Method for Classification of Cancers with Partial Least Squares Regression as Feature Selector with Kernel SVM

Method for Classification of Cancers with Partial Least Squares Regression as Feature Selector with Kernel SVM

Nimrita Koul
School of Computer Science and Engineering
REVA University
Bangalore, India
nimrita.koul@reva.edu.in

Sunilkumar S Manvi
School of Computer Science and Engineering
REVA University
Bangalore, India
ssmanvi@reva.edu.in

Bryan Gardiner
School of Computing, Engineering and Intelligent Systems
Ulster University
Northern Ireland, United Kingdom
b.gardiner@ulster.ac.uk

***Abstract*— Classification of cancers according to the exact place of origin in the body is an important research problem that is being addressed both clinically and computationally. Application of data science and machine learning to a vast volume of imaging and genomics data regarding cancers has enabled computational researchers to accurately classify the tumor samples according to their place of origin. In current work, we developed a method to classify tumors using partial least squares regression as feature selector and support vector machine classifiers. We have evaluated our approach on three cancer gene expression datasets and found classification accuracies of 100% in some circumstances. The comparison is conducted with standard classification methods like Decision Trees and simple Support Vector Machines with respect to standard performance parameters and the time taken for classification. The comparison in terms of training and testing accuracies and the time taken for classification results show that our method performs consistently better than conventional methods.**

Keywords— Classification, Machine Learning, Support Vector Machines, Regression Analysis

I. INTRODUCTION

FOR many years, clinical research has been focusing on the study of cancers and drug discovery towards their treatment. Accurate differentiation among various types of cancers is advantageous in deciding the most appropriate course of treatment for the patient with minimum generalized toxicity. Tumors have been found to be molecularly varied in terms of types and subtypes, therefore, would benefit by a varied course of treatment and drugs. The course of treatment for a cancer hugely depends on its accurate diagnosis and classification. This is also critically important when pursuing drug discovery research. DNA microarray technology [1] has provided genome wide expression levels of genes corresponding to normal and cancerous tissues. This data is used by computational researchers to analyze various patterns, abnormalities, influences and regulations among the genes [1,2,3,4]. Researchers have developed various methods for dimensionality reduction [2,3], analysis and classification of this data. These methods help with not only classification of various tumor types but also subtype classification.

The methods for classification include machine learning, information theoretic methods, statistical and probabilistic methods. Machine learning approaches like artificial neural networks (ANN), support vector machines (SVM), random forest (RF), etc., are used widely for classification tasks for cancer gene expression profiles [2, 13, 14, 20].

The method proposed in this paper uses partial least squares regression for feature selection and provides a simple and straight forward algorithm for selecting the most relevant features. This approach works well with both two class as well as multiclass cancer classification tasks. The classification of cancer is sometimes a two-class classification problem, and often a multi-class problem. Some classification methods work well for binary classification while some methods give better results for multi-class problems. Gene expression data [1] is a very high dimensional data, the sample size is much less than the dimensionality. Most of the genes present in datasets are irrelevant to classification [21-25] of the cancers, hence, they need to be identified and dropped from the set, reducing the overall dataset dimensionality, thus, reducing the computational overload during analysis. This is done by the process of feature selection or gene selection. Since cancer classification is a biological problem that is being solved computationally, the results obtained must be relevant biologically.

The rest of the paper is organized as follows: Section II discusses the background and the problem statement, Section III talks about the proposed work, and Section IV discusses the results & a discussion thereon. In Section V we conclude the paper and provide the scope for future work.

II. BACKGROUND AND PROBLEM STATEMENT

A. Background

A human cell has its functioning and life cycle encoded into its nucleus in the form of DNA and RNA which are the repeating sequences of four bases. Central dogma of cell biology tells that the process of transcription generates RNA from DNA and the process of translation generates amino acid sequences, hence the proteins, from RNA. RNA is of two main types - the messenger RNA (mRNA) and the transfer RNA (tRNA). Gene expression is the process of transcription of DNA into mRNA. Therefore, the gene expression value is an indicator of the number of copies of RNA manufactured with the cell and also the protein

synthesized. Various conditions of environment, disease, stress, lifestyle, etc. have been shown to change the expression levels, hence behavior of a cell. Thus, gene expression patterns can be used to distinguish between a normal and a diseased state of a cell. A DNA array containing m number of genes on a chip, produces m number of expression ratios. Numerator of this ratio is the expression value of a gene 'k' in diseased state or abnormal circumstances, and the denominator represents the expression value of the same gene under the controlled settings or normal conditions.

