Linear regression-based feature selection for microarray data classification

Article *in* International Journal of Data Mining and Bioinformatics · August 2015

DOI: 10.1504/IJDMB.2015.066776

3 authors, including:

Md Abid Hasan
University of California, Riverside
14 PUBLICATIONS **132** CITATIONS
SEE PROFILE

Md Kamrul Hasan
Islamic University of Technology
205 PUBLICATIONS **1,686** CITATIONS
SEE PROFILE

Linear Regression Based Feature Selection for Microarray Data Classification

Abstract: Predicting the class of gene expression profiles helps improve the diagnosis and treatment of diseases. Analyzing huge gene expression data otherwise known as microarray data is complicated due to its high dimensionality. Hence the traditional classifiers do not perform well where the number of features far exceeds the number of samples. A good set of features help classifiers to classify the dataset efficiently. Moreover, a manageable set of features is also desirable for the biologist for further analysis. In this paper, we have proposed a linear regression based feature selection method for selecting discriminative features. Our main focus is to classify the dataset more accurately using less number of features than other traditional feature selection methods. Our method has been compared with several other methods and in most of the cases the classification accuracy is higher using less number of features than the other popular feature selection methods.

Keywords: Regression analysis, feature selection technique, large dataset classification

1. Introduction

Data mining is a multidisciplinary effort to extract important knowledge from biologically embedded data. The maturation of large data sets in different domain baffled the researchers with great challenges [1]. With the recent advancement in gene expression microarray technology, scientists have the opportunity to observe the complex relation between various genes in a genome by simultaneously measuring the expression levels of thousands of genes [2]. This gene expression profiles has turned into an interest of extensive study in biomedical research; especially in classification of patient samples. One of the most common approaches in this area is binary classification, which distinguishes between two types of samples: *case* samples (individuals that carry some illness) and *control* samples (healthy individuals). Other type of binary class datasets might contain two types of malignant samples where both type of samples are suffering from some illness but with different types. Different learning algorithms can effectively classify between the classes by training a classifier.

With the help of microarray technology, thousands of gene expression can be measured simultaneously which saves a lot of time and expenses, but the number of samples in any microarray experiment is far less than the number of features (genes). This asymmetry of data poses a serious challenge for standard learning algorithms as they are not suitable handling imbalanced datasets. The *high-dimensionality* of microarray dataset often leads the classifier to discrepant results [3]. Computational learning theory suggests that the search space is exponentially large in terms of N, the number of features, and the required number of samples for reliable learning about given phenotype is on the scale of $O(2^N)$ [4]. Even the minimum requirement ($M = 10 \times N$) as a statistical "rule of thumb" is impractical for microarray dataset [5]. That is why a small number of discriminative genes often a requisite for better classification. Moreover, a compact set of discriminative genes are desirable by the biologists because of the heavy expenses associated with follow-up biological or clinical validation of the selected genes [2]. The advantages of feature selection are: (1) Removal of irrelevant and redundant data. (2) Improvement of the efficiency of the classifier both in time and space. (3) Features selected are more likely to be responsible for the disease. (4) It helps classifiers with improved generalization ability by avoiding over-fitting [4].

Feature selection techniques broadly fall into two models: *wrapper* model and *filter* model [6]. Wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the acceptability of a selected subset. This process is computationally very expensive. On the other hand, filter method selects the subset of features as a pre-processing step which is

independent of the classifier. Filter method has been a popular choice due to its quality of being computationally simple, fast and the ability to handle extremely large-scale datasets. Furthermore, any classifier can be used after the feature selection process is finished [7].

In this study, we developed an improved feature selection technique based on linear regression analysis. The method aims to obtain discriminative features (genes) by measuring the deviation of the 'case samples' features from the standard regression line. The standard regression line is drawn using the control samples considering these as a standard data that will be used for comparison in a *case-control* dataset. While in case of two malignant types of datasets, one type of data can be considered as a standard or base dataset and thus make the regression line using that type of data. This method selects the discriminative features by acquiring knowledge from the control samples and applies that knowledge to find features from the case samples those vary from the model. We tested the performance of the proposed method by six classifiers: Support Vector Machine (SVM), Random Forest, KStar (K*), Naive Bayes and AdaBoost. We have compared our proposed feature selection method with ReliefF algorithm which evaluates the discriminative power of individual genes without handling gene redundancy [8]. ReliefF searches for the nearest neighbours of samples of different classes and weights gene according to how well they differentiate them. The other algorithm is CFS [9], which exploits heuristic subset search based on some correlation measure. Correlation between the genes and discriminative power of individuals are the measures of the goodness of this algorithm. Other techniques used for feature selection are based on chiSquared value, gain ratio value and Information gain value [10, 11]. In our proposed method, the relationship of a gene with respect to all other genes is considered for building the regression model as well as selecting the features for classification.

