Combined mRMR Filter and Sparse Bayesian Classifier for Analysis of Gene Expression Data

Mehran Soltani
Dept. of Electrical Engineering
Amirkabir University of Technology
Tehran, Iran
Mehran.soltani1994@gmail.com

Mohammad Hasan Shammakhi
Dept. of Electrical Engineering
Amirkabir University of Technology
Tehran, Iran
Mh.shammakhi@aut.ac.ir

Saeed Khorram
Dept. of Electrical Engineering
Amirkabir University of Technology
Tehran, Iran
Khorram.saeed@aut.ac.ir

Hamid Sheikhzadeh
Dept. of Electrical Engineering
Amirkabir University of Technology
Tehran, Iran
hsheikh@aut.ac.ir

***Abstract*—Many disorders can be diagnosed by analysis of gene expression microarrays and this can save lots of lives. However, as gene expression data have high dimensions, establishing a method to identify the genes related to the target disease still remains a challenge, because it should provide a well-grounded prediction about the disease status. To this end, the best subset of genes should be distinguished for the classification task. In this paper, we have introduced a new framework for the analysis of gene expression data. Our proposed algorithm tries to find the best feature subset, in two main stages. First, an information theoretic forward feature selection algorithm called mRMR (minimum redundancy, maximum-relevancy) is used to find a candidate set for best features. In the next stage, the RVM (Relevance Vector Machine) classifier which is well suited for gene data analysis is utilized. The RVM has frequent privileges over other classifiers, namely, it can return a membership probability for each class that can be very vital for diagnosis of dramatic diseases, and it can also lead to a more sparse approach to fit a model over the training data which will help to avoid overfitting, etc. The Experimental results showed that the proposed algorithm outperforms the previous works in both classification accuracy and sparsity of the model.**

Keywords— Gene expression; mRMR; RVM; Feature selection; Sparse model;

I. INTRODUCTION

Recently, DNA gene expression microarrays have been extensively used in bioinformatics, aiming to diagnose disorders and diseases, e.g. discriminating cancers from normal tissues, distinguishing one cancer sub-type from another, etc. These microarrays include expression of thousands of genes simultaneously, while only a small number of them have a strong correlation with the targeted phenotype [1]. To that end, a requisite issue in the analysis of this kind of data is to find an appropriate algorithm for selection of the most important features.

In the feature selection process, the purpose is to reduce the dimension of data by removing the most irrelevant or redundant features. Evidently, this can lead to a reduction in computational cost, as well as improvement in classification accuracy [2]. Generally, three types of feature selection methods have been investigated in the literature: wrappers, filters and embedded methods. Wrapper methods search for the features based on a specific learning algorithm while in filter methods, feature selection is independent of any specific learning algorithm and is generally considered as a preprocessing step [3]. In addition, embedded methods [4] select the features during the learning process. It is worth mentioning that Wrapper methods generally perform better but with more computational complexity.

Furthermore, various approaches have been used for gene selection in classification of cancers. Ding and Peng [2], [5] have presented a filter-based method of minimum-redundancy, maximum relevancy (mRMR) to find the optimal subset of genes. Additionally, another algorithm proposed by Li [6] have combined the k-nearest neighbor with genetic algorithms. Guyon *et al.* [7] proposed a method based on SVM which is called Recursive Feature Elimination (RFE). Although the aforementioned approaches showed great abilities in dimension reduction of gene expressions data, the proposed algorithm outperforms the previous works in terms of accuracy and sparsity.

In this paper, we have investigated the combination of a filter-based method with a wrapper algorithm to benefit from both high classification accuracy and low complexity. The proposed method tries to find the best subset of features in two stages. First, we have used the statistical forward feature selection method, mRMR, to find a candidate for the best subset of features. This method selects the genes based on their highest relevancy with the target class, as well as their least similarity to each other. In the next stage, the most irrelevant features are eliminated for the given set through a backward-selection scheme regarding the classification error. In this stage, we have used the RVM classifier [8], [9] as a sparse kernel based Bayesian algorithm.

Furthermore, another kernel based algorithm is the SVM [10] which have been used widely in classification of gene expression data showing satisfactory results [11], [12], [13]. Therefore, to demonstrate the generalizability of the proposed method, the SVM classifier is also used in the proposed method. This also resulted in a good performance comparing to previous works.

