1, genes that have the highest number of bits equal to 1 are selected for the minimum subset. If more than one gene have the same number of bits 1, gene with the lowest POS-score is selected. For further details on POS-score, see [8].
3) Using logical AND operator, the gene masks of the remaining genes are updated and Step 2 is repeated to select the second gene. This process is iterated until the desired number of genes in the minimum set is selected or the genes have no ones in their gene masks.

In the second stage, the proposed method, GClust, divides the remaining genes that are not selected in the minimum subset of genes in the first phase into a certain number of clusters. For this purpose, one of the most popular descent clustering methods, the k-means algorithm is used. This algorithm uses the Euclidean distance to find dissimilarity among genes. The Euclidean distance between two genes x_j and x_j' is given as

$$d_{jj'} = d(x_j, x_j') = \sum_{i=1}^{n} (x_{ij} - x_{ij}')^2 = \|x_j - x_j'\|^2.$$

The within-point scatter for genes clusters is

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(j)=k} \sum_{C(j')=k} \|x_j - x_j'\|^2,$$
$$= \sum_{k=1}^{K} p_k \sum_{C(j')=k} \|x_j - \bar{x}_k\|^2,$$

where $\bar{x}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{nk})$ is the mean vector for the kth cluster, and $p_k = \sum_{j=1}^{p} I(C(j) = k)$ is the number of genes in the kth cluster.

The above criterion is minimized by assigning the p genes to the K clusters in such a manner that within each cluster the average dissimilarity of the genes from the cluster mean is minimized, where cluster mean is defined by genes in the given cluster.

For solving

$$C^* = \min_{C} \sum_{k=1}^{K} p_k \sum_{C(j)=k} \|x_j - \bar{x}_k\|^2, \tag{1}$$

an iterative descent algorithm can be utilized by noting that for any set of genes p'

$$\bar{x}_{p'} = \arg\min_{t} \sum_{j\in p_{\prime}} \|x_j - t\|^2. \tag{2}$$

By solving the enlarged optimization problem

$$\min_{C, \{t_k\}_1^K} \sum_{k=1}^{K} p_k \sum_{C(j)=k} \|x_j - t_k\|^2, \tag{3}$$

the value of C^* can be obtained.

The following iterative algorithm is used for genes assignment to the respective clusters.

1) For deciding on a given cluster assignment C, the expression given in (3), i.e. the total cluster variance is minimized w.r.t $(t_1, \ldots, t_K)$ giving the means of the currently assigned clusters (2).
2) Given $(t_1, \ldots, t_K)$, the current set of means, (3) is minimized by assigning each gene to the nearest (current) cluster mean. That is

$$C(j) = \underset{1\leq k\leq K}{\arg\min} \|x_j - t_k\|^2.$$

3) The above two steps are repeated until the genes assignment do not change.

The next step in the second stage of the proposed method is to select the most informative gene from each cluster. As a cluster consists of similar objects, it is worthwhile to mention that within the same cluster genes will be collinear. Keeping this into consideration, the propsed method uses the least absolute shrinkage and selection operator (lasso) [29] method to select the most informative gene. The lasso method shrinks the coefficients of non-informative genes exactly to zero by penalizing them and retains only those ones in the model that mainly regulate the target class. This method uses the following expression to estimate the genes coefficients

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0+\beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

In the above expression, β_j's are genes coefficients, λ is the size of penalty on the parameters and y_i is the binary response. For further details on the lasso procedure see [29].

Based on the above discussion, the proposed method, GClust, takes the following steps to select the most informative genes.

1) Select the minimum subset of genes that unambiguously assign maximum number of observations in the training data to their correct classes based on the greedy search approach given in [8].
2) Divide the remaining genes which are not selected in the minimum subset of genes into a certain number of clusters using K-means algorithm. The value of K depends on the number of genes the user wants to select. For example, if the user wants to select 20 genes and the minimum set has 2 genes, then the remaining genes will be divided into 18 clusters to get the total of 20 genes required.
3) Using the lasso method, select the most informative gene from each cluster discarding the non-informative ones.
4) Combine the minimum subset of genes with those selected in Step 3 to form the final set of genes.

The propsed method is novel in the sense that it uses the greedy search approach in conjunction with clustering that ensures high generalization power and at the same time minimizes redundancy in the selected set of genes. The number of genes selected in the minimum set ranges from 1 to 3 in the benchmark datasets considered in this paper that are differentially expressed, thus reducing the chances of multicollinearity. As the rest of the genes are selected via clustering and lasso, therefore, the final set of genes is less redundant as compared to genes selected by the competitors. This has been shown by constructing correlation matrices for the selected genes via all the methods in the next section.

IV. EXPERIMENTS AND RESULTS

For assessing various gene selection methods, one can assess the accuracy of a classifier applied after the gene selection procedure, where classification is based only on the chosen genes. Such evaluation can verify the discriminative ability of the selected genes. Jirapech and Aitken [30] have assessed various gene selection methods given in [31] and have shown that they can have a significant effect on a classifier's accuracy. This approach has been used in various studies including [20] and [32].

In the current paper, experiments are conducted using 7 gene expression datasets in which the proposed GClust method is validated by comparing it with other five state-of-the-art gene selection methods including proportional overlapping analysis (POS) [8], masked painter approach [20], Wilcoxon Rank Sum technique (Wil-RS) [33], [34], relative simplicity (RS) [21] method, double kernel-based clustering method for gene selection (DKBCGS) [28] and the minimum subset of genes selected via the greedy approach [8]. The performance is assessed by calculating the classification accuracy, sensitivity and Brier score from two different classifiers, random forest (RF) [35] and support vector machine (SVM). A brief description of these classifiers is given below.

Random forest [35] is an ensemble learning method consisting of a sufficiently large number of classification trees that allows the trees to vote for the target class based on the majority rule. Trees are grown on bootstrap samples taken from the given training data where binary splits are made, unlike the standard decision tree, on a random sample of features from the total features set at each node of the tree. This is done so as to inculcate additional randomness in the base learners, the classification trees.