B. Related Works

Golub et al. [1], did early work in the field of computational cancer data analysis for classification of cancers using information theoretic measures. They classified the gene expression profiles into two subtypes of leukemia, Acute Lymphocytic Leukemia (ALL) and Acute Myelocytic Leukemia (AML).

Guyon et al. [2], used a wrapper approach called recursive feature elimination with a SVM classifier as a validator of the selected feature set. Although filter approaches work effectively for the task of cancer classification, the wrapper methods like recursive feature elimination select the subset most suitable to the machine learning task to be attempted. Lê Cao et al. [3] also employed a wrapper method which is a variation of Partial least squares regression, known as "sparse partial least squares discriminant analysis" for feature selection from public microarray datasets. This approach was found to be computationally effective and gave more interpretable results. In [4], M. Masud Rana et al. have used a wrapper method called "minimum redundancy maximum relevance" for selecting feature subsets containing features with least mutual redundancy and highest relevance to the output class label. Nancy et al. [5], applied an evolutionary algorithm called Ant colony optimization combined with adaptive network fuzzy inference system for automatic and fast feature selection. This is a complex method and incurs a heavy computational overload. Shukla et al. [6], have used a distributed approach to feature selection for cancer subtype detection for "Diffuse Large B cell Lymphoma (DLBCL)" dataset and the classification is performed using SVMs, naïve Bayes, k - nearest neighbor, as well as the Decision Tree classifiers. H. Yanhao [7], have studied sparsity of gene expression profiles and proposed a sparse group LASSO technique for feature selection. They employ SVM as classifier. They have tested the approach on gene expression data, MiRNA and DNA methylation data as well. Xiaohong et al. [8], developed a two-phase feature selection for the task of cancer classification. Their algorithm iteratively reduces the size of selected genes, with a result of the approach finding less than 1.5% of the original number of features as selected features. Morais-Rodrigues et al. [9], have used logistic regression on breast cancer gene expression data for classification without applying feature selection. Loey et al [10], have used information gain (IG) for feature selection from breast and colon cancer gene expression datasets. The selected genes are further shortlisted using the grey wolf optimization (GWO). The classification is conducted using SVMs on breast and colon data, which improved the stability of classification accuracy and feature selection. This also is a two phase wrapper feature selection approach. Akhand et al. [11], have employed feature selection by minimum redundancy maximum relevance algorithm and applied various classifiers like a neural network classifier, Naïve Bayes, Decision Tree on benchmarks datasets. Almugren et al. [12], present a survey of recent hybrid methods that have been used for feature selection and classification.

C. Problem Statement

To differentiate profiles into two categories – cancerous or non-cancerous is a two class classification problem. When the objective is to differentiate the profiles into subtypes of a cancer, the problem becomes a multiclass classification problem. This problem is mathematically stated as follows: Let the random variables X_1, X_2, X_m represent the expression ratios of the genes $G_1, G_2 \ldots G_m$ respectively. The domain of X_i is the range of expression values of G_i. The label of each gene expression profile is its class, represented by C. A random variable, C takes one of the K possible values, if there are K subtypes of a cancer.

The expression profile of a sample S is a tuple *'t'*, of size m, containing the expression values of all *'m'* genes of a genome in that sample.

$$t = \{X_1.value, X_2.value, \ldots., X_m.value\} \quad (1)$$

Mathematically, the cancer classification problem can be stated as follows:
Consider a training set

$$T = \{(t_1, c_1), (t_2, c_2), (t_3, c_3), \ldots, (t_n, c_n)\} \quad (2)$$

Where n is the sample size, t_i is m-dimensional gene expression profile for *'m'* genes,

$$t_i = (t_i.X_1, t_i.X_2, \ldots, t_i.X_m\} \quad (3)$$

'c_i' is the label of ith tuple, $c_i \in dom(C)$.

As sample size, m, is much smaller than number of genes, therefore $m \gg n$. R is the test set = $\{r_1, r_2, .., r_k\}$ where each r_i is the full expression profile of length *'m'* of gene *'i'*. The form of each *'r_i'* is a tuple of length m, such that

$$r_i = \{r_i.X_1, r_i.X_2, \ldots, r_i.X_m\} \quad (4)$$

Where X_i is an expression value of gene *'i'*. The problem is to find a classification function that maximizes the accuracy of classification on test set R.

