Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 4 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Discriminative Gene Selection Employing Linear Regression Model

By Abid Hasan, Shaikh Jeeshan Kabeer, Md. Abdul Mottalib & Kamrul Hasan
University of Technology

Abstract - Microarray datasets enables the analysis of expression of thousands of genes across hundreds of samples. Usually classifiers do not perform well for large number of features (genes) as is the case of microarray datasets. That is why a small number of informative and discriminative features are always desirable for efficient classification. Many existing feature selection approaches have been proposed which attempts sample classification based on the analysis of gene expression values. In this paper a linear regression based feature selection algorithm for two class microarray datasets has been developed which divides the training dataset into two subtypes based on the class information. Using one of the classes as the base condition, a linear regression based model is developed. Using this regression model the divergence of each gene across the two classes are calculated and thus genes with higher divergence values are selected as important features from the second subtype of the training data. The classification performance of the proposed approach is evaluated with SVM, Random Forest and AdaBoost classifiers. Results show that the proposed approach provides better accuracy values compared to other existing approaches i.e. Relief F, CFS, decision tree based attribute selector and attribute selection using correlation analysis.

Keywords : *linear regression, feature selection, microarray dataset, classification.*

GJCST-C Classification : *D.2.2*

Strictly as per the compliance and regulations of:

Discriminative Gene Selection Employing Linear Regression Model

Abid Hasan [α], Shaikh Jeeshan Kabeer [σ], Md. Abdul Mottalib [ρ] & Kamrul Hasan [ω]

Abstract - Microarray datasets enables the analysis of expression of thousands of genes across hundreds of samples. Usually classifiers do not perform well for large number of features (genes) as is the case of microarray datasets. That is why a small number of informative and discriminative features are always desirable for efficient classification. Many existing feature selection approaches have been proposed which attempts sample classification based on the analysis of gene expression values. In this paper a linear regression based feature selection algorithm for two class microarray datasets has been developed which divides the training dataset into two subtypes based on the class information. Using one of the classes as the base condition, a linear regression based model is developed. Using this regression model the divergence of each gene across the two classes are calculated and thus genes with higher divergence values are selected as important features from the second subtype of the training data. The classification performance of the proposed approach is evaluated with SVM, Random Forest and AdaBoost classifiers. Results show that the proposed approach provides better accuracy values compared to other existing approaches i.e. Relief F, CFS, decision tree based attribute selector and attribute selection using correlation analysis.

Generalterms : algorithms, design, verification.
Keywords : linear regression, feature selection, microarray dataset, classification.

I. Introduction

The explosive growth and developments of microarray applications have enabled biologists and data mining engineers to study and observe thousands of gene expression data at the same time. Various attribute selection methodologies have been applied in the field of microarray data and in this particular case it is termed as gene selection as illustrated in [1]. Microarray data analysis has paved the way to cancer, tumor and other disease classification methods that can be used for subsequent diagnosis or prognosis. The problem of microarray data are many fold, firstly not all the data are relevant and often only a small portion of the data is related to the purpose of interest moreover noise and inconsistent data are prominent which hampers the search for the best genes for selection and classification [2]. However the major difficult aspect of microarray data is that the genes numbering in the thousands far outweighs the number of samples number in the lower hundreds if not less. This makes the task of building effective models particularly difficult and poses over fitting problems where the model does not perform well for novel patterns [3]. Thus feature selection methods being developed should be efficient in handling these issues.

Feature selection techniques can be generally divided into two broad categories depending on how the selection process interacts with classification model [4]. The first is the filter method where the importance of a feature is determined by scoring all the features based on their inherent attribute and retaining a portion of the features with higher scores while the low scoring features are removed as shown in many works including [5] and [6]. Filter methods are simple, fast and they do not require consultation with the classifier however the most obvious drawback is that it examines each feature individually and hence cannot harness the combined predictive power of features The second feature selection methodology is the wrapper model where a classification model is built by using a set of training set of features whose class labels are known and then the search for the optimal subset of features is done by repeatedly generating and evaluating possible feature using against the well known classifiers [7]. As the search for the solution is built into the classification process and as it considers the combinative predicting power of gene subsets the convergence time is higher the methods are usually complex.