Support vector machine (SVM) [36] is one of the most powerful machine learning methods that divides the given training data by finding a hyper-plane between the two classes that has the widest margins. Observations that lie on the edges of the plane are called the support vectors and are used to estimate the target class of the test observation.

Table 1 gives a brief summary of the datasets characteristics. The table shows number of genes, number of samples and class sizes for each data set. All the datasets are binary class problems taken from various open sources given in the last column of the table. The given datasets are divided into training and testing parts for feature selection and assessment via the classification algorithms. Seventy percent of each of the data sets is taken as training part and the remaining is used for testing purposes. Five hundred runs of the

TABLE 1. Summary of the benchmark datasets.

Datasets	Samples	Genes	Class sizes	Source
nki	144	76	96/48	[37]
Colon	62	2000	40/22	[25]
Breast	78	4948	34/44	[38]
Leukeamia	68	7029	47/25	[39]
Ova Ovary	542	10937	379/163	[41]
GSE4045	37	22215	29/8	[40]
AP Breast Colon	1081	22215	617/464	[41]

split sample analysis were performed for each combination of dataset and feature selection algorithm, with classifiers considered. Random forest is executed using the R package `randomForest` [42] with the default parameters values, i.e. `ntree=500`, `mtry=`$\sqrt{p}$ and `nodesize=1`. For support vector machine, the `kernlab` [43] R package is used with default settings.

Using the same training part of each data for the given combination of feature selection method and classification algorithm, 20 genes are selected by all the methods considered for training the classifiers and average values of the performance metrics, i.e. classification accuracy, sensitivity and Brier score are calculated using the testing parts. Results of GClust and the other methods considered are given in Tables 2, 3, 4, 5, 6, 7 and 8.

TABLE 2. Comparison of the GClust method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on leukaemia dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.9980**	0.9930	0.9021	0.9560	0.9526	0.9817	0.9542
	Sen.	**0.9966**	0.9931	0.9350	0.9614	0.9129	0.9234	0.9348
	BS	**0.0078**	0.0141	0.0772	0.0337	0.0641	0.0158	0.0109
SVM	Acc.	**0.9956**	0.9937	0.9358	0.9772	0.9549	0.9836	0.9045
	Sen.	**0.9966**	0.9715	0.9308	0.9262	0.7349	0.7769	0.9212
	BS	**0.0053**	0.0418	0.0828	0.0598	0.0895	0.0428	0.0240

TABLE 3. Comparison of the proposed method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on nki dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.8408**	0.6252	0.5909	0.6137	0.5884	0.6143	0.7809
	Sen.	**0.6142**	0.2545	0.2533	0.2430	0.4678	0.5436	0.5120
	BS	0.1569	0.0749	0.2652	**0.0637**	0.1230	0.0760	0.2129
SVM	Acc.	0.7250	**0.7580**	**0.7580**	0.6763	0.7200	0.7475	0.6094
	Sen.	0.3137	0.4484	0.3991	**0.4754**	0.4567	0.4643	0.2501
	BS	0.1817	**0.0058**	0.1665	0.0564	0.0534	0.0513	0.2674

From Table 2, it is clear that GClust method shows the best performance among all the methods on leukaemia dataset. The accuracy of the proposed method is 0.9980, while the accuracy of POS, Wilcoxon, MP, DKBCGS, RS and greedy minimum set (Min) are 0.9930,0.9021, 0.9560, 0.9526, 0.9817 and 0.9542, respectively. Thus, the GClust method has higher accuracy rate as compared to the other methods. Similarly, the performance of the proposed method and other three methods have been checked on the SVM classifier. With the SVM classifier, the proposed method showed better results than the others.

Sensitivity and BS of the proposed method is 0.9966, and 0.0078 respectively, while those of POS, Wilcoxon, MP,

TABLE 4. Comparison of the GClust method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on breast cancer dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.7874**	0.6973	0.6560	0.6788	0.6765	0.6866	0.7010
	Sen.	**0.8470**	0.8023	0.7824	0.7906	0.8165	0.7999	0.7527
	BS	**0.1586**	0.1996	0.2172	0.1810	0.2023	0.2006	0.2011
SVM	Acc.	**0.8087**	0.6416	0.5966	0.6229	0.6442	0.6426	0.7826
	Sen.	**0.8866**	0.7782	0.7443	0.7611	0.7349	0.7769	0.7394
	BS	0.1504	**0.0440**	0.2426	0.0630	0.0466	0.4511	0.2211

TABLE 5. Comparison of the GClust method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on colon dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.8716**	0.8674	0.8702	0.8544	0.8375	0.8480	0.8654
	Sen.	**0.8238**	0.7797	0.8200	0.7807	0.7996	0.8102	0.7180
	BS	**0.1044**	0.1115	**0.1044**	0.1219	0.1599	0.1126	0.1041
SVM	Acc.	**0.8620**	0.8357	0.8533	0.8555	0.7967	0.8250	0.8165
	Sen.	0.7372	0.7228	0.7722	**0.8139**	0.7865	0.7968	0.6737
	BS	0.0981	0.0134	0.1214	0.1077	0.0604	**0.0125**	0.1351

TABLE 6. Comparison of the GClust method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on GSE4045 dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.9680**	0.7797	0.7466	0.7855	0.7592	0.7698	0.8445
	Sen.	**0.8484**	0.1582	0.1762	0.1354	0.5187	0.5908	0.5353
	BS	**0.0411**	0.1447	0.1495	0.1635	0.1473	0.1458	0.1077
SVM	Acc.	**0.9830**	0.7766	0.7592	0.7933	0.7611	0.7706	0.8670
	Sen.	**0.9824**	0.0003	0.0018	0.0840	0.6894	0.7347	0.4364
	BS	**0.0080**	0.1712	0.1886	0.1274	0.1738	0.1723	0.0858