D. Our Contributions

The original contributions of this paper to the field of feature selection are as follows –

1. We have modified Partial least squares regression algorithm to work on very high dimensional cancer gene expression datasets. We obtained 5, 10, 20 and 50 top ranked genes from three benchmarks cancer gene expression data sets.
2. We have done an exhaustive comparison on testing accuracy obtained from the SVM classifier by training it on very few genes from each gene expression dataset to establish the effectiveness of feature selection technique used.

We have compared the performance of 4 different kernels of SVM when trained very few features and established that linear kernel performs the best with just 20 features in 2 of the 3 datasets used

III. PROPOSED WORK

In this section, we present the proposed approach for feature selection and classification of cancer gene expression datasets. Details of the proposed algorithm, the datasets used, and experimental details are all outlined.

A. Datasets

We used 3 publicly available cancer gene expression datasets to test the classification performance of our algorithm, with each dataset representing a different type of cancer, namely, leukemia[1, 23], small round blue cell tumors (SRBCT) [24] and colon cancer [25]. Details of the datasets are as follows.

TABLE 1
DETAILS OF DATA SETS USED

Dataset	No. of Genes	No. of Samples	No. of Classes
Leukemia	3571	72	2
SRBCT	2308	83	4
Colon	2000	62	2

B. Proposed Method

1) Feature Selection

Feature selection on high throughput gene expression data selects relevant genes that influence the diagnosis of samples or characterize the disease. We have proposed the use of partial least squares regression (PLS) [15] for relevant feature selection from gene expression data. PLS uses latent variables as indicators of the relationships among response variables and predictors. It is resistant to multi-collinearity, noise, high dimensionality and the cases when the number of dimensions is much higher than the number of samples as is the typical case with gene expression data. In this regression, the predictors matrix *X* with dimensions *n*k* and target matrix Y with dimensions *n*m* can be modelled as follows –

$$X = TP^T + E \quad (5)$$
$$Y = UQ^T + F \quad (6)$$
$$u_a = b_a t_a + h \quad (7)$$
$$a = 1, 2, 3, \ldots . A$$

Where *A* is the number of "latent variables", $T = (t_1, t_2 \ldots tA)$ and $U = (u_1, u2 \ldots u_A)$ are latent variable scores of *X* and *Y*. T is a projection of X and U is a projection of Y with same dimensions as A and Y respectively. *P* and *Q* are orthogonal loading matrices calculated by non-linear iterative partial least squares method. E and F are errors with a random variables with normal distribution. Eq. 5 and Eq. 6 are the outer relation between *T* and *U*, b_a is the regression coefficient of u_a. Eq. 7 is the inner relation between *U* and *T, E, F*, and h is the error in *X, Y* and u_a.

Cross validation procedure has been employed to reduce the prediction error. The proportion of variance explained by each latent variable determines the total number of latent variables used in PLS regression and it is an important parameter that determines accuracy of prediction.

The contribution of each gene to the class label is determined by decomposing the sum of squares of the gene expression values, where the total sum of squares of latent variables has two components: "sum of squares of regression"; and "sum of squares of error". Sum of squares (SS) is the square of difference between actual value of predicted variable, *y*, and its mean.

$$SS = SSR + SSE \quad (8)$$

Where SS or SS(Y) is the "total sum of squares" of latent variables, SSR is sum of squares of regression, SSE is sum of squares of error.

The importance of each gene in the input gene expression matrix is calculated by PLS as given in (9).

$$GIj = \sqrt{\frac{k \sum_a w_{ja}^2 \; b_a^2 \quad t_a^T t_a}{\sum_a b_a^2 t_a^T t_a}} \quad (9)$$

Where GI_j is the gene importance of *jth* gene in the gene expression dataset, w_{ja} is the weight of the *jth* gene to the a^{th} latent variable. The weight is obtained by applying the nonlinear iterative partial least squares algorithm given as follows –

1. Given the gene expression matrix X, select a column vector xi and copy it to vector u;
2. Project matrix X onto u to find the loading v;
3. Normalize v to length 1;
4. Copy the old scores in u to another vector u_{old} and project X to v to find updated score vector u;
5. For convergence check, calculate difference vector d between previous scores and current ones. If magnitude of d is larger than the threshold, repeat from step 2;
6. Remove estimated product of the scores and loadings from X and store the remaining values in matrix E;
7. Repeat the procedure with E as new X to find other principal components.