These two kernel-based algorithms find the output based on a linear combination of some kernel functions:

$$y(\mathbf{x} \mid \mathbf{w}) = w_0 + \sum_{n=1}^{N} w_n \, \mathrm{K}(\mathbf{x}, \mathbf{x_n}) \, , \qquad (1)$$

where $\boldsymbol{w} = (w_1, w_2, \dots, w_N)^T$ is the vector of weights and $\mathrm{K}(\mathbf{x}, \mathbf{x}_n)$ is a predetermined nonlinear kernel function.

The SVM results in a sparse model by simultaneously minimizing the error on training data and maximizing the margin between the two classes. This algorithm constrains the width of the margin by selecting some data points which are called support vectors.

The RVM using a Bayesian framework, offers some advantages over the SVM. The SVM algorithm is prone to overfitting as the number of the support vectors can be very large. In contrast, the RVM algorithm is more robust against overfitting by defining some zero-mean Gaussian priors over the model weights governed by hyperparameters. Also, the prediction of RVM is probabilistic for regression and classification with much fewer nonzero kernel functions, i.e., as mentioned in [8], the RVM results in a sparser model in comparison with SVM.

The structure of the paper is as follows. First, in Section II and III, mRMR feature selection and RVM classifier are introduced. Afterwards, in Section IV, the proposed method is discussed in more details. Subsequently, Section V, evaluates the experiment results on two popular gene expression data sets. Finally, Section VI concludes the paper.

II. MRMR FEATURE SELECTION

In Mutual information based feature selection algorithms, the main purpose is to find a feature set S with m features having maximum dependency on the target class c. This criterion called “Maximal-Dependency” tries to maximize the following function, D:

$$\max \;\; D(S,c) \quad D = I(\{x_i, i = 1,2,\ldots,m\}; c) \, , \qquad (2)$$

where the mutual information $I(x,y)$, for two given random variables x and y, is defined in terms of their probability density functions $p(x)$, $p(y)$ and $p(x,y)$ as:

$$I(x,y) = \int\!\!\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \qquad (3)$$

Using maximal-dependency criterion defined in (2), mutual information relationship in (3) can be written as:

$$\begin{aligned} I(S_m; c) &= \int\!\!\int p(S_m, c) \log \frac{p(S_m, c)}{p(S_m)p(c)} dS_m dc \\ &= \int\!\!\int p(S_{m-1}, x_m, c) \log \frac{p(S_{m-1}, x_m, c)}{p(S_{m-1}, x_m)p(c)} \\ &= \int\!\!\int p(x_1, \ldots, x_m) \log \frac{p(x_1, .. x_m, c)}{p(x_1, .. x_m)p(c)} dx_1 .. dx_m dc \end{aligned} \qquad (4)$$

It is often difficult to calculate the joint probability function because of two main reasons: firstly, there is an inadequate number of samples for computation of joint probability distribution. Secondly, to compute the joint probability function of two variables, the inverse of the covariance matrix is required, which in high dimensions, it is computationally expensive. Therefore, Max-Dependency criterion is practically not efficient.

To tackle this problem, another approach called “Maximum-Relevancy” is introduced in [5]. This approach selects the features based on maximization of function $D(S, c)$ as the mean of the mutual information of all features $\{x_i\}$ with target class c. By considering $\{S\}$ as the selected subset, the D function is redefined as:

$$\max \;\; D(S,c) \quad , \text{where} \;\; D(S,c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, c) \qquad (5)$$

One can realize that maximum-relevancy criterion finds the most relevant feature subset in respect of the target class c. However, the dependency between the selected features is still inevitable. Moreover, elimination of redundant features results in a more compact feature set without any sensible change on the discrimination analysis. Therefore, the following “Minimal-Redundancy” can be applied to find mutual information between different pairs of features:

$$\min \;\; R(S) \quad , \; R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \qquad (6)$$

The mRMR feature selection is defined by combination of mentioned criterions:

$$\max \quad \Phi(D, R) \, , \qquad \Phi = D - R \qquad (7)$$

Suppose that m-1 features are selected. The task is to find the m^{th} feature from $\{X - S_{m-1}\}$. The selected feature should satisfy maximization of $\Phi(.)$. Therefore, the m^{th} feature will be selected by maximization of the following condition [5]:

$$\max_{x_j \in X - S_{m-1}} [I(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_i, x_j)] \qquad (8)$$

Different wrappers can be combined with this procedure to have both forward and backward feature selection. This approach will decrease both the model complexity and classification error. In this paper, we have combined the mRMR with a backward wrapper method using the sparse Bayesian classifier, Relevance Vector Machine, which is introduced in more detail below.