TABLE 7. Comparison of the GClust method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on AB Breast Colon dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.9886**	0.9684	0.9024	0.9332	0.9446	0.9350	0.8646
	Sen.	0.9505	**0.9526**	0.8870	0.9492	0.9292	0.9481	0.7901
	BS	**0.0216**	0.0238	0.0734	0.0272	0.0485	0.0282	0.1204
SVM	Acc.	**0.9704**	0.9684	0.8792	0.8396	0.8404	0.9601	0.7302
	Sen.	**0.9440**	0.9420	0.9335	0.8061	0.9328	0.8865	0.8294
	BS	0.0218	**0.0008**	0.0890	0.0142	0.0935	0.0503	0.1178

TABLE 8. Comparison of GClust method with POS, Wilcoxon, MP, DKBCGS, RS and the minimum set by greedy methods on OVA Ovary dataset using RF and SVM classifiers.

Classifier		Methods						
		GClust	POS	Wilcoxon	MP	DKBCGS	RS	Min
RF	Acc.	**0.9710**	0.9384	0.7538	0.8674	0.8935	0.8312	0.8004
	Sen.	**0.9074**	0.6780	0.4733	0.6210	0.4686	0.8279	0.7305
	BS	0.0687	0.0466	0.1450	0.2129	**0.0384**	0.1559	0.1059
SVM	Acc.	**0.9450**	0.9384	0.8244	0.8964	0.9139	0.8699	0.8178
	Sen.	**0.8261**	0.6780	0.5496	0.6151	0.5842	0.6339	0.5098
	BS	0.0158	**0.0010**	0.1522	0.0876	0.1046	0.2174	0.1390

DKBCGS, RS and greedy minimum set (Min) are 0.9931, 0.9350, 0.9614, 0.9129, 0.9234 and 0.9348, and 0.0141, 0.0772, 0.0337, 0.0641, 0.0158 and 0.0109, respectively. Thus, the results revealed that GClust outperforms the other methods in terms of all the metrics used.

Table 3 shows the performance of the methods with random forest and SVM classifiers using nki dataset. From the table it is clear that the proposed method has better performance in terms of sensitivity and accuracy. The sensitivity of the proposed method is 0.6142, while that of POS, Wilcoxon, MP DKBCGS, RS and greedy minimum set (Min) are 0.2545, 0.2533, 0.2430, 0.4678, 0.5436 and 0.7879, respectively. The BS of the proposed method is 0.1590, while that of the POS, Wilcoxon, MP, DKBCGS, RS and greedy minimum set (Min) are 0.07497, 0.2652, 0.0637, 0.1230, 0.0760 and 0.2129, respectively. The classification accuracy of the proposed method is the highest among all the methods, i.e. 0.8408, while the classification accuracy of POS, Wilcoxon, MP, DKBCGS, RS and greedy minimum set (Min) are 0.6252, 0.5909, 0.6137, 0.5884, 0.6143 and 0.7809, respectively. Thus overall, the results indicate that the proposed method has performed better than the others. The BS of the POS method is smaller than those of the other methods in the case of nki data set. The SVM classifier results suggest that the POS method is the best in terms of BS and classification accuracy.

From Table 4, random forest classifier results indicate that the accuracy and sensitivity of the proposed method are higher than those of the other five methods. Besides, the results of the proposed method in terms of BS are also better than the others. Similarly, in the case of the SVM classifier, sensitivity of the GClust method is higher than those of the other methods. The GClust method gave 0.8866 sensitivity on the SVM classifier using breast cancer dataset. The BS of the POS method is smaller than those of the other methods. The accuracy of the proposed method is higher than the other five methods using the SVM classifier.

Table 5 displays the results of the methods on colon dataset via RF and SVM classifiers. The results indicate that, overall, the GClust method is better than the other methods. Both sensitivity and BS of the proposed method are better than those of the other five approaches in the case of random forest classifier. RS performed well in terms of BS via SVM,

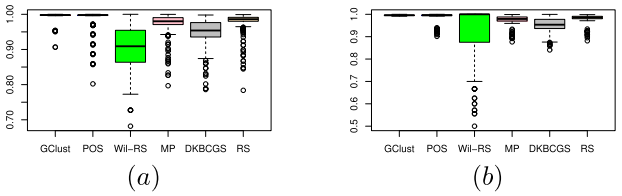

FIGURE 2. Box plots for classification accuracies via (a): RF, (b): SVM classifiers on leukaemia data with selected genes by the given methods.

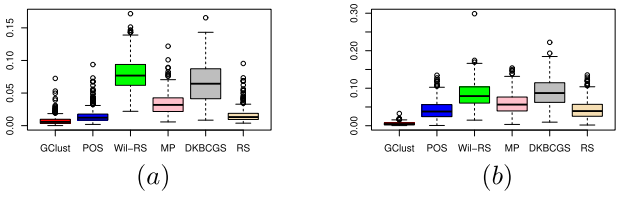

FIGURE 3. Box plots for Brier score via (a): RF, (b): SVM classifiers on leukaemia data with selected genes by the given methods.

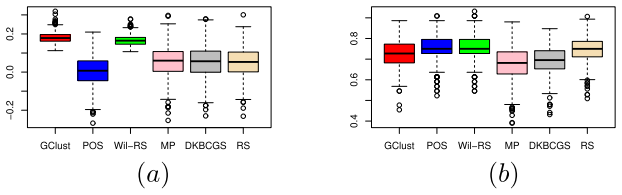

FIGURE 4. Box plots for Brier score via (a): RF, (b): SVM classifiers on nki data with selected genes by the given methods.