Therefore, the gene matrix X was broken into a score matrix T, a loadings matrix P and an error matrix E. The class matrix Y is broken into components **U** and **Q** and the error term **F**. PLS minimizes the norm of F while retaining correlation of X and Y by means of inner relation U. Optimum number of principal components, a, is determined by cross validation. The best model is one that has the least value of predictive error sum of squares and the minimum size of selected geneset.

2) Classification

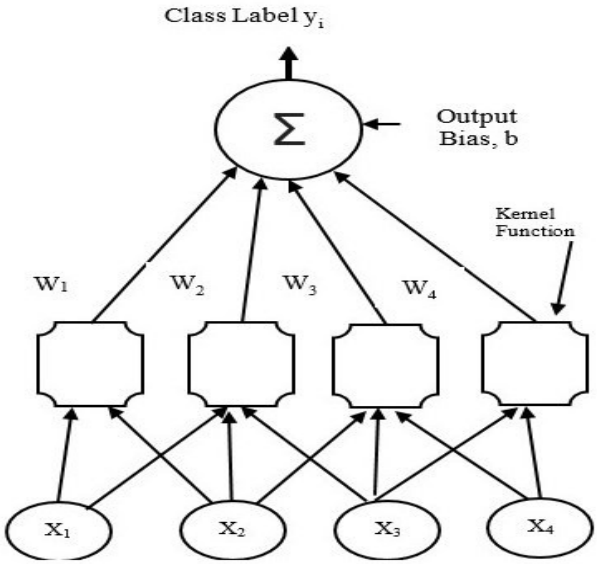
For classification of gene expression profiles, an SVM is used. An SVM is a kernel method that uses dot products of original data in a higher dimensional feature space, and often used with linear, sigmoid, quadratic, polynomial, radial basis function (RBF) or exponential RBF kernels. The kernel function determines the feature space used by the classifier. SVMs use the principle of large margin of separation for classification into separate classes. Thus, the SVM is an extended perceptron with margins, feature expansion and kernel trick. The margin is the shortest perpendicular distance

of an input vector from the decision or class boundary. A positive margin means a correctly classified input vector while a negative margin means a wrongly classified vector.

Feature expansion is achieved by adding new features, which are products of higher powers of original input features. The classifier function is now a nonlinear function for original input features but linear for the expanded feature space. The decision boundary is now a nonlinear surface. A polynomial feature expansion of degree k on an input feature vector of size '*n'* gives $O(n^k)$ new features. RBFs generate a much larger number of new features. In order to optimize in the presence of a large number of features, the kernel trick is used as follows:

1. Compute one Lagrange multiplier for each input gene expression vector;
2. Calculate optimal weights in terms of linear combinations of rows;

This algorithm balances two goals of minimizing the error in prediction and keeping the model as simple as possible using regularization, known as structural risk minimization. The classifier applies a kernel trick for high dimensional gene expression data which is linearly inseparable, which means that there is no need to compute exact data transformations, i.e. only inner products in higher dimensions is computed and getting inner products is easier than getting exact data points in higher dimensions.



3) Cross Validation

In order to prevent over fitting while training the classifier, each of the input gene expression datasets was divided into a training set (T) and testing set (R), with a 70 percent and 30 percent ratio respectively. K-fold cross validation was performed on the data, with k = 5 and k = 10.

4) Performance Parameters

In the case of classification algorithms, a true positive (TP) is considered when the algorithm correctly predicts a label as positive, a true negative (TN) occurs when the model correctly predicts the negative class of the sample, i.e. both the actual and the predicted class labels are negative, a false positive (FP) is an output when the algorithm predicts a label as positive when it is actually negative, and a false negative (FN) is the case when the algorithm predicts the label to be negative when it actually is positive. A confusion matrix is an n*n table that shows correlation between the actual labels and the predicted labels for n classes; 'n' is 2 for binary classification. The algorithm reaches training convergence as the training loss and validation loss stabilize. The proposed classification algorithm was evaluated on the basis of the following parameters –