III. RELEVANCE VECTOR MACHINE

In the RVM-based classification, case [8], [9], considering a two-class problem, it is desired to find the probability that an input x belongs to a specific target label $t \in \{0,1\}$.

To construct a kernel-based model for the classification task, a logistic sigmoid function is applied to the linear combination of basis functions as:

$$y(x, w) = \sigma(w^T \varphi(x)) \, , \qquad (9)$$

where $\sigma(.)$ is the sigmoid function:

$$\sigma(y) = 1/(1+\exp(-y)) \tag{10}$$

The RVM assumes zero-mean Gaussian independent priors for the weights:

$$p(\mathbf{w}\,|\,\boldsymbol{\alpha}) = \prod_{i=1}^{M} N(w_i\,|\,0,\alpha_i^{-1}) \ , \tag{11}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ is the vector including the inverse of the variance of the distribution α_i, i.e. the precision of the corresponding w_i. They are called hyperparameters of the model.

The RVM estimates the value of these hyperparameters by maximizing the marginal likelihood $p(\{t_n\}_{n=1}^{N}\,|\,\boldsymbol{\alpha})$. During this process, many of the α_is tend towards infinity, that is their corresponding weights tend towards zero. In this algorithm, the distribution over the observations is considered as a Bernoulli function in the following form:

$$P(t\,|\,\mathbf{w}) = \prod_{n=1}^{N} \sigma\{y(x_n;\mathbf{w})\}^{t_n}[1-\sigma\{y(x_n;\mathbf{w})\}]^{1-t_n} \tag{12}$$

The final stage is to solve an optimization problem and find the values of the parameters ($\boldsymbol{w}$) and hyperparameters ($\boldsymbol{\alpha}$). According to [8], a Laplace approximation is used because the weights cannot be integrated analytically. For this purpose, using an initialized value of $\boldsymbol{\alpha}$, the Bernoulli distribution in (10) is approximated with a Gaussian distribution. Thereby, the resulting approximation of the posterior distribution by Expectation- Maximization (EM) algorithm gives the mean and covariance of the Laplace approximation in the following forms:

$$\mathbf{w}^{*} = \mathbf{A}^{-1}\mathbf{\Phi}^{T}(\mathbf{t}-\mathbf{y}), \tag{13}$$

$$\sum = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{B}\mathbf{\Phi}+\mathbf{A})^{-1} \ , \tag{14}$$

Where $\mathbf{A} = diag(\alpha_i), \mathbf{B}$ is a $N \times N$ diagonal matrix with elements $b_{\mathrm{n}} = y_{\mathrm{n}}(1-y_n)$, $\boldsymbol{y} = (y_1, y_2, \dots, y_N)$, and $\mathbf{\Phi}$ is the design matrix with elements $\Phi_{ni} = \phi_{ni}(x_n)$. Hence, considering (1), we obtain $\mathbf{\Phi} = \mathbf{K}$, where $\mathbf{K}$ is the symmetric $(N+1)\times(N+1)$ kernel matrix with elements $K(x_n, x_m)$.

Using the Laplace approximation to evaluate the marginal likelihood and defining $\gamma_i = 1 - \alpha_i \sum ii$, the new value for α_i is obtained by:

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^{*})^2}, \tag{15}$$

and by defining:

$$\hat{\mathbf{t}} = \mathbf{\Phi}\mathbf{w}^{*} + \mathbf{B}^{-1}(\mathbf{t}-\mathbf{y}), \tag{16}$$

the approximate log marginal likelihood can be written in the form:

$$\ln\, p(\mathbf{t}\,|\,\boldsymbol{\alpha},\beta) = -\frac{1}{2}\left\{N\ln(2\pi) + \ln|\mathbf{C}| + (\hat{\mathbf{t}})^{\mathrm{T}}\mathbf{C}^{-1}\hat{\mathbf{t}}\right\}, \tag{17}$$

where

$$\mathbf{C} = \mathbf{B} + \mathbf{\Phi}\mathbf{A}\mathbf{\Phi}^{\mathrm{T}} \tag{18}$$