Studies such as [8] and [9] have shown that the biological state of individuals is defined by their gene expression values. Therefore genes which have different expression profiles are more likely to properly identify biological states than genes having similar expression profiles. In this paper a linear regression model is proposed where one class of training dataset is considered as the base condition and generates the regression coefficients for each of the genes in the base class. Using the regression coefficients of the base condition a regression representation for the other class is generated and the difference in expression profiles between the genes of the base and non-base classes are measured. Genes with higher difference in expression profiles are given more importance and scoring of genes are generated. The base class serves as domain knowledge that is used to guide the search for discriminating genes in the dataset, [10] and [11] are works where domain knowledge was used to search for

Author α σ ρ ω : Department of CSE Islamic University of Technology, Gazipur 1704, Bangladesh. E-mails : aabid@iut-dhaka.edu, sjkabeer@iut-dhaka.edu, mottalib@iut-dhaka.edu, hasank@iut-dhaka.edu

the best features. The detailed procedure of the proposed method is provided in the following section. In the simulation and result analysis section it is seen that very high classification accuracy rates are achieved using only a very small number of the genes and the proposed method generated better results compared to other filtering approaches. The proposed approach has been applied on 6 microarray datasets and their effectiveness was determined by testing them in three different types of classifiers: Support Vector Machine (SVM), Random Forest and AdaBoost.

This paper is divided into 4 sections with section 1 giving an overview of the working domain and very brief introduction to the proposed approach. Section 2 elaborates the proposed approach in detail. Section 3 covers the simulation and result analysis part of the research where the proposed method is compared with Relief F, CFS, Chi-Squared value and Gain Ratio; it is seen that the proposed approach performs better then these existing methods. Section 4 provides the conclusion and provides scope for further research or development of this research work.

II. Proposed Approach

a) Theoretical Background

Linear regression is a statistical approach that can be used for predicting and forecasting. It has been traditionally used to model relationships between a set of explanatory variables $A=\{a_1,a_2,...,a_n\}$ and the output variable b_x. The idea is to derive a model using which the predictor or the output variable can be estimated using the explanatory variables [12]. In traditional feature selection applications the set of features are the input variables and the class labels are the output variables. Considering one feature a, the hypothesis function for this simple linear regression is

$$b_x = x_0 + x_1 a \tag{1}$$

where x_o and x_1 are the parameters and b_x is the predictor variable. The objective is to find the values of the parameters so that it best fits the data in the training set such that the features of unknown samples can be used for classification. x_j should be chosen such that $b_x(a)$ is as close to the training data (a,b) such that the following cost function is minimized

$$F(x_0,x_1)=\frac{1}{2m}\sum_{i=1}^{m}(b_x(a^i)-b^i)^2 \tag{2}$$

Here $F(x_0,x_1)$ is the cost function and m is the total number of samples in the training dataset. It is apparent that real world applications will require consideration of more than one feature and hence the hypothesis function will become

$$b_x(a) = x_0 + x_1 a_1 + x_2 a_{02} + x_3 a_3 + ... + x_n a_n \tag{3}$$

for convenience its assumed that $a_0 = 1$, therefore the feature vector A and parameter vector X becomes

$$A=\begin{bmatrix} a_0 \\ a_1 \\ a_3 \\ a_4 \\ . \\ . \\ . \\ a_n \end{bmatrix} \qquad X=\begin{bmatrix} x_0 \\ x_1 \\ x_3 \\ x_4 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

$b_x(a)$ can now be written as

$$b_X(a) = X^T A \tag{4}$$

X^T is the transpose of parameter vector and A is the vector of explanatory variables. So the corresponding cost function $F(X)$ which needs to be minimized for multiple variables, is the following:

$$F(x_0,x_1,x_2,...,x_n)=\frac{1}{2m}\sum_{i=1}^{m}(b_x(a^i)-b^i)^2 \tag{5}$$

Gradient descent is a very popular approach that has been used in many researches including linear regression. From earlier discussion it is clear that the idea is to minimize the cost function $F(X)$. Gradient descent algorithm helps find the parameter value which leads to the minimum cost. The representation of equation 5 in partial derivative term is

$$x_j := x_j + \alpha\frac{1}{m}\sum_{i=1}^{m}(b_x(a^i)-b^i)a_j^i \tag{6}$$

The algorithm starts with an arbitrary value x_j and keeps on changing by simultaneously updating x_j for $j=0,1,2,...,n$ until convergence for each of the x_j occurs.