1. Accuracy (ACC) – Classification accuracy is the ratio of number of correctly classified samples in the input data set to total number of samples in it. A true positive (TP) is a classifier output where actual class is positive and prediction is also positive. A true negative (TN) is a classifier output where the actual class is negative and classifier also predicts a negative class. A false positive (FP) is a when actual class is negative but the classifier predicts positive, similarly a false negative (FN) is when actual class is positive but the classifier predicts negative.

$$ACC = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad (10)$$

2. Precision – Precision of a classification model measures the frequency of correct prediction of the positive class.

$$PR = \frac{(\mathrm{TP})}{(\mathrm{TP}+\mathrm{FP})} \quad (11)$$

3. Recall – Recall of a classifier measures the ratio of correctly predicted labels in the positive class.

$$RC = \frac{(\mathrm{TP})}{(\mathrm{TP} + \mathrm{FN})} \quad (12)$$

4. F1 Score – This metric is the harmonic mean of precision and recall, given by

$$F1 = 2 * \frac{(\mathrm{PR})}{(\mathrm{PR}+\mathrm{RC})} \quad (13)$$

Figure 1. Schematic of SVM Classifier

Figure 1 shows the structure of a Support Vector Machine (SVM) classifier. X_i *are* the input vectors that are transformed into higher dimensions by the Kernel functions into a space where linear separation can be carried out. The transformed vectors are multiplied by weights, the output of an SVM is the weighted sum of all inputs and a bias. As presented in Figure 2, feature selection through partial least squares regression is applied on the input gene expression data with original number of genes available in the data sets. Four kernels were applied to the SVM for classification with the entire original gene set, then with 5, 10 and 20 top selected genes. The resulting performance of the classifier throughout all experiments conducted is presented in Section IV.

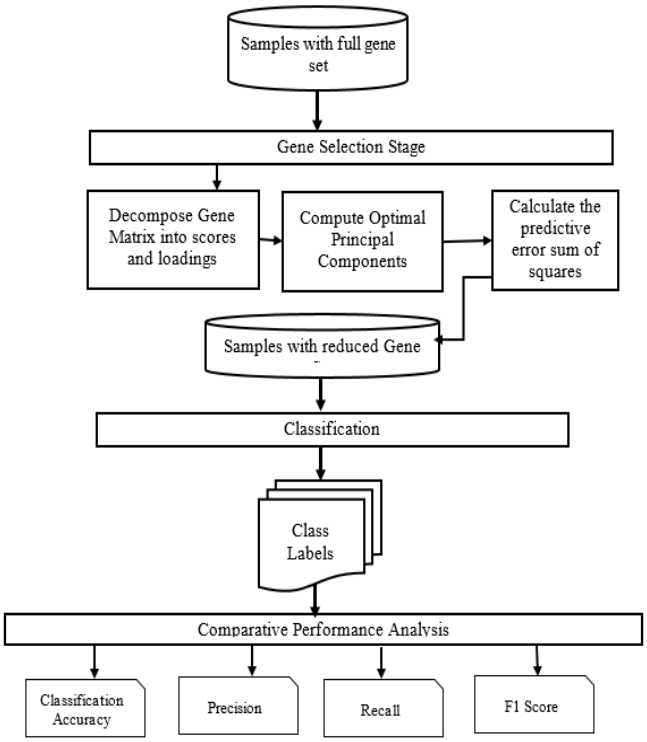


Figure 2. Schematic Representation of the proposed approach

IV. Results and Discussion

Three input gene expression datasets, viz. Leukemia SRBCT and colon were fed as input to four classification algorithms, viz. SVM with linear kernel, SVM with RBF kernel, SVM