IV. COMBINED MRMR FILTER AND SPARSE BAYESIAN RVM

In this paper, our goal is to present an algorithm to find a compact set of features for gene expression data sets. The mRMR technique discussed earlier finds candidates for the best subset incrementally [5], but as in forward selection algorithms, finding the best subset of features is not guaranteed. To remove the most irrelevant features amongst the chosen set, a forward-backward feature selection scheme is used. The proposed method, depicted in Fig. 1, tries to find the best subset of features in two steps. In the first step, we extract a candidate set using the mRMR from the training data (the forward phase). Next step includes backward wrapper method using the RVM classifier which gradually removes the most irrelevant feature from the candidate set, and finally gives the order of the most compact subsets (the backward phase). At the end, compactness of these feature subsets will be evaluated using test data. In the following, we present our algorithm, Filter-RVM, step by step with more details:

A. Extracting a candidate for best feature set

The processes to find the candidate set can be illustrated in three steps:

1. Use mRMR algorithm to find the best sequence features from the training data progressively.
2. Compute cross-validation classification error on first n sequential feature sets {S1 S2 . . . Sn} (find error set $e = \{e_1, e_2, \dots, e_n\}$).
3. Compare the error of these sequential subsets of features in order to find the k with minimum error, $e_k = \min(e)$, among all feature sets.

Now, we can select the subset S_k as our candidate set for the next stage analysis.

B. Searching for most compact subset

After dimension reduction of the data in the previous stage, a backward wrapper approach to find the most compact subset is used. The backward wrapper tries to find and remove the most redundant features at a time from the feature set. If we consider S_k as our current feature set, the wrapper eliminates one feature. After evaluation of classification error for each S_{k-1}, i.e., finding e_{k-1}, the feature which results in the minimum classification error on validation data is excluded from the remaining subsets. This process goes on until the best subsets are selected.

V. DATA SETS AND EXPERIMENTAL RESULTS

A. Data sets and preprocessing step

To evaluate the presented algorithm, we have used two different gene expression data sets. The data sets and their characteristics are described below:

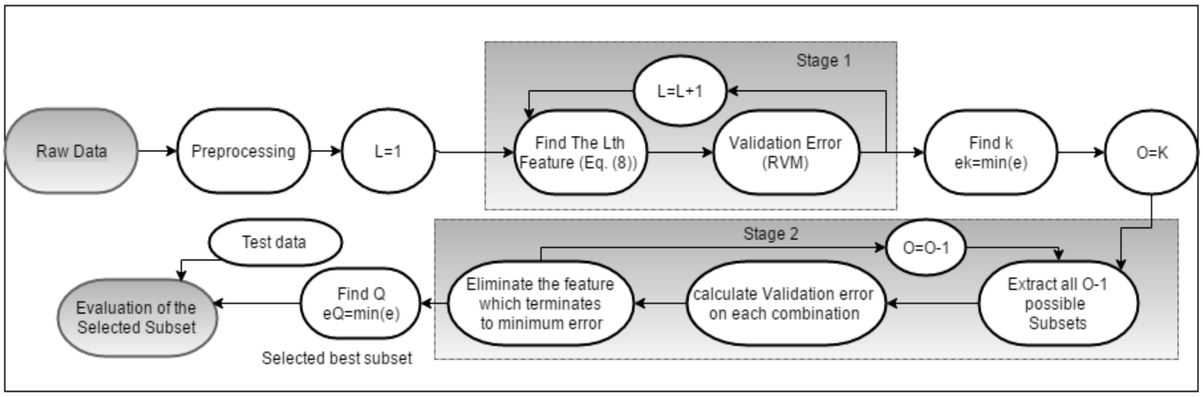
Stage 1
L=L+1
Raw Data
Preprocessing
L=1
Find The Lth Feature (Eq. (8))
Validation Error (RVM)
Find k ek=min(e)
O=K
Stage 2
Test data
O=O-1
Evaluation of the Selected Subset
Find Q eQ=min(e)
Eliminate the feature which terminates to minimum error
calculate Validation error on each combination
Extract all O-1 possible Subsets
Selected best subset

Fig. 1. Diagram of the proposed feature selection method

- Kent-Ridge Colon dataset [14] contains 62 colon-cancer samples collected from the target patients. These samples contain two thousand genes with the most reliable measurement in expression levels. Among them, 40 tumor biopsies are from tumors ("negative" labels) while the other 22 normal biopsies are form the healthy parts of the colons of the same patients ("positive" labels).
- The Leukemia dataset [15] is taken from Leukemia patients. These data set includes 7130 genes of 72 patients which is divided into two classes: ALL (Acute Lymhoblastic Leukemia) and AML (Acute Myeloid Leukemia). 47 patients are distinguished as ALL and remaining 25 patients are distinguished as AML.