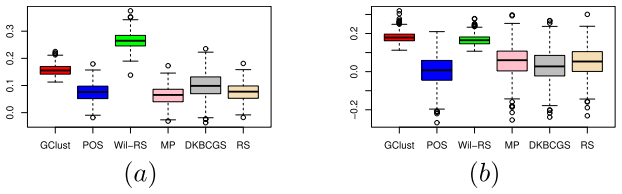

FIGURE 5. Box plots for Brier score via (a): RF, (b): SVM classifiers on nki data with selected genes by the given methods.

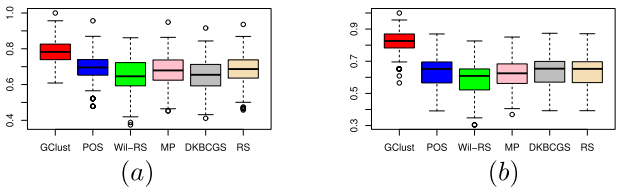

FIGURE 6. Box plots for classification accuracy via (a): RF, (b): SVM classifiers on breast cancer data with selected genes by the given methods.

Wilcoxon gave similar result to that of the GClust in terms of BS via random forest and MP method performed better than the other methods in term of sensitivity via SVM classifier.

Table 6 displays the results of the methods on GSE4045 dataset with the selected 20 genes. In this case, GClust has outperformed all the other methods in terms of all the performance metrics used for both RF and SVM classifiers.

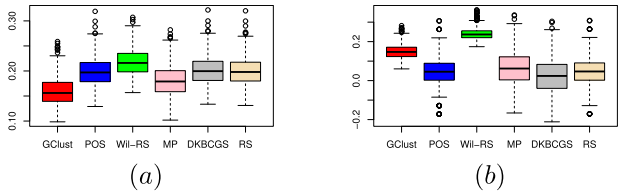

FIGURE 7. Box plots for Brier score via (a): RF, (b): SVM classifiers on breast cancer data with selected genes by the given methods.

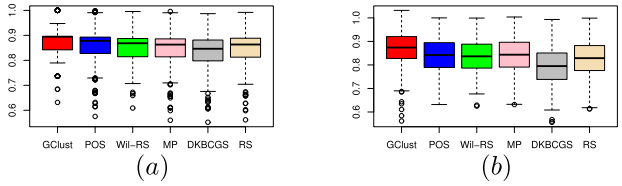

FIGURE 8. Box plots for classification accuracy via (a): RF, (b): SVM classifiers on colon data with selected genes by the given methods.

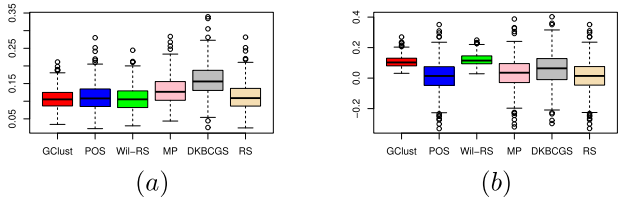

FIGURE 9. Box plots for Brier score via (a): RF, (b): SVM classifiers on colon data with selected genes by the given methods.

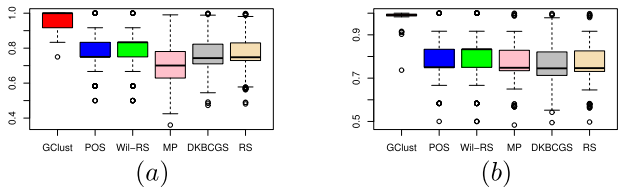

FIGURE 10. Box plots for classification accuracy via (a): RF, (b): SVM classifiers on GSE4045 data with selected genes by the given methods.

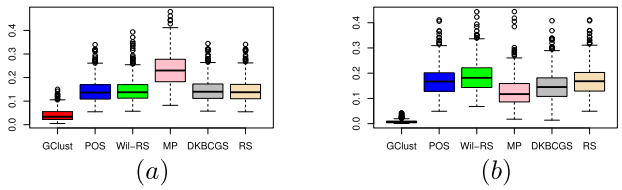

FIGURE 11. Box plots for Brier score via (a): RF, (b): SVM classifiers on GSE4045 data with selected genes by the given methods.

From Tables 7 and 8, it is obvious that the proposed method is giving better results than the others on most of the statistics considered.

To further assess the proposed method, GClust, in comparison with the other methods, boxplots of the results obtained, i.e. classification accuracy and Brier score, from all the 500 runs made for both SVM and random forest classifiers are constructed. All the methods are allowed to select 20 genes

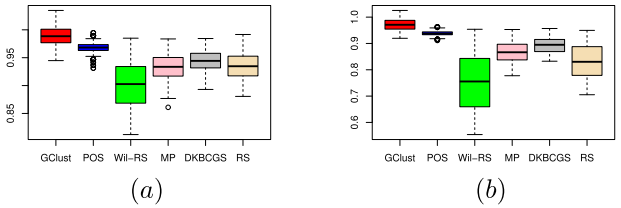

FIGURE 12. Classification accuracy of RF classifier on (a):AB Breast Colon (b): OVA Ovary dataset.

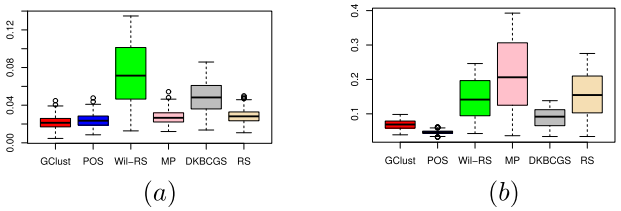

FIGURE 13. Brier scores of RF classifier on (a):AB Breast Colon (b): OVA Ovary dataset.