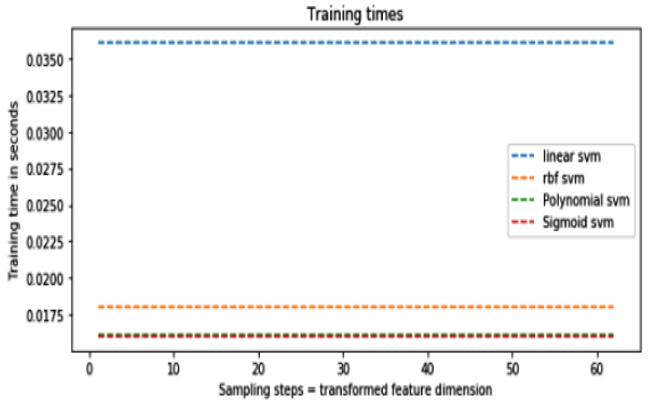


Figure 3. Comparison of Training Times using 4 SVM kernels on Colon dataset

with polynomial kernel and SVM with Sigmoid kernel to note the classification accuracy and other performance parameters defined in Section II. Thereafter, the feature selection was applied on the datasets to obtain top ranked 5, 10, 20 and 50 genes. Next, all four kernel SVMs were applied on reduced feature input sets and the performance parameters were recorded. Feature selection module was implemented in R and the classification module was implemented in Python. On application of the proposed PLS regression based feature selection approach to the three datasets we obtained 4 gene subsets for each of the input data sets. For the Leukemia dataset, the gene IDs in each subset selected by the PLS were subset with 5 best genes, subset with 10 best genes, subset with 20 best genes and the subset with 50 best genes. Similarly for the other two datasets, we obtained four gene subsets for evaluation with the SVM classifier. Figures 3, 4 and 5 show the comparison of training times required for each SVM kernels on the colon, SRBCT and leukemia datasets respectively. In all datasets, we observe that the linear kernel takes the maximum time while the sigmoid kernel takes the minimum amount of time for completing the training.

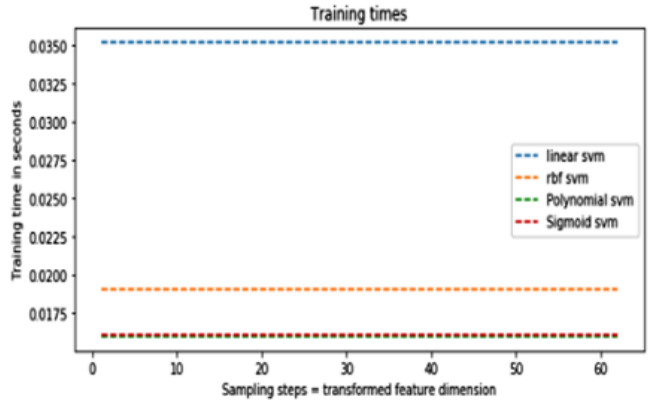


Figure 4 . Comparison of Training Times using 4 SVM kernels on SRBCT dataset

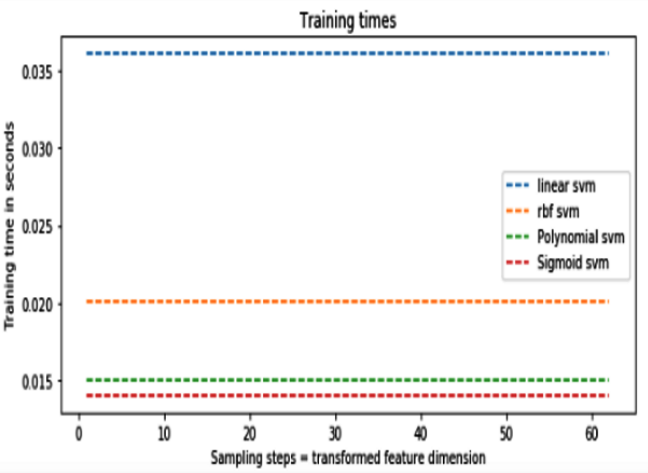


Figure 5. Comparison of Training Times using 4 SVM kernels on Leukemia dataset

In Table 2 we have presented a comparison of classification accuracy on test data set for three input gene expression datasets. We trained the classifiers using the original set of features in each of the three data sets, then using top 5 features, top 10 features, top 20 features and top 50 features. In all cases of input feature set, the Linear SVM, RBF SVM, Polynomial SVM with degree 4 and Sigmoid SVM classifiers were employed. It is observed that which we get almost 100% accuracy with full feature set, top 20 features and top 50 features also give almost 100% accuracy with Linear SVM for SRBCT and Colon datasets.