The original values of gene distributions in these two data sets are continuous. In the mRMR process, it is required to calculate the joint probability density between gene distributions in data sets, but due to low number of samples, it is impractical to get a good estimate of the probability density function. Therefore, after the data is normalized in a min-max scheme, the gene values are then discretized uniformly to five value points.

B. Results and Comparisons

First, we conducted our experiment using mRMR-SVM algorithm, as our baseline method. The mRMR feature selection method was applied to the preprocessed data. Afterwards, to obtain the best number of features, the subsets derived from mRMR algorithm were evaluated by Leave-One-Out Cross Validation (LOOCV) using SVM classifier. Next, the same procedure was taken by replacing the SVM with the RVM classifier. The classification error rate obtained by these two methods are depicted in Fig. 2 (for Kent-Ridge Colon dataset) and Fig. 3 (for Leukemia dataset). As expected, the mRMR-RVM showed less error rate; as it is a more sparse model that can better capture the structure of the data by choosing proper Relevance Vectors. Adding the RVM to the model makes the system more robust to overfitting as the number of Relevance Vectors is fewer than the number of Support Vectors, Table I.

Further, the proposed Filter-RVM method is evaluated. As in the previous section, the optimum number of features were obtained by LOOCV scheme on the output feature sets of mRMR algorithm. Afterwards, the backward feature selection method is applied to reduce the dimension of the data. In each phase, the most irrelevant feature to the target classes is determined by the RVM classifier. The feature is removed from the subset for further progress. This procedure goes on until only one feature is remained. Moreover, the backward feature selection filtering is also applied on mRMR-SVM algorithm (Filter-SVM). The classification error rates illustrated in Fig.2 and Fig. 3 verify the effectiveness of the proposed filtering method on both mRMR-RVM and mRMR-SVM algorithms. The classification accuracy of 94% for Colon dataset and 100% for Leukemia dataset is obtained by the proposed algorithm using only 7 and 3 gene features, respectively.

To compare our presented algorithm with previous works, we note that El Akadi *et al*. [13], using a two-stage mRMR-GA algorithms, obtained 92% accuracy with 40 features for Colon

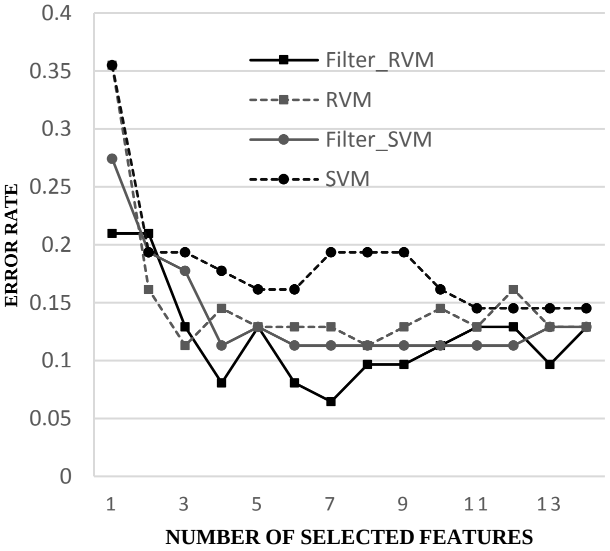

Fig. 2. Classification error rate on Kent-Ridge Colon dataset.

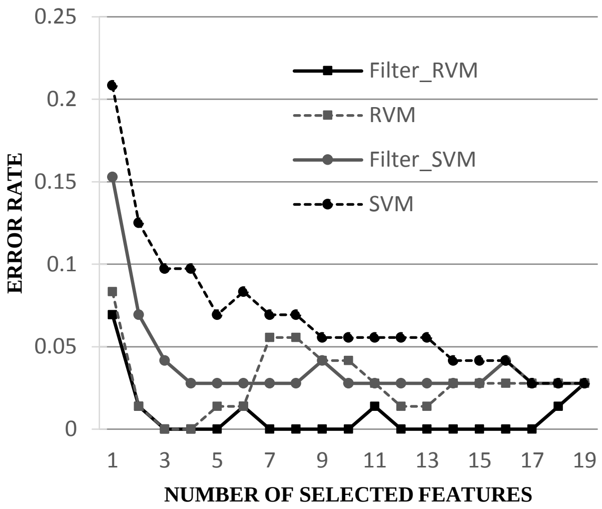

Fig. 3. Classification error rate on Leukemia dataset.