On some previous work of linear regression, Huang and Pan [12] point out a close connection of weighted voting scheme [17], the compound covariant method [16], and the shrunken centroids method [25] and tried to classify using the general framework of linear regression In [13] the authors also used linear regression model which can be used to search the motifs that maximize the fit to the regression model [14, 15]. There have been some other applications in practice. Hedenfalk *et al.* [16] studied on classification of hereditary breast cancer with/without BRCA1 and BRCA2 mutations using cDNA microarrays whereas Golub *et al.* [17] considered distinguishing two subtype of leukemia, acute myeloid leukemia (AML) and Acute Lymphoblastic Leukemia (ALL), using Affymetrix high-density oligonucleotide microarrays containing probes for 6817 genes. In microarray dataset the number of features far exceeds the number of samples and it is impossible to fit a linear regression model directly using all genes. For example Golub *et al.* [17] choose about 1100 out of 6817 genes based on their correlation with the class distinction, and Hedenfalk *et al.* [16] used a standard *t*-test with the significant level $\alpha = 0.0001$ to select genes with cross-validation, resulting in very few genes.

In this study we have used the linear regression as a statistical approach for gene selection based on their divergence level from the standard gene expression regression line. The reason behind choosing linear regression over polynomial regression is, the level of expression does not vary much and the expression best fits the linear model. For comparing the divergence of gene expression level, we built the standard regression model by selecting a subtype of training dataset and then compared the expression values of the genes statistically on the other subtype of data and thus the differentially expressed genes are chosen as features (i.e. the discriminating genes).

The proposed method is a feature selection technique which is independent of any classification algorithm's help. It endues some knowledge about a subtype of dataset and using that knowledge the algorithm finds the set of features by applying the acquired knowledge to the other subtype. This method can be used as a pre-processing step of any classification procedure. That is why we have used different classifiers to evaluate the performance with other feature selection methods.

2. Methods

We first provide some background about linear regression analysis. Then we will describe our proposed algorithm based on linear regression analysis.

2.1. Linear regression analysis

An approach to model the relationship between scalar dependent variable y and one or more explanatory variables X. The primary purpose of the regression analysis is the development of an equation that can be used for predicting values on some dependent variables, given explanatory variables for all members of a population. Regression analysis helps to understand how typical values of dependent variable change when one or more variables are fixed. It estimates the conditional expectation of the dependent

variable given the independent variables. We then fit the linear regression model into a learning algorithm. It's the job of that learning algorithm to output a function otherwise known as hypothesis function; that takes input variable *x* and give output variable. The hypothesis function for single variable is represented by

$$h_\theta = \theta_0 + \theta_1 x \quad (1)$$

here θ_i are the parameters. This model is called linear regression model with one variable or univariate linear regression. We need to find the values of the parameters that best fits the data in the training dataset. In other word, we need to choose θ_0 and θ_1 in such a way that h_θ is close to *y* in our training data (*x, y*). So we need to find θ_0 and θ_1 so that the cost function

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \quad (2)$$

is minimized. *m* is the total number of samples in the training set. Now linear regression model with *n* variables is

$$h_\theta = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n \quad (3)$$

and the cost function becomes

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \quad (4)$$

where θ is a set of parameters $\{\theta_0, \theta_1, \cdots, \theta_n\}$.

To minimize the cost function $J(\theta)$ an algorithm can be implemented which helps find the values of θ so that the cost is minimum. Applying the gradient descent algorithm for the cost function in Equation 4, the parameter values can be obtained and thus the cost can be minimized. By defining the cost function into partial derivative term, the cost function can be written as

$$\frac{\delta}{\delta\theta_j}J(\theta) = \frac{\delta}{\delta\theta_j}\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \quad (5)$$

Now by taking partial derivative of θ_0 case, θ_1 case up to θ_n case repeatedly, the parameter values will be obtained.

$$\theta_j := \theta_j + \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i - y^i)x_j^i) \quad (6)$$

The process iterates and updates the values of θ_i until convergence. α is the learning rate. A smaller α value represents small descent and longer iteration, while a bigger α value represents bigger steps in the descending process. In this study we have chosen α = 0.01. Repeatedly θ_i is calculate until the $J(\theta)$ is minimized. During this process the θ_i values are updated simultaneously.