TABLE 2
COMPARISON OF CLASSIFICATION ACCURACY ON 3 DATASETS

Leukemia Dataset				
Number of Features Selected	**Linear SVM**	**RBF SVM**	**Polynomial SVM degree 4**	**Sigmoid SVM**
Full Feature Set	100%	98%	100%	77%
Top 5	75%	78%	76%	58%
Top 10	90%	90%	85%	85%
Top 20	91%	100%	77%	90%
Top 50	100%	89%	84%	95%
SRBCT Dataset				
Number of Features Selected	**Linear SVM**	**RBF SVM**	**Polynomial SVM degree 4**	**Sigmoid SVM**
Full Feature Set	100%	100%	100%	77%
Top 5	78%	64%	76%	76%
Top 10	100%	92%	85%	79%
Top 20	100%	100%	77%	53%
Top 50	100%	100%	90%	70%
Colon Dataset				
Number of Features Selected	**Linear SVM**	**RBF SVM**	**Polynomial SVM degree 4**	**Sigmoid SVM**
Full Feature Set	92%	85%	85%	70%
Top 5	73%	62%	77%	62%
Top 10	80%	79%	85%	62%
Top 20	100%	85%	80%	77%
Top 50	100%	95%	100%	65%

From the results, we find that the features selected using PLS regression give a good classification performance for these datasets. We have observed classification testing accuracy of 100% on the colon dataset with Linear Kernel with just 20 selected features, a testing accuracy of 100% on both the SRBCT and leukemia dataset with just 10 selected features and linear kernel. We can say that the linear kernel SVM works best for the classification task used on gene expression datasets with just 10 features. The proposed approach works effectively in the high dimensional gene expression data classification with nonlinear relationship among the genes, with calculations using support vectors reducing the memory consumption. The proposed method is flexible and extensible as new and custom kernel functions can be easily applied. Since gene expression data typically contains a much greater number of features than the samples, the regularization is needed to avoid overfitting.

TABLE 3

COMPARISON OF AVG CLASSIFICATION ACC FOR FEATURE SELECTION AND CLASSIFICATION COMBINATIONS SELECTING 10 BEST FEATURES

Feature Selection Method	Classification Algorithm	(Training Accuracy, Testing Accuracy) for best 10 features
PLS	Linear SVC	(88%, 100%)
PLS	DT	(100%,100%)
PLS	RF	(46%, 23%)
Mutual Information	Linear SVC	(60%, 61%)
Mutual Information	DT	(100%,100%)
Mutual Information	RF	(82%,84%)
RFE	Linear SVC	(47%, 46%)
RFE	DT	(100%, 100%)
RFE	RF	(84%, 92%)

In Table 3, we have shown the average classification accuracy obtained for 9 combinations of feature selection and classification algorithm pairs for the three datasets – leukemia, SRBCT and colon cancer. For this comparison, the classification algorithms were trained on the ten best features selected by the feature selection algorithm. This comparison presents both the training and testing accuracy.

V. Conclusion

Computational cancer classification is an important and complex research problem given the nature of input data. In this paper, we have used microarray gene expression data as input for classification into tumorous and non-tumorous classes in two datasets and multiclass classification in SRBCT dataset as it involves four sub classes. The input data is high dimensional and the number of samples is much less. We applied partial least squares regression to reduce the number of genes required for classification, where a SVM with various kernels was used as the classification algorithm. We demonstrated the higher accuracy levels achieved with the proposed method when compared to other existing methods. Best performance is achieved when utilizing the top 10 selected features in SRBCT and Leukemia datasets and at 20 best features for colon dataset. Linear kernel is seen as the best method based on classification accuracy in all datasets. As future work, the method will be tested on the combination of multi-modal data i.e. gene expression data along with RNA sequence data, further improving clinical relevance of the obtained results.

Acknowledgment

Authors thank the Department of Science and Technology (DST), Government of India, for financially supporting this work under the DST-ICPS grant scheme for the year 2018.

References

[1] G T.R. Golub, D.K. Slonim, P. Tamayo, M. Gaasenbeek C. Huard, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, pp. 531–537, 1999.

[2] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines", *IEEE Trans. Knowl. Data Eng*., vol. 25, no. 1, pp 1-14, 2002.

[3] K. Lê Cao, S. Boitard, S., PBesse, "Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems", *BMC Bioinformatics*, vol. 12, no. 253, 2011.

[4] M. Rana., K. Ahmed, "Feature Selection and Biomedical Signal Classification Using Minimum Redundancy Maximum Relevance and Artificial Neural Network", *Proceedings of International Joint Conference on Computational Intelligence. Algorithms for Intelligent Systems*, Springer, Singapore, 2020.

