ISSN 2063-5346

GENE SELECTION FOR TUMOR CLASSIFICATION BASED PARTIAL LEAST SQUARES DIMENSION REDUCTION METHOD

Dr. T.Shanmuavadivu, MCA., Ph.D.

Article History: Received: 01.02.2023 Revised: 07.03.2023 Accepted: 10.04.2023

Abstract

Analyzing gene expression data from DNA microarrays by commonly used classifiers is a hard task, because there are only a few observations but with thou-sands of measured genes in the data set. Partial least squares based dimension reduction (PLSDR) is superior to handling such high dimensional problem, but irrelevant features will introduce errors into the dimension reduction process and reduce the classification accuracy of learning machines. Here feature selection is applied to filter the data and an algorithm named PLSDR is described by integrating PLSDR with gene selection, which can effectively improve classification accuracy of learning machines. Feature selection is performed by the indication of t-statistics scores on standardized probes. Experimental results on seven microarray data sets show that the proposed method PLSDR is effective and reliable to improve the generalization performance of classifiers.

Keywords: Partial Least Squares, Dimension Reduction, Gene Selection

Assistant Professor, PG Dept of Computer Science, Arulmigu Palaniandavar Arts College for Women, Palani,Tirunelveli, Tamil Nadu, India-627012. Email: jansi_cse@msuniv.ac.in*@gmail.com*

DOI:10.31838/ecb/2023.12.s1-B.376

1. INTRODUCTION

DNA microarray experiments are used to collect information from tissue and cell samples regarding gene expression differences for tumor diagnosis. The output of micro array experiment is summarized as an $n \times p$ data matrix, where n is the number of tissue or cell samples, p is the number of genes (features). Here, p is always much larger than n, which hurts the generalization performance of most classification methods. To overcome this problem, we can either select a small subset of interesting genes (gene selection, feature selection) or construct K new components summarizing the original data as well as possible, with $K < p$ (dimension reduction, feature extraction).

Gene selection has been studied extensively in the last few years. The most commonly used procedures of gene selection are based on a score which is calculated for all genes individually and genes with the best scores are selected. Gene selection procedures output a list of relevant genes which can be experimentally analyzed by biologists. The method is often denoted as univariate gene selection, whose advantages are its simplicity and interpretability. However, interactions and correlations between genes are omitted during gene selection, although they are of great interest in system biology. Further- more, gene selection often fails to pick relevant genes, because the score they assign to correlated genes is too similar, and none of the genes is strongly preferred over another.

Partial Least Squares (PLS) was firstly developed as an algorithm performing matrix decompositions, and then was introduced as a multivariate regression tool in the context of chemometrics, PLS has also been found to be an effective dimension reduction technique for tumor discrimination Nguyen and Rocke proposed to use PLS for dimension reduction as a preliminary step for binary and multi-class classification. Barker and Rayens examined PLSDR in a formal statistical manner. Boulesteix and Dai *etc*. compared PLS with some of state-of-the-art classification and dimension reduction methods. Zeng *etc*. introduced PLS into the field of text classification as a text representation method . All these works have demonstrated the outstanding performance of PLSDR.

Considering of the fact that gene selection and dimension reduction algorithms have complementary advantages and disadvantages. Dimension reduction algorithms thrive on correlation among features but fail to remove irrelevant features from a set of complex features. Feature selection algorithms fail when all the features are correlated but do well with informative features. It would be an interesting work to combine irrelevant gene elimination and dimension reduction Nguyen and Rocke considered a preliminarily gene selection is helpful for PLSDR due to the fact that good prediction performance relies on good predictors, and some authors used t-statistic feature ranking method to select some significant genes before dimension reduction. Meanwhile, PLSDR also seems working well with the whole gene set benefited from its high computational efficiency. Furthermore, Boulesteix objected the usage of any preliminary feature selection for PLSDR they considered PLSDR benefits little from the preliminary selection of genes and the computational cost is too big if the cross-validation is used to select gene. But the comparative experiments to validate their notions are lacked. Therefore, few researchers had examined the effect of preliminarily gene selection to PLSDR, and how many genes should be selected is also remained as a puzzle. To examine the influence of gene selection to PLSDR by comprehensive experiments on real

microarray data. Furthermore, based on the idea that irrelevant genes are always harmful for classification, we propose a novel algorithm named PLSDR^g which integrating PLSDR with gene selection, which reduces the irrelevant genes by the indication of random features. Some notions used in this work are clarified here. Expression levels of p genes in n microarray samples are collected in an $n \times p$ data matrix $X = (x_{ij})$, $1 \le i \le n$, $1 \le j \le p$; of which an entry x_{ij} is the expression level of the jth gene in the ith microarray sample. As we only consider binary classification problems, the labels of the n microarray samples are collected in the vector $\mathbf{y}$. When the ith sample belongs to class one, the element y_i is 1; otherwise it is -1. Besides, II • II denotes the length of a vector. X^T represents the transpose of X, X^{-1} represents the inverse matrix of X. The matrices X and $\mathbf{y}$ used in the following are assumed to be centered to zero mean by each column.. Partial Least Squares Based Dimension Reduction PLS is a class of techniques for modeling relations between blocks of observed features by means of latent features. The underlying assumption of PLS is that the observed data is generated by a system or process which is driven by a small number of latent (not directly observed or measured) features. Therefore, PLS aims at finding uncorrelated linear transformations (latent components) of the original predictor features which have high covariance with the response features. Based on these latent components, PLS predicts response features $\mathbf{y}$, the task of regression, and reconstruct original matrix X, the task of data modeling, at the same time.