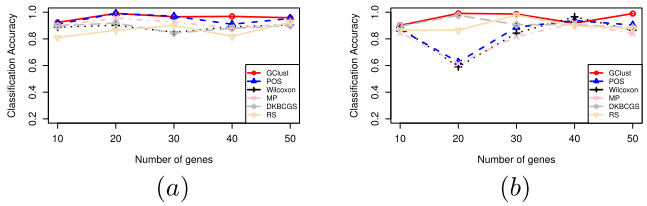

FIGURE 14. Classification accuracy of RF classifier on (a): leukemia dataset (b): nki dataset, with various number of selected genes by the methods.

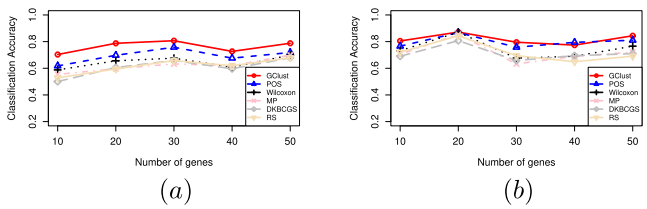

FIGURE 15. Classification accuracy of RF classifier on (a): breast cancer dataset (b): colon dataset, with various number of selected genes by the methods.

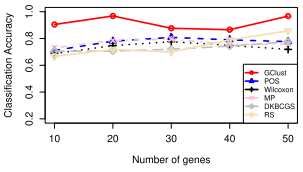

FIGURE 16. Classification accuracy of RF classifier on GSE4045 dataset with various number of selected genes by the methods.

and then the reduced datasets are used with the classifiers to calculate the desired performance measures. These plots are shown in Figure 2-13. The plots show that the proposed GClust method is better than the rest of the methods considered in most of the cases.

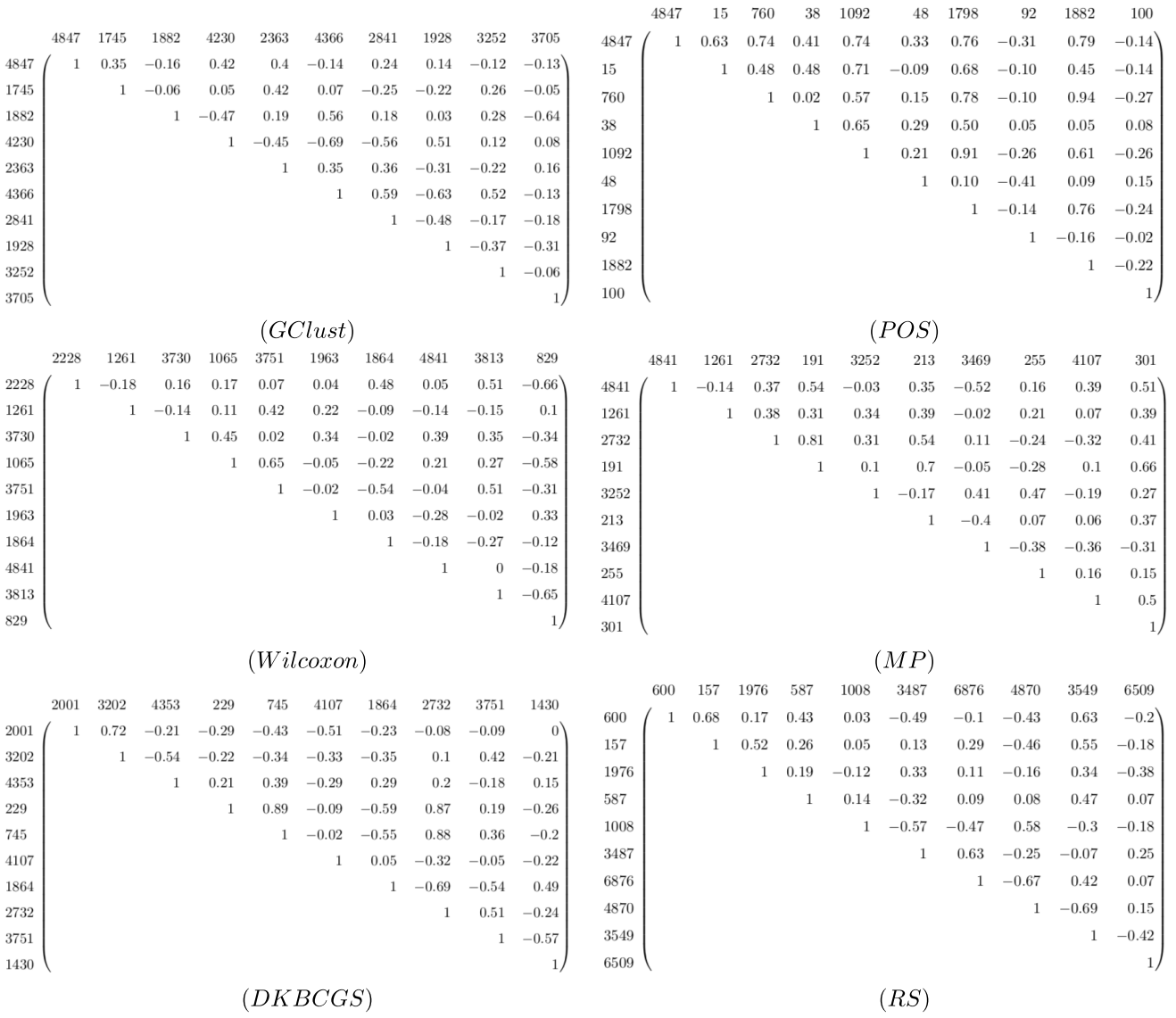

FIGURE 17. Correlation matrices of genes selected by the method applied on leukemia dataset. Row and column labels indicate gene indices.

To assess the stability of the proposed GClust approach in comparison with the other approaches, the gene selection methods were allowed to select different number of genes and the classifiers' performance on the reduced data are recorded. This is shown in Figures 14, 15 and 16. The number of genes selected i.e. 10, 20, 30, 40 and 50 are shown on the x-axis while classification accuracy is shown on the y-axis. The figures demonstrate that GClust gives the most stable results among all the other methods.