[5] S.G. Nancy, K. Saranya, S. Rajasekar, "Neuro-Fuzzy Ant Bee Colony Based Feature Selection for Cancer Classification", *Springer Innovations in Communication and Computing. Springer, Cham* 2020.

[6] A.K. Shukla, D. Tripathi, "Detecting biomarkers from microarray data using distributed correlation based gene selection", *Genes & Genomics*, 2020.

[7] K. Kourou, G. Rigas, C. Papaloukas, M. Mitsis, D.I. Fotiadis, "Cancer classification from time series microarray data through regulatory Dynamic Bayesian Networks", *Computers in Biology and Medicine*, 2020.

[8] H. Yanhao, X. Lihui, K Chuanze, W. Minghui , M Qin, Y Bin, "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso*", Journal of Theoretical Biology*, vol. 486, 2020.

[9] H. Xiaohong , L. Dengao, L. Ping, Li Wang, "Feature selection by recursive binary gravitational search algorithm optimization for cancer classification", *Soft Computing ,* vol. 24, no. 6, pp. 4407–4425, 2020.

[10] F. Morais-Rodrigues, R. Silverio-Machado, R.B. Kato, D.L.N. Rodrigues, J. Valdez- Baez, V. Fonseca, E.J. San, L.G.R. Gomes, R.G. dos Santos, M. Vinicius Canário Viana, J. da Cruz Ferraz Dutra, M. Teixeira Dornelles Parise, D. Parise, F.F. Campos, S.J. de Souza, J.M. Ortega, D. Barh, P. Ghosh, V.A.C. Azevedo, M.A. dos Santos, "Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression", *Gene Gene,* vol. 5, 2019.

[11] M. Loey, M. Wajeeh Jasim, M. Hazem EL-Bakry, M. Hamed N. Taha, N. Eldeen M. Khalifa, "Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques", *Symmetry,* vol. 12, no. 408, 2020.

[12] M. A. H. Akhand, Md. Asaduzzaman Miah, Mir Hussain Kabir, M. M. Hafizur Rahman, "Cancer Classification from DNA Microarray Data using mRMR and Artificial Neural Network", *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 7, 2019.

[13] N. Almugren, H. Alshamlana, "Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification", *IEEE Access*, vol. 7, pp. 75833-44, 2019.

[14] Y.A. Zakariyal, M. Hisyam Lee, "A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification", *Advances in Data Analysis and Classification*, vol. 13, pp: 753–771, 2019.

[15] P., Geladi, B. Kowlaski, "Partial least square regression: A tutorial. Analytica Chemica Acta", vol. 35, pp. 1–17, 1986.

[16] M. Sarah Ayyad, A. I. Saleh, M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors Technique", *BioSystems*, vol. 176, pp: 41–51, 2019.

[17] A. Russul , H. Jingyu, H. Azzawi, X. Yong, "A novel gene selection algorithm for cancer classification using microarray datasets", *BMC Medical Genomics*, vol. 12, no. 10, 2019.

[18] S.A. Medjahed, T. A. Saadi, A. Benyettou, A. Ouali, "Kernel-based learning and feature selection analysis for cancer diagnosis", Appl *Soft Comput*, vol. 51, pp 39–48, 2017.

[19] P. Mundra, J. Rajapakse, "SVM-RFE withmRMR filter for gene Selection", *IEEE Transactions on. Nano. Biosci*, vol. 9, no. 1, pp. 31-37, 2010.

[20] L. Wang, F. Chu, W Xie, "Accurate cancer classification using expressions of very few genes", *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 40–53, 2007.

[21] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M Schwab, C.R Antonescu, C. Peterson et al,. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural Networks", *Nat. Med.*, 7, 673-679., 2001.

[22] U Alon , N Barkai , D A Notterman, et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arra", *Proc Natl Acad Sci USA*, vol. 96 pp. 6745–50, 1999.

[23] Leukemia Data Set https://web.stanford.edu/~hastie / CASI_files/ DATA/ leukemia.html

[24] SRBCT Data Set https://research.nhgri.nih.gov /microarray/Supplement/

[25] Colon Dataset http://genomics-pubs.princeton.edu/oncology/

[26] Sebastian Student and Krzysztof Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data", Biology Direct, vol 7, no. 33, 2012.