Let matrix $T = [\mathbf{t}_1, \ldots, \mathbf{t}_K] \in \mathrm{R}^{n \times K}$ represents the n observations of the K components which are usually denoted as latent features or scores. The relationship between T and

X is defined as:

$$T = XV$$

where $V = [\mathbf{v}_1, \ldots, \mathbf{v}_K] \in \mathrm{R}^{p \times K}$ is the matrix of projection weights. PLS determines the projection weights V by maximizing the covariance between the response and latent components.

Based on these latent components, X and $\mathbf{y}$ are decomposed as:

$$X = TP^T + E$$

$$\mathrm{Y} = TQ^T + \mathrm{f}$$

where $P = [\mathbf{p}_1, \ldots, \mathbf{p}_K] \in \mathrm{R}^{p \times K}$ and $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_K] \in \mathrm{R}^{1 \times K}$ are denoted as loadings of X and $\mathbf{y}$ respectively. Generally, P and Q are computed by Ordinary Least Squares (OLS). E and f are residuals of X and y respectively. By the decomposition of X and $\mathbf{y}$, response values are decided by latent features not by X (at least not directly). It is believed that this model would be more reliable than using OLS model on X directly, because the latent features are coincided with the true underlying structure of original data.

The major point of PLS is the construction of components by projecting X on the weights V. The classical criterion of PLS is to sequentially maximizing the covariance between response $\mathbf{y}$ and latent components. Ignoring the miner differences among these algorithms, the most frequently used PLS1. PLS1 determines the first latent component $\mathbf{t}_1 = X\mathbf{w}_1$ by maximizing the covariance between $\mathbf{y}$ and $\mathbf{t}_1$ under the constraint of II $\mathbf{w}_1$ II= 1. The corresponding objective function is:

$$\mathbf{w}_1 = \arg \max(\mathrm{cov}(X\mathbf{w}, \mathbf{y}))$$

$$\mathbf{w}^T\mathbf{w}=1$$

The maximization problem of

Equation (3) can be easily solved by the Lagrange multiplier method.

$$\mathbf{w}_1 = X^T \mathbf{y} / \text{ II } X^T \mathbf{y} \text{ II}$$

To extract other latent components, PLS1 model the residual matrices, which could not be modeled by previous latent features, as new X and $\mathbf{y}$ sequentially. To obtain the residuals, PLS1 deflate matrices X a nd $\mathbf{y}$ by subtracting their rank-one approximations based on $\mathbf{t}_1$.

2. LITERATURE REVIEW

Literature Review contains a critical analysis and the integration of information from a number of sources, as well as a consideration of any gaps in the literature and possibilities for future research.

(i) CLASSIFICATION BASED

Classification learning can deal with more than two class instances. In practice, the learning process of classification is to find models that can separate instances in the different classes using the information provided by training instances. Thus, the models found can be applied to classify a new unknown instance to one of those classes. Putting it more prosaically, given some instances of the positive class and some instances of the negative class, and it can be used as a basis to decide if a new unknown instance is positive or negative . This kind of learning is a process from general to specific and is supervised because the class membership of training instances is clearly known. In contrast to supervised learning is unsupervised learning, where there are no pre-defined classes for training instances. The main goal of unsupervised learning is to decide which instances should be grouped together, in other words, to form the classes.

(ii) SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) is a kind of a blend of linear modeling and instance-based learning which uses linear models to implement nonlinear class boundaries. It originates from research in statistical learning theory . An SVM selects a small number of critical boundary samples from each class of training data and builds a linear discriminant function (also called maximum margin hyperplane) that separates them as widely as possible. The selected samples that are closest to the maximum margin hyperplane are called support vectors.

3. PLSDR with irrelevant genes elimination

PLSDR is famous for its computational efficiency which can handle thousand of genes at one time. However, researchers often neglect removing irrelevant features, which is an interesting and critical issue for its application. gene selection has the following benefits.

(a) Gene selection improves the classification accuracy. In general, original microarray data sets have some irrelevant and noise genes, which will influence the performance of dimension reduction. In practical, biologists often expect noises are reduced, at least in some extent, during the stage of dimension reduction. But, if some irrelevant and noise genes are reduced beforehand, we can expect the performance of dimension reduction will be improved. We will try to examine this statement in our experiments.

(b) Gene selection avoids high computational complexity. Any additional gene selection procedure will bring some extra computation, but the computational complexity must not be too high. Boulesteix objected the preliminarily gene selection for PLSDR mainly because of the huge computational complexity of cross-validation. Gene selection improves the interpretability of the components. The meanings of the components are always

difficult to be interpreted in dimension reduction. Biologists often analysis the relation between extracted components and original features by the coefficients, but it is obscured by the huge number of genes. Reducing the number of original features is obviously helpful when the components are needed to be related with original genes manually.

PLSDR Algorithm:

Input: Training data set X with **y**

Step-1: Construct 100 standardized random features, get the mean value δ of their t- statistic scores with **y**;

Step-2: Compute t-statistic scores of genes in X, eliminate those whose t-statistic scores are no greater than δ;

Step-3: Train the PLSDR model N on the data subset as output.

Output: The PLSDR model N

Name	Samples	Class Ratio	Features
Breast Cancer	**97**	**46/51**	**24,481**
CNS	**60**	**21/39**	**7,129**
Colon	**62**	**22/40**	**2,000**
DLBCL	**47**	**23/24**	**4,026**
Leukemia	**72**	**25/47**	**7,129**
Ovarian	**253**	**91/162**	**15,154**
Prostate	**136**	**59/77**	**12,600**

Table-1: Class ratio of Gene Selection