Correlation matrices for the top 10 genes selected by all the methods on leukemia and breast cancer datasets are also constructed. The matrices given in Figures 17 and 18 are the correlation matrices constructed from genes selected by all the six methods applied to leukemia and breast cancer datasets, respectively. Row and column labels in the matrices show gene indices. The matrices show that genes selected by the GClust method are less correlated than those selected by the other methods.

V. DISCUSSION

Keeping in view the need for handling high dimensional datasets with tens of thousands of features, this paper has proposed a novel method, GClust, for features selection to reduce data dimensions. In the current paper gene expression data generated via microarray technology are considered for analysis. The proposed method starts initially by identifying genes that can assign the maximum number of observations to their correct classes using a greedy approach. The greedy approach exploits interquartile range and proportional overlapping analysis for the identification. Genes selected in this way make the minimum set of genes. The proposed method then divides the remaining genes that are not selected in the

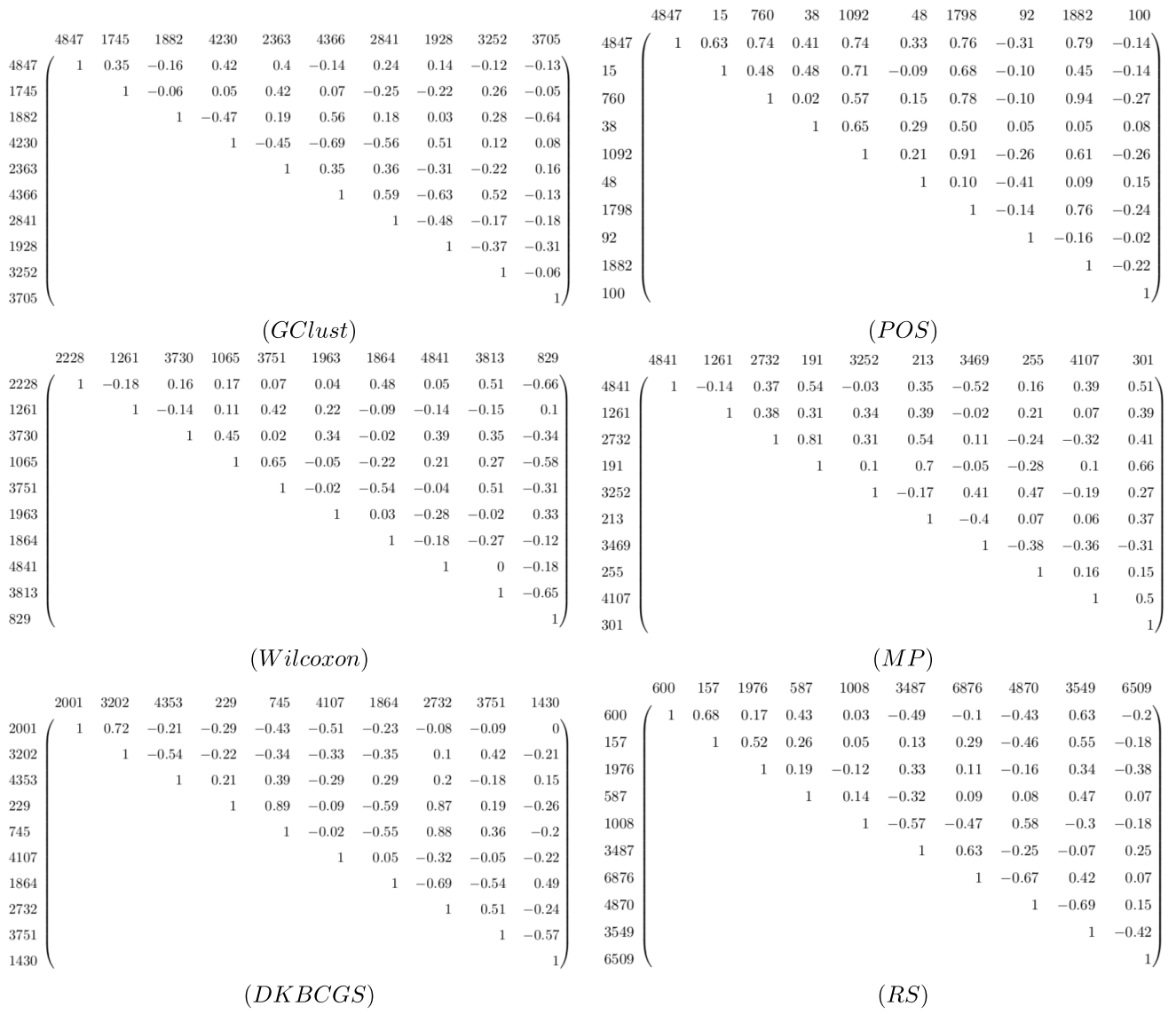

FIGURE 18. Correlation matrices of genes selected by the method applied on leukemia dataset. Row and column labels indicate gene indices.

minimum subset of genes, into a certain number of clusters. From each cluster the most informative gene is selected via the lasso method and combined with genes in the minimum subset to form the final set of genes. A total of 7 gene expression datasets are considered for assessing the proposed method. Twenty genes are selected by the method with the training parts of the datasets and data with reduced dimensions are used to obtain classification accuracy, sensitivity and Brier score via random forest and support vector machine classifiers on the testing parts. The results of the proposed method are compared with those of proportional overlapping analysis (POS), masked painter approach, Wilcoxon Rank Sum technique (Wil-RS), relative simplicity (RS) method, double kernel-based clustering method for gene selection (DKBCGS) and the minimum subset of genes selected via the greedy approach. A total of 500 random partitions have been made to calculate the performance metrics. Boxplots from the results have been constructed and average values of the results are shown for the method against the other methods. The analyses have revealed that the proposed method has outperformed the other methods in most of the cases. GClust has achieved the highest values of classification accuracy most of the times for the datasets considered among all the other methods via both random forest and support vector machine classifiers. The proposed method has also used the Brier score as performance measure which carries small values when the predicted class membership probabilities are close to the actual response values expressed in the 0, 1 form. This gives a measure of the degree of belief in the predicted probability of an observation belonging to a particular target class. GClust has outperformed the other methods by achieving the smallest values of Brier score in most of the cases given in the paper.