2.2. Proposed method for feature selection

Using the basic purpose of linear regression analysis, we have predicted the features from a subtype of a dataset based on the knowledge acquired from its counterpart. As previously discussed, one of the subtypes of control dataset or the subtype of a both malignant dataset is used to build the linear regression model for the feature prediction. Gradient descent algorithm was used for finding the parameters values of a linear regression model that yields lowest $J(\theta)$ value (Equation 6) by simultaneously updating θ_j values. The classic regression model contains an output variable *y* which is defined by one or more input variable(s) *X*. In this study we do not have a single output variable rather every gene is considered as an output variable one at a time. While a gene is considered as a dependent variable, the other set of features act as explanatory variables to best fit the regression line. This procedure is done for all the features individually. Using all the features in the dataset to make the regression line is not feasible. For that reason, we have eliminated the redundant features by measuring the similarity of their expression values (steps 1-4 of Algorithm 1). Considering each of the genes in the base subtype as an output variable one at a time and all other genes as input variables, the parameters are computed. The expression for each of the genes can be represented as

$$\begin{aligned}
G_\theta(1) &= \theta_0 g_0 + \quad \theta_2 g_2 + \theta_3 g_3 + \cdots + \theta_n g_n \\
G_\theta(2) &= \theta_0 g_0 + \theta_1 g_1 + \quad \theta_3 g_3 + \cdots + \theta_n g_n \\
G_\theta(3) &= \theta_0 g_0 + \theta_1 g_1 + \theta_2 g_2 + \quad \cdots + \theta_n g_n \\
&\vdots \\
G_\theta(n) &= \theta_0 g_0 + \theta_1 g_1 + \theta_2 g_2 + \cdots + \theta_{n-1} g_{n-1}
\end{aligned}$$

This process is repeated for all the genes in the corresponding subtype until $G_\theta(n)$ Considering known output variable $G_\theta(i)$ and input variables $\{g_1, g_2, \cdots, g_n\}$, we compute the parameters θ_i for each of them. In this way we obtain a $(n \times n)$ parameter matrix (Θ). Using the transpose of this matrix we statistically predict the gene expression values in the other subtype of the training dataset (Equation 7).

$$G'_\theta(i) = \Theta_i^T g'_i \quad (7)$$

Θ^T is the transpose of the Θ matrix obtained previously and g'_i is the set of expression values of the

dataset on which the regression model is applied. The divergence of the expression values from the actual gene expression values gives us the hint of distinctiveness of that gene and a possibility of that gene to be considered as a potential feature. The complete procedure of feature selection is depicted in Algorithm 1.

Algorithm 1

Input: Two subtypes of training datasets D_1 and D_2. N is the number of features and M is the number of samples.

Output: Set of features.

1: calculate the mean values for D_1 and D_2 on every features

2: find the difference of mean values between D_1 and D_2

3: sort the features based on their difference of mean values

4: remove the features with smaller difference (redundant or similarly expressed)

5: apply linear regression model considering each feature as a output variable and all other features as input variables

6: compute parameters (θ_i) for each of the G_θ using gradient descent algorithm

7: build parameter ($n \times n$) matrix (Θ) using D_1

8: compute the $G'_\theta(i)$ value for each gene using Θ^T applied on D_2

9: calculate the divergence of expression values from the base model

10: sort the genes based on their divergence values from the regression line in descending order

11: select the top-*n* most important of features for classification

The algorithm works in twofold. At first a feature reduction technique based on the knowledge of gene expression is applied after dividing the training dataset into two subtype, D_1 and D_2. Biologically, from the concept of microarray gene expression analysis, genes with similar expression do not bear much information about its distinctiveness. The idea is simple; if a particular gene expressed in both the samples then that gene will not give us any useful information about the specialty of it. That is why algorithm 1 removes the redundant features by measuring the similarity of their expression values (line 1-4). It also creates the platform for the linear regression to be applied on D_1 to build the parameter matrix Θ (line 5-7). The main purpose of linear regression is for predicting the values. We tried to predict the expression of genes in D_2 based on the regression line generated using D_1 dataset. As both subtypes contain same number of genes, and if there is no dissimilarity between them, all the genes in both the subtypes should express likewise. As this is not the case, which is how distinctive genes can be found by measuring the divergence of gene expression values from the regression line using Equation 7 (line 8-9). Based on the divergence values, most discriminative feature set is selected (line 10-11). Most deviated genes are considered as most important and most distinctive genes.