b) Linear Regression on Microarray Dataset

Linear regression is a statistical approach that can be used for microarray datasets provides gene expression values for different samples. Using gene expression values to find out features and hence to classify novel samples is a common approach; however the application of linear regression to this task is a relatively fresh approach. In this proposed method the gene expression values of one class of samples of a two class microarray training dataset is used as the base class. Using this portion of dataset, a model is built which acts as the domain knowledge of the dataset.

Using this model the divergence in comparison to the gene expression values in the other class of the training dataset is measured. This tells us how much the latter deviates or diverges from the base class. From earlier discussions, equation 3 gives the expression for $b_x(a)$ where $A = \{a_1, a_2, ..., a_n\}$ are the features; for microarray applications, the genes are considered as features. As before $X = \{x_0, x_1, x_2, ..., x_n\}$ represents the parameters of the linear regression equation. In the proposed approach parameters X is being calculated by the cost function (equation 6) for each of the gene in the base class subtype of training dataset. Once we get the $n \times n$ parameter matrix X, it is applied to calculate $b'_x(i)$; the gene expression values in the non-base subclass of the training dataset. For each gene $b'_x(i)$ is calculated using all the gene data expect for its own hence

$$b'_x(1) = x_0 a_0 + x_2 a_2 + x_3 a_3 + ... + x_n a_n$$
$$b'_x(2) = x_0 a_0 + x_1 a_1 + x_3 a_3 + ... + x_n a_n$$
$$b'_x(3) = x_0 a_0 + x_1 a_1 + x_2 a_2 + ... + x_n a_n$$
$$\vdots$$
$$b'_x(n) = x_0 a_0 + x_1 a_1 + x_2 a_2 + ... + x_{n-1} a_{n-1}$$

These $b'_x(i)$ represent the statistical values of expression for each gene in the non-base subtype of the training dataset.

c) Proposed Algorithm

In our proposed method, basic idea of linear regression has been used. We have tried to predict a potential feature from one of the subtypes of microarray training datasets using the knowledge acquired from the other subtype of the same training dataset. At first the microarray dataset is divided into two segments test and training dataset in the similar way as most supervised learning algorithm does. One of the biggest problems of microarray data; redundancy has been handled by measuring the similarity in expression values of the genes in both types. We have eliminated those genes having similar expression values considering their ineffectiveness as important features for classification. Moreover, removing these genes gives the algorithm an efficient way of starting feature selection procedure. Training samples are then divided into two subtypes: S_1 and S_2 representing two subtypes of training data: base type and non-base type, built based on their class information. Next the parameter vector X for S_1 is generated using equation 6 and from the parameter vector X, $b'_x(a)$ is calculated for S_2. After the divergences and the differences are calculated, genes are sorted according to difference values in the descending order. From the sorted list of genes $N(N = 10,20,30,...,100)$ highest ranked genes are chosen and their classification accuracy is evaluated using different classifiers. Section 3 shows the detailed performance evaluation of the proposed approach and its superiority compared to other existing feature selection methods.

III. Materials and Methods

To find out how the proposed algorithm works, we have established the experiments using four different microarray datasets. We have compared our proposed feature selection algorithm with several other attribute selection procedures. Following sections describe a short description of microarray datasets and performance evaluations of the proposed method.

a) Datasets

The datasets are obtained from different authors. Datasets are converted into convenient way for this particular research.

The original prostate dataset was used in [13]. The dataset contains the 12,533 gene expression measurements of 102 samples. 50 of these 102 samples contain normal tissues not containing prostate tumor while 52 had prostate tumor.

Prostate cancer dataset was originally taken from dataset GSE2443 [14]. The dataset contains 12,627 gene expression values of 20 samples. Among them 10 samples contain androgen dependent tumor while other 10 contain androgen-independent tumor.