Similar conclusion could be drawn regarding the efficiency of the GClust method based on sensitivity analysis given in the paper. Moreover, the proposed method is the most stable method among all the other methods and genes selected by GClust have smaller correlations as compared to those selected by the other methods.

The intuition behind the efficient performance of the proposed method is that it first selects genes that can unambiguously assign maximum number of observations to their correct classes and then discard those genes that are redundant and cause the problem of multicollinearity. This is achieved by using the lasso method to select genes from the clusters formed in the second phase of the algorithm. Application of the lasso method to clusters might be time consuming if the number of clusters or size of individual clusters or both are large. This step can be made faster with the help of parallel computing using the R programming language as given in the R package `parallel` [44], for instance.

There could be several possibilities for future work in the direction of the proposed method. A possibility for further improvements in the proposed method might be using correlation analysis in conjunction with cluster analysis for genes selection from the set of genes not selected in the minimum subset of genes. The number of genes in the current paper is chosen to be 20 for simplicity, this could be optimised by deciding on the optimum number of clusters with the help of using an objective function on clustering like the Elbow method [45], for example. The case of continuous predictor variables as given by the expression values of genes taken from tissue samples of human bodies is considered in this paper. The method could be modified in a way so as to cover the case of categorical variables in the predictor set, such as gender, eye/hair colour, age group, etc. Moreover, the work proposed in this article can also be extended to solve big data problems as given in [46]–[49]. One can also modify the proposed method to deal with multi-class problems.

REFERENCES

[1] J. R. Díaz-Uriarte and S. A. de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinform.*, vol. 7, no. 1, p. 3, 2006.

[2] P. Li, C. Dai, and W. Wang, "Inconsistent data cleaning based on the maximum dependency set and attribute correlation," *Symmetry*. vol. 10, no. 10, p. 516, 2018.

[3] L. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 1, pp. 40–53, Jan. 2007.

[4] F. Chu and L. Wang, "Applications of support vector machines to cancer classification with microarray data," *Int. J. Neural Syst.*, vol. 15, no. 6, pp. 475–484, 2005.

[5] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recognit.*. vol. 43, no. 8, pp. 2763–2772, Aug. 2010.

[6] M. Yamada, J. Tang, J. Lugo-Martinez, E. Hodzic, R. Shrestha, A. Saha, H. Ouyang, D. Yin, H. Mamitsuka, C. Sahinalp, and P. Radivojac, "Ultra high-dimensional nonlinear feature selection for big biological data," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1352–1365, Jul. 2018.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2017.

[8] O. Mahmoud, A. Harrison, A. Perperoglou, A. Gul, Z. Khan, M. Metodiev, and B. Lausen, "A feature selection method for classification within functional genomics experiments based on the proportional overlapping score," *BMC Bioinform.*, vol. 15, no. 1, p. 274, 2014.

[9] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *J. Amer. Stat. Assoc.*, vol. 105, no. 490, pp. 713–726, 2010.

[10] C. Augenstein, N. Spangenberg, and B. Franczyk, "Applying machine learning to big data streams: An overview of challenges," in *Proc. ISCMI*, vol.105, Nov. 2017, pp. 25–29.

[11] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, Jan. 2002.

[12] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.

[13] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Inf. Sci.*, vol. 179, no. 13, pp. 2208–2217, 2009.

[14] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.

[15] L. Breiman, *Classification and regression trees*. Evanston, IL, USA: Routledge, 2017.

[16] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognit.*, vol. 40, no. 11, pp. 3236–3248, Nov. 2007.

[17] J. C. H. Hernandez, B. Duval, and J.-K. Hao, "A genetic embedded approach for gene selection and classification of microarray data," in *Proc. Eur. Conf. Evol. Comput., Mach. Learn. Data Mining Bioinform.*, vol. 11. Berlin, Germany: Springer, 2007, pp. 90–101.

[18] M. Dramiński, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, no. 1, pp. 110–117, Jan. 2007.

[19] A. Ultsch, C. Pallasch, E. Bergmann, and H. Christiansen, "A comparison of algorithms to find differentially expressed genes in microarray data," in *Advances in Data Analysis, Data Handling and Business Intelligence*. Berlin, Germany: Springer, 2009, pp. 685–697.

[20] D. Apiletti, E. Baralis, G. Bruno, and A. Fiori, "MaskedPainter: Feature selection for microarray data analysis," in *Proc. EMBS*, 2007, pp. 4227–4230.

[21] Y. Chen, L. Wang, L. Li, H. Zhang, and Z. Yuan, "Informative gene selection and the direct classification of tumors based on relative simplicity," *BMC Bioinform.*, vol. 17, no. 1, p. 44, 2016.

[22] M. Marczyk, R. Jaksik, A. Polanski, and J. Polanska, "Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition," *BMC Bioinform.*, vol. 14, no. 1, p. 101, 2013.

[23] J. Lu, R. Kerns, S. Peddada, and P. R. Bushel, "Principal component analysis-based filtering improves detection for Affymetrix gene expression arrays," *Nucl. Acids Res.*, vol. 39, no. 13, p. e86, 2011.

[24] W. Talloen, D.-A. Clevert, S. Hochreiter, D. Amaratunga, L. Bijnens, S. Kass, and H. W. H. Göhlmann, "I/NI-calls for the exclusion of non-informative genes: A highly effective filtering tool for microarray data," *Bioinformatics*, vol. 23, no. 21, pp. 2897–2902, 2007.