We have made our standard training subtype in different combinations of training genes. Thus we considered the best set of genes those fit regression model as our standard model. After the number of genes reduced by similarity measurement, we worked with 50% of the features for building regression model and then 10% of genes for creating different subset of features. However, 20% genes were too many to consider without much improvement in result and 5% genes were too less for consideration.

After the feature set has been selected, we can simply pick top-*n* features with highest divergence. The performance of the classification and feature selection method depends on the number of features. Too few features give the classifier insufficient classification power while too many features may cause over-fitting and add noise. In our analysis, we have used several set of features. Initially the performance improves with the increase of the number of features but the improvement becomes steady with the gradual increase of the number of features in most of the cases.

3. Results

To find out the acceptance of the proposed feature selection method, we have setup the experiment with six microarray datasets. In this process, we have compared our algorithm with 5 other attribute selection techniques: ReliefF, Correlation based feature selection (CFS), two attribute selection measures using decision tree: using gain ratio value and information gain value and selecting attributes using correlation analysis: Chi-Squared measure. As previously stated, the proposed feature selection technique can be used as a pre-processing step in any classification task. To prove that statement, we have used different classifiers i.e. SVM, Random forest, AdaBoost, Naive Bayes and KStar (K*).

3.1. Datasets

The datasets are obtained from different authors. Each of the dataset is a two class microarray data. We have separated them into training datasets and test datasets. The datasets are converted into the convenient format for performance evaluation without losing any information. The list of datasets appears in Table 1.

Table 1 Microarray Datasets used for the experiment. *S* is the number of samples and *P* is the number of probes (features).

Dataset	Ref.	*S*	*P*	Classes
Leukaemia	[18]	128	12625	B-cell and T-cell
Colon	[19]	62	2000	Negative and positive
DLBCL	[20]	77	7070	DLBCL and FL
Lung cancer	[21]	181	12533	MPM and ADCDA
Prostate	[22]	102	12533	Normal and tumor
Prostate cancer	[23]	20	12625	Dependent ad independent

3.2. Performance Evaluation Criteria

Using the benchmark of 6 datasets, we tested 5 classifiers and 5 feature selection algorithms. We used MATLAB for the implementation of the proposed algorithm and publicly available weka tool [24] for the classifier implementation. For SVM classifier we have used 10-fold cross validation. The random forest procedure was run with 10 trees. AdaBoost classifier uses 10 iteration and weighted threshold of 100. For K* globalBlend parameter values is set to 20.

3.3. Performance evaluation

We have tested our method with different number of features to test the performance of the method. Figure 1 shows the comparison in classification accuracy among the attribute selection methods. With few exceptions, the proposed attribute selection method shows better results. Wherever the proposed method produces a bit less accuracy than other attribute selection methods, it compensates by selecting less number of features. SVM, Random forest and AdaBoost demonstrated better performance than Naive Bayes and K* for most of the datasets. Classification accuracy for Leukaemia dataset is 100% using the features selected by proposed feature selection technique. Moreover, the credibility of the proposed method can be inferred from the classification accuracy of prostate dataset (98.36%) and prostate cancer dataset (91.67%) which is substantially higher than that of other feature selection methods. Although the difference in classification accuracy for DLBCL (93.48%) and Lung cancer dataset (97.98%) are not significant than the other methods, however, considering the number of features used for classification, the final outcomes stands out.