TABLE I.
Comparison of sparsity between mRMR-SVM and mRMR-RVM for best two feature sets. #RV is the number of relevance vectors and #SV is the number of support vectors.

Best Subset	**Colon Dataset**		**Leukemia Dataset**	
	#RV	#SV	#RV	#SV
best subset 1 (RVM)	2	34	2	36
best subset 2 (RVM)	4	31	2	32
best subset 1 (SVM)	2	34	2	26
best subset 2 (SVM)	5	29	2	26

dataset and 100% with 15 features for Leukemia dataset whereas Lee *et al.* in [16] have obtained 1.39% error rate for Leukemia dataset. Rathore *et al.* [11] has used an ensemble classification with 10-fold cross validation algorithm for Colon data set and obtained 96% accuracy with 50 features.

Although a 100% classification accuracy had been obtained on Leukemia dataset, our proposed algorithm outperformed the previous works in terms of sparsity and feature set compactness. This enhances the interpretability of the model and reduces the computational complexity which plays a prominent role in gene expression data analysis.

VI. Conclusion

In this paper, we proposed a method for feature subset selection of gene expression microarrays which uses a cascade of two feature selection steps. After a candidate best feature set is selected by mRMR algorithm, our method tries to reduce the inter-subset distance of the features, expecting to find the best compact gene array which has the most effect on the target classes. According to the previous section, our proposed algorithm showed the accuracy of 94% on Colon and 100% on Leukemia datasets with an extremely sparse model, that is only 2 Relevance Vectors with 7 features and 2 Relevance Vectors with 3 features are selected, respectively. In addition, our model presented a probabilistic output, unlike the previous works, in the classification task which is a supreme lead in diagnosis of tragic diseases, e.g. cancers.

References

[1] T. R. Golub, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, vol. 286, no. 5439, pp. 531–537, 1999.

[2] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003.

[3] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, no. 1-2, pp. 273–324, 1997.

[4] I. Guyon and A. Elisseeff, “An Introduction to Feature Extraction,” Feature Extraction Studies in Fuzziness and Soft Computing, pp. 1–25.

[5] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

[6] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, “Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method,” Bioinformatics, vol. 17, no. 12, pp. 1131–1142, Jan. 2001.

[7] I.Guyon, J.Weston, V.Barnhill, V.Vapnik, “Gene selection for cancer classification using support vector machines,” J Mach Learn Res 46:389–422, 2002.

[8] M. Tipping, “Sparse Bayesian learning and the relevance vector machine,” J. Mach. Learn. Res., vol. 1, pp. 211–244, Jun. 2001.

[9] Bishop CM, Pattern Recognition and Machine Learning, Springer, USA, 2006.

[10] V. Vapnik, S. Golowich, and A. Smola, “Support vector method for function approximation, regression estimation, and signal processing,” in Proc. Annu. Conf. Neural Inform. Process. Systems, vol. 9.Dec1997, pp. 281–287.

[11] S. Rathore, M. Hussain, and A. Khan, “GECC: Gene Expression Based Ensemble Classification of Colon Samples,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no. 6, pp. 1131–1145, Jan. 2014.

[12] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” Bioinformatics, vol. 16, no. 10, pp. 906–914, Jan. 2000.

[13] A. E. Akadi, A. Amine, A. E. Ouardighi, and D. Aboutajdine, “A two-stage gene selection scheme utilizing MRMR filter and GA wrapper,” Knowledge and Information Systems, vol. 26, no. 3, pp. 487–500, Oct. 2010.

[14] Kent Ridge Colon Cancer Dataset. Available at http:// mldata.org/ repository/data/viewslug/colon-cancer-kent-ridge/.

[15] Golube Leukemia cancer Dataset. Available at http://portals.broadinstitute.org/cgibin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63

[16] Y. Lee and C.-K. Lee, “Classification of multiple cancer types by multicategory support vector machines using gene expression data,” Bioinformatics, vol. 19, no. 9, pp. 1132–1139, Dec. 2003.