[25] 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," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[26] A. Ultsch, C. Pallasch, E. Bergmann, and H. Christiansen, "A comparison of algorithms to find differentially expressed genes in microarray data," in *Advances in Data Analysis, Data Handling and Business Intelligence*. Berlin, Germany: Springer, 2009, pp. 685–697.

[27] M. Vaidya and P. S. Kulkarni, "Innovative technique for gene selection in microarray based on recursive cluster elimination and dimension reduction for cancer classification," *Int. J. Innov. Res. Adv. Eng.*, vol. 1, no. 6, pp. 209–213, Jul. 2014.

[28] S. Liu, C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, and M. Dehmer, "Feature selection of gene expression data for Cancer classification using double RBF-kernels," *BMC Bioinform.*, vol. 19, no. 1, p. 396, 2018.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[30] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," *BMC Bioinform.*, vol. 6, no. 1, p. 148, 2005.

[31] Y. Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, "RankGene: Identification of diagnostic genes based on expression data," *Bioinformatics*, vol. 19, no. 12, pp. 1578–1579, 2003.
[32] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[33] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.
[34] C. Liao, S. Li, and Z. Luo, "Gene selection for cancer classification using Wilcoxon rank sum test and support vector machine," in *Proc. Int. Conf. Comput. Inf. Sci.* Berlin, Germany: Springer, Nov. 2006, pp. 57–66.
[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[36] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[37] J. J. Goeman, "L_1 penalized estimation in the Cox proportional hazards model," *Biometrical J.*, vol. 52, no. 1, pp. 80–84, 2010.
[38] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: A multiple random validation strategy," *Lancet*, vol. 365, no. 9458, pp. 488–492, Feb. 2004.
[39] 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, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.
[40] P. Laiho, A. Kokko, S. Vanharanta, R. Salovaara, H. Sammalkorpi, H. Järvinen, J.-P. Mecklin, T. J. Karttunen, K. Tuppurainen, V. Davalos, S. Schwartz, Jr., D. Arango, M. Mäkinen, and L. A. Aaltonen, "Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis," *Oncogene*, vol. 26, no. 2, pp. 312–320, 2007.
[41] G. Stiglic and P. Kokol, "Stability of ranked gene lists in large microarray analysis studies," *J. Biomed. Biotechnol.*, vol. 2010, May 2010, Art. no. 616358.
[42] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[43] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "Kernlab—An S4 package for kernel methods in R," *J. Stat. Softw.*, vol. 11, no. 9, pp. 1–20, 2004. [Online]. Available: http://www.jstatsoft.org/v11/i09/
[44] R Foundation for Statistical Computing, Vienna, Austria. (2017). *R: A Language and Environment for Statistical Computing*. [Online]. Available: https://www.R-project.org/
[45] P. Bholowalia and A. Kumar,"EBK-means: A clustering technique based on elbow method and k-means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 17–24, 2014.
[46] M. Marjani, F. Nasaruddin, A Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob, "Big IoT data analytics: Architecture, opportunities, and open research challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017.
[47] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.
[48] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 777–797, 2017.
[49] A. Akusok, K.-M. Björk, Y. Miche, and A. Lendasse, "High-performance extreme learning machines: A complete toolbox for big data applications," *IEEE Access*, vol. 3, pp. 1011–1025, 2015.

ZARDAD KHAN received the master's degree in statistics with distinction from the University of Peshawar, Pakistan, in 2008, the M.Phil. in statistics from Quaid-i-Azam University Islamabad, Pakistan, in 2011, and the Ph.D. degree in statistics from the University of Essex, U.K., in 2015. He is an Assistant Professor of statistics with Abdul Wali Khan University Mardan, Pakistan. He has also done a 1 year postdoctoral study from the University of Essex, UK, and is currently a Visiting Fellow in the same University. Zardad's focus is on machine learning, applied statistics, computational statistics, biostatistics, graphical modelling, and survival analysis.

MUHAMMAD NAEEM received the bachelor's degree in statistics, the master's in statistics from Islamia College University Peshawar, Pakistan, in 2010 and 2014, respectively. He is pursuing the Ph.D. degree with the Department of Statistics, Abdul Wali Khan University Mardan, Pakistan. His focus is on linear models, machine learning, applied statistics, and computational statistics.

UMAIR KHALIL received the bachelor's degree in statistics, the master's and Ph.D. degrees in statistics from the University of Peshawar, Pakistan, in 2000, 2003, and 2013, respectively. He is an Assistant Professor of statistics with Abdul Wali Khan University Mardan, Pakistan. His focus is on biostatics, applied statistics, survival analysis, quality control, and computational statistics.

DOST MUHAMMAD KHAN received the bachelor's degree in statistics, the master's and Ph.D. degrees in statistics from the University of Peshawar, Pakistan, in 2000, 2003, and 2012, respectively. He is an Assistant Professor of statistics with Abdul Wali Khan University Mardan, Pakistan. His focus is on robust statistics, applied statistics, survival analysis, statistical inference, and computational statistics.

SAEED ALDAHMANI received the bachelor's degree in statistics from United Arab Emirates University, UAE, in 2007, the master's degree in statistics from Macquarie University, Australia, in 2010, the master's degree in applied finance from Western Sydney University, Australia, in 2011, and the Ph.D. degree in statistics from the University of Essex, U.K., in 2017. He is currently an Assistant Professor of statistics with United Arab Emirates University. His focus is on graphical models, biostatistics, applied statistics in finance, and computational statistics.

MUHAMMAD HAMRAZ received the bachelor's degree in statistics, the master's and M. Phil. degrees in statistics from Quaid Azam University Islamabad, Pakistan, in 2004, 2007 and 2010 respectively. He is an Assistant Professor of statistics with Abdul Wali Khan University Mardan, Pakistan. His focus is on linear models, biostatics, applied statistics, and computational statistics.