The lung cancer dataset contains two types of cancer: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of lung. Among 181 tissue samples, 31 of them had MPM and 150 of them had ADCA. Each of the samples was described by 12,533 gene expression value [15].

The colon dataset was used in [16]. The dataset contains 62 samples collected from colon cancer patients. 40 tumor biopsies are from tumors and 22 normal biopsies are from healthy parts of colon of same patients. The number of genes used in this expression is around 2000.

b) Performance Evaluation

The implementation of the proposed algorithm of feature selection was done on MATLAB and the performance evaluation of the selection set of features for classification was performed on publicly available weka tool [16]. We have used 10 fold cross-validation for SVM classifier. The random forest procedure was run with 10 trees and AdaBoost classifier uses 10 iteration and weighted threshold of 100.

i. *Results*

The classification accuracy of the features selected by proposed method and its comparison with

other method for different datasets is given in the following tables. *N* represents the number of features used by the classifier for classification.

Table 1 : Prostate dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random Forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	100	80.33	150	81.97	100	88.52
CFS	17	67.21	17	59.02	17	67.21
Chi-Squared value	17	68.85	17	60.66	17	65.57
GainRatio Value	1190	83.61	1190	72.13	1190	85.24
Proposed	30	90.16	20	86.66	50	98.36

Table 2 : Prostate cancer dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random Forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	200	58.33	150	66.67	100	50.00
CFS	44	58.33	44	25.00	44	33.33
Chi-Squared value	44	58.33	44	66.67	44	41.67
GainRatio Value	188	58.33	188	58.33	188	25.00
Proposed	20	91.67	20	75.00	10	83.33

Table 3 : Lung cancer dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	100	93.96	200	93.96	100	97.98
CFS	37	89.93	37	91.27	37	93.30
Chi-Squared value	37	89.93	37	92.62	37	94.63
GainRatio Value	705	91.27	705	95.30	705	97.31
Proposed	30	97.98	40	97.98	30	97.98

Table 4 : Colon dataset classification accuracy

Attribute selection method	Classifiers					
	SVM		Random forest		AdaBoost	
	N	Acc (%)	*N*	Acc (%)	*N*	Acc (%)
ReliefF	200	87.88	200	84.85	200	72.73
CFS	8	69.7	8	66.67	8	57.58
Chi-Squared value	8	81.82	8	69.70	8	63.64
GainRatio Value	62	81.82	62	66.67	62	69.69
Proposed	10	84.85	20	75.76	20	78.79

Another aspect of the proposed feature selection method is; we have not used any threshold for how many features for classification will be selected. Several different subsets of features have been used for classification and thus select the best subset based on its classifying ability. Figure 1 shows the error rate in classification by the classifier with any particular feature subset. For a particular feature selection *j* using a particular classifier *i,* the average error rate is calculated by $\frac{1}{3}\sum_{k} Accuracy(i, j, k)$ since we are evaluating the performance of a particular selector with different number of features.

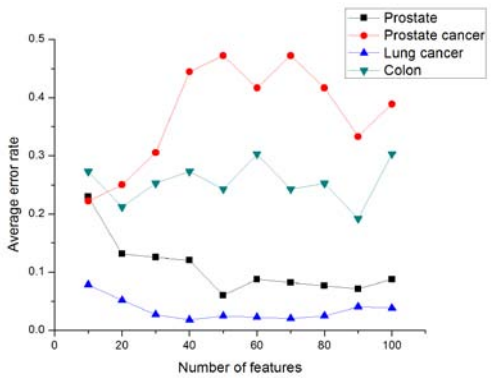

Figure 1 : Average error rate for different set of features

With the increase of the number of features, the error rate is decreased for most all the datasets. However, for prostate cancer dataset, although the error rate increases with first few subsets of features but at the end it too shows the same characteristics as the other microarray datasets shows.

ii. *Discussion*

We have proposed a new approach of feature selection using linear regression analysis. The algorithm works twofold. At the initial stage of the algorithm, we have eliminated redundant gene by measuring the similarity in expression values. Linear regression analysis then applied on one subtype (base type) of the training dataset to build the regression model. This model then applied on the other subtype (non-base type) of the training dataset to find out the divergence of the expression values of genes in that subtype. The more deviation shown by the gene, the more important it is considered as a feature. This way set of features selected for classification of the datasets.