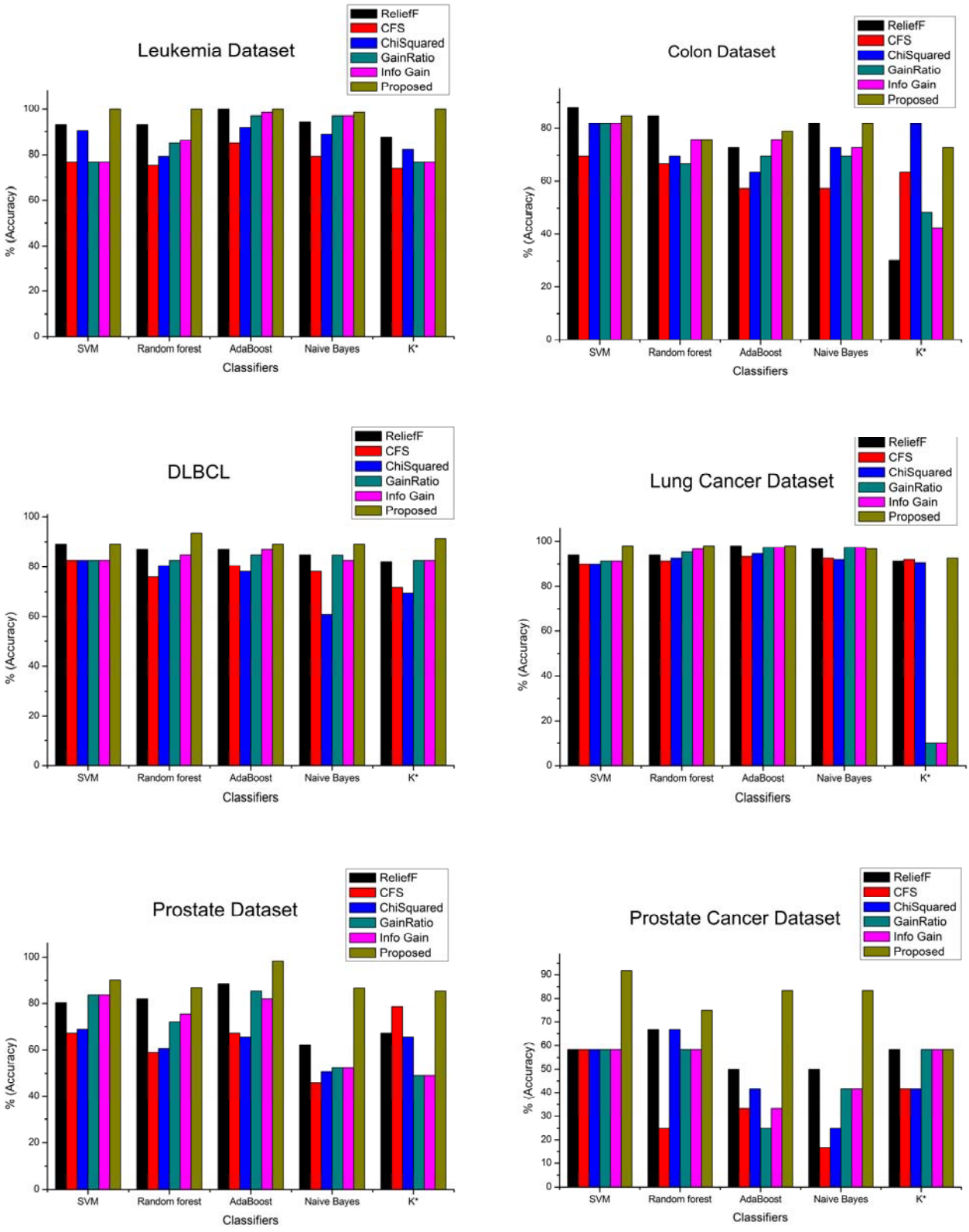
Figure 2 depicts the utilization of number of features, selected by different attribute selection methods for classification. The comparison shown in the figure approves the claim of using fewer features for better classification accuracy by the proposed feature selection method. The effect of change in number of features is depicted in Figure 3. All of the datasets show accuracy alteration with variety in number of features. In most of the datasets the effect of number of feature change becomes steady after some number. In our case the shift stabilizes after we select 60 features with an exception for Prostate cancer dataset. However, for Prostate cancer too, the best accuracy was gained for 30 features.

An efficient way of selecting small number of genes is always difficult due to the heavy expenses associated with the validation of genes. However, we have applied a forward feature selection approach for finding the optimal set of features. After extracting useful features by different attribute selection method and the proposed method, a random set of features are selected at the first iteration for classification accuracy measurement. The next set of features is adopted only if the result produced by the latter is better than the former. Otherwise the original set is kept. Although no biological validation has been made about the selected features, we have trusted on the best of set of features as they demonstrated their ability to classify accurately. In Table 2, we have provided the name of the genes (as given in the datasets) responsible for the best classification accuracy, selected by our proposed method.

4. Discussion

We have proposed a new algorithm for feature selection based on linear regression analysis. The method works by measuring the deviation of genes in experimenting type from the standard type. The algorithm handles one of the classic problems of microarray dataset - redundancy. Eliminating unnecessary genes by measuring their similarity in their expression values, results a considerable less number of features (gene) to work with using linear regression model. The second section of the algorithm implements the regression model into the second subtype of case dataset to measure the divergence of the gene expression values from the base regression line. After the features have been selected, we have used a verity of classifier to find out the acceptability

Figure 1 Classification accuracy for six different microarray datasets



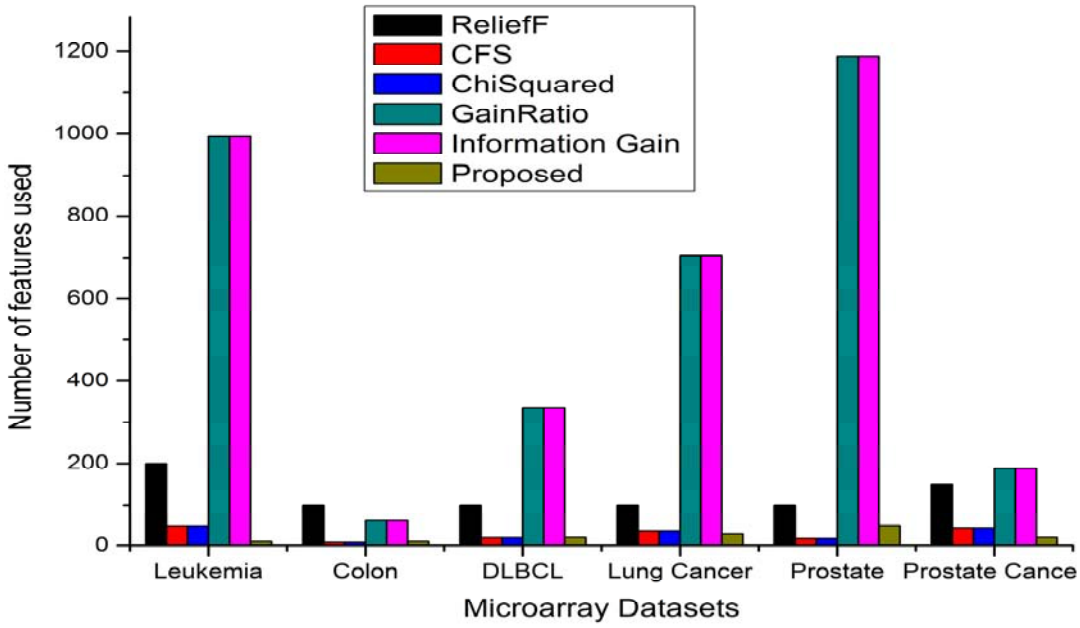


Figure 3 Classification accuracy for different subset of features

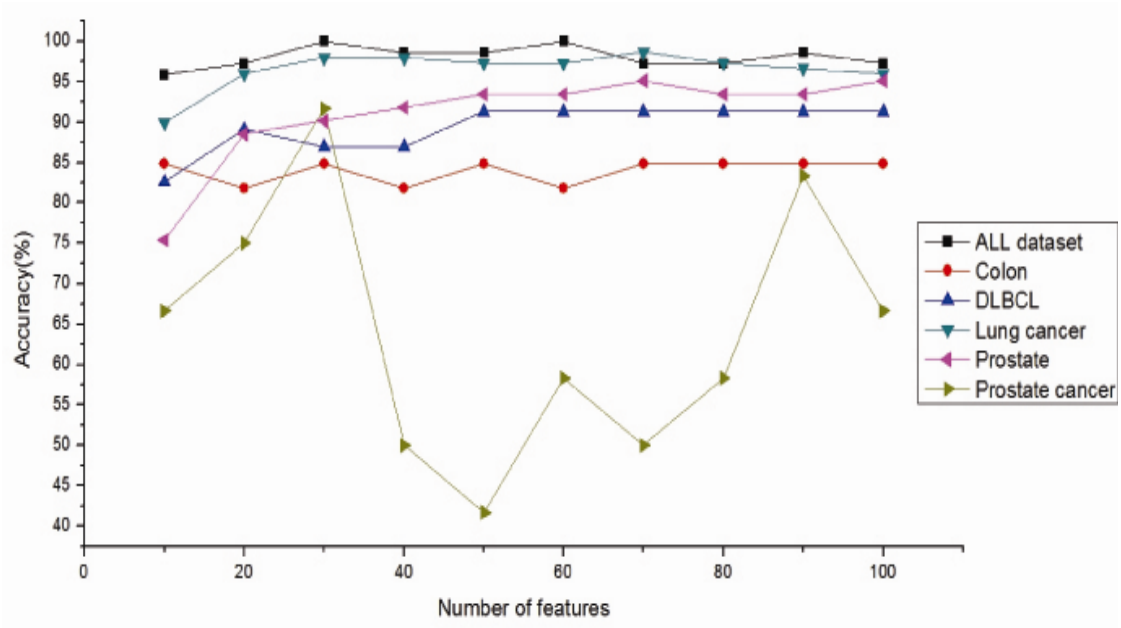


Table 2 Genes selected as features. (Gene names as given in the dataset)

Dataset	Genes selected as features
Leukaemia (10)	3663_at, 36108_at, 38585_at, 38511_at, 39389_at, 40567_at,40916_at, 38604_at, 34168_at, 37544_at
Colon (10)	attribute119, attribute807, attribute878, attribute306, attribute22, attribute143, attribute33, attribute4, attribute1810, attribute5
DLBCL (20)	D86974_at, J05070_at, L38941_at, U62317_rna3_at, M12529_at,V00594_s_at, X15183_at, M22382_at, M98539_at, D79205_at, M63379_at, L20941_at, J00105_s_at, M63438_s_at, X82240_rna1_at, U14970_at, U78027_rna3_at, HG4319-HT4589_at, X00351_f_at, M12886_at
Lung cancer (30)	33500_i_at, 33274_f_at, 33273_f_at, 3350_r_at, 33499_s_at, 41164_at, 37864_s_at, 35083_at, 37004_at, 41827_f_at, 31557_at, 32438_at, 33377_at, 38691_s_at, 41165_g_at, 38126_at, 37383_f_at, 33656_at, 33674_at, 34105_f_at, 31956_f_at, 327_f_at, 32744_at, 31950_at, 37954_at, 428_s_at, 35474_s_at, 31951_s_at, 1612_s_at, 32748_at
Prostate (30)	1922_g_at, 37747_at, 1009_at, 34987_s_at, 39921_at, 31963_at, 32150_at, 35316_at, 1895_at, 35279_at, 33995_at, 40283_at, 1806_at, 180_at, 40449_at, 31958_i_at, 39474_s_at, 39799_at, 40795_at, 32756_at, 36131_at, 217_at, 32336_at, 31386_at, 32750_r_at, 3840_r_at, 32053_at, 36987_at, 768_at, 615_s_at
Prostate cancer (20)	207783_x_at, 200094_s_at, 207574_s_at, 36711_at, AFFX-M27830_5_at, 33323_r_at, 200799_at, 221607_x_at, 221791_s_at, 200633_at, 213214_x_at, 35666_at, 209118_s_at, 202376_at, 208718_at, AFFX-r2-Hs28SrRNA-3_at, 208904_s_at, 20439_s_at, 211983_x_at, 201033_x_at

of the proposed method on test datasets. Our test on real case-control dataset shows an advantage of the new method over other filtering approaches.

We focus here on the better performance of classification with fewer features. We have seen that in most of the cases the proposed method produces better results with fewer numbers of features. In few cases the accuracy is similar to that of other methods. For example, our proposed approach demonstrated classification accuracy of 84.85% for colon dataset which is similar to that of other feature selectors and a bit less than the classification accuracy performed by ReliefF attribute selector (87.88%). However, the result is comparable and the number of features used by the proposed method (10 genes) is much less than that of ReliefF based feature selector (200 genes). Number of features used for prostate dataset classification by the proposed approach is higher (30 genes) than CFS based feature selection technique (17 genes). However the classification accuracy of prostate dataset (98.36%) using proposed feature selector far exceeds the performance of CFS feature selector (67.21%). The complete demonstration of the classification performance and the required number of features are summarized in Figure 1 and Figure 2. Figure 3 shows the number of features required for best performance. A small set of genes are desirable for the biologist for further analysis which our proposed method can provide. For further analysis the selected genes are listed in Table 2.

The proposed method selects features for classifying binary class microarray dataset. Although nearly accurate classification by the selected features annunciates the acceptance of the method, but the procedure still lacks in handling multiclass dataset classification. So, the incorporation of necessary techniques for multiclass dataset classification is being conceived. For multiclass classification, two of the multiple classes will be considered for creating regression model and selecting features through that process as described in this paper. The whole procedure will be continued for every possible combination of the classes. Each set of selected features will then be considered and the overlapped genes will be the final set of features. The computational time required for this process is not efficient enough. And it demands further analysis for better performance.

5. Conclusion

Feature selection using linear regression analysis appears promising. However future work should examine the validity of the results (i.e. the selected genes) from the biological point of view. Incorporation of other parameters needs to be considered for more realistic performance by the proposed model. In some special cases i.e. an increased number of

discriminative genes, requires efficient computation. More study about the datasets and interactions among the genes or the group of genes along with the expression values might give us more information about distinctiveness of a feature. Moreover an automated threshold selector might be helpful for selecting the exact number of features for the better result.

The algorithms are freely available at http://www.4shared.com/rar/EXITqY9T/Algorithm_Codes.html

References

1 Han J., Kamber M. (2001), "Data Mining: Concepts and Techniques", *Morgan Kaufman, second edition*.

2 Yu L. (2008), "Feature selection for genomic data analysis" *Taylor & Francis Group, LLC,* Binghamton University

3 Mitchell T. M. (1997), "Machine Learning", *McGraw-Hill,* New York.

4 Gutkin M., Shamir R., Dror G. (2009), "Simpls: A method for feature selection in gene expression-based disease classification", *PLoS ONE* vol. 4 (7).

5 Yu L., Liu H. (2004), "Efficient feature selection via analysis of relevance and redundancy", *Jounal of Machine Learning Research*, vol. 5, pp. 1205-1224.

6 Kohavi R., John G. H. (2000), "Multiple comparison in induction algorithms" *Machine Learning,* vol. 38 (3), pp. 309-338.

7 Saeys Y., Inza I., Larranaga P. (2007), "A review of feature selection techniques in bioinformatics", *Bioinformatics,* vol. 23, pp 2507-2517.

8 Robnik-Sikonja M., Kononenko I. (2003), "Theoritical and empirical analysis of relief and relieff", *Machine Learning,* vol. 53, pp. 23-69.

9 Hall M. A., "Correlation-based feature selection for discrete and numeric class machine learning", in: *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 359-366

10 Liu H., Motoda H. (2008), "Computational methods of feature selection", *Chapman &Hall/CRC.*

11 Sharma A., Imoto S., Miyano S. (2012), "A top-r feature selection algorithm for microarray gene expression data", *IEEE Transaction on computational biology and bioinformatics.* Vol. 9 (3), pp. 754-64.

12 Huang X., Pan W. (2003), "Linear regression and two-class classification with gene expression data", *Bioinformatics,* vol. 19, pp. 2072-2078.

13 Logson B. A., Mezey J. (2010), "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations", *PLoS Comput Biol*, vol. 6.

14 Keles S., Lann M. v. d., Eisen M. B. (2004), "Identification of regulatory elements using a feature selection method", *Bioinformatics*, vol. 18, pp. 2553-2561.

15 Tadesse, M. G., Vannucci M., Lio P. (2004), "Identification of DNA regulatory motifs using Bayesian variable selection", *Bioinformatics,* vol. 20, pp. 2553-2561.

16 Hedenfalk I., Duggan D., Chen Y., Radmacher M., Bittner M., Simon R., Meltzer P., Gustenson B., Esteller M., Kallioniemi O. P. (2001), "Gene expression profiles in heredity breast cancer", *N. Engl. J. Med,* vol. 344 pp. 539-548.

17 Golub T. R., Slonim D. K., Tamayo P. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, vol. 286, pp. 531-537.

18 Sabina C., Li X., Gentlemen R., Vitale A., Wang K. S., Foa F. R. M., Ritz J. (2005), "Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic pattern that identify lineage deviation and distinct mechanism of transformation", *Cancer Research,* vol. 11(20), pp. 7209-19.

19 Alon U., Barkai N., Notterman D., Gish K., Ybarra S., Mack D., "Broad pattern of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucliotide arrays" in: *Proceedings of National Academy of Science,* vol. 96, pp. 6745-6750

20 Shipp M. A., et. al. (2002), "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning", *Nat Med,* vol. 8, pp. 68-74.

21 Gordon G. J., Jensen R. V., Hsiao L. L., Gullans S. R., Blumenstock J. E., Ramaswamy S., Richards W. G., Suqarbaker D. J., Bueno R. (2002), "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma", *Cancer Research,* vol. 62, pp. 4963-4967.

22 Sing D., Febbo P. G., Ross K., Jackson D. G., Manola J., Ladd C., Tamayo P., Renshaw A.