Our main focus in this study is to classify accurately with less number of features. Table 1–4 shows the superiority of the classification accuracy by the features selected by the proposed method for different classifiers. Although, for colon dataset, classification accuracy by the features selected by ReliefF and CFS approach shows better result than the proposed method. However the result is still comparable and the number of feature selected by the proposed method is considerably fewer than the other method of attribute selection. Also Figure 1 summarizes the effect of different feature subsets for classification.

IV. Conclusion

Linear regression based feature selection shows promising results in classification of microarray datasets. The proposed approach might be applied on more microarray datasets and the results obtained might be used to improve some of the parameters of the proposed method. The results will also help to understand the performance of the proposed approach on a broader scale. The proposed approach can also be extended for multiclass approaches to be applied in other data mining domains. In the future Incorporation of other knowledge might help the proposed method to enhance the performance and significance of the result.

References Références Referencias

1. Kohavi, R and John, G.H. 1997. Wrappers for Feature Subset Selection, *Artificial Intelligence* Vol. 97 (1-2) (Dec. 1997), 273 - 324.
2. Iñaki Inza, Pedro Larrañaga, Rosa Blanco, Antonio J. Cerrolaza 2004. Filter versus wrapper gene selection approaches in DNA microarray domains, *Artificial Intelligence in Medicine* Vol 31 (2) (June 2004), 91–103.
3. Beatrice Duval and Jin-Kao Hao 2009, Advances in meta heuristics for gene selection and classification of microarray data, Briefings *in Bioinformatics* Vol. 11 (1) (July 2009), 127-141.
4. Yvan Saeys, Iñaki Inza, and Pedro Larrañaga 2007. A review of feature selection techniques in bioinformatics, *Bioinformatics* Vol.23 (19), 2507 -2517.
5. Hall M 1999. Correlation-based feature selection for machine learning. *PhD Thesis*. Department of Computer Science, Waikato University (1999).
6. Yu L, Liu H 2004. Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 2004; Vol. 5, 1205-1224.
7. Iñaki Inza, Basilio Sierra, Rosa Blanco, Pedro Larrañaga 2002. Gene Selection by Sequential Search Wrapper Approaches in Microarray Cancer Class Prediction, *Journal of Intelligent & Fuzzy Systems,* Vol. 12(1), 25-33.
8. Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J.S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M. Perou, Per E. Lønning, Patrick O. Brown, Anne-Lise Børresen-Dale, and David Botstein, 2003. Repeated Observation of breast tumor subtypes in independent gene expression data sets, *PNAS,* Vol. 100 (14), 8418-8423.
9. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*. (Oct 15) Vol. 286 (5439), 531-7.
10. Ting Yu, Simeon J. Simoff and Donald Stokes 2007. Incorporating Prior Domain Knowledge into a Kernel Based Feature Selection Algorithm. *Lecture Notes in Computer Science,* Vol. 4426/2007, 1064-1071.
11. Ofir Barzilay, V.L. Brailovsky1999. On domain knowledge and feature selection using a support vector machine. *Pattern Recognition Letters,* Vol. 20, (5), (May 1999), 475–484.
12. X. Yan and X. Su. Linear Regression Analysis. *World Scientific,* 2009.
13. D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P.w. Kantoff, T. R. Golub, W. R. Seller, 2002. Gene expression correlated of clinical prostate cancer behavior. *Cancer cell*. Vol.1 (2). p.p. 203-9
14. C. J. Best, J. W. Gillespie, Y. Yi, G. V. Chandramouli, et al. 2005, Molecular alternations in primary prostate cancer after androgen ablation therapy. Clin cancer Res. Vol. 1(11), 6823-34.
15. G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. R. Ramaswamy, W.

G. Richards, D. J. Suqarbaker, R. Bueno, 2002. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and meothelioma. *Cancer Research.* Vol, 62. p.p. 4963-4967

16. U. Alon, N. Barakai, D. Notterman, K. Gish, S. Ybarra, D. Mack 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *In Proceedings of National Academy of Science*. Vol. 96(12), 6745-6750
17. I. H. Witten and E. Frank 2005, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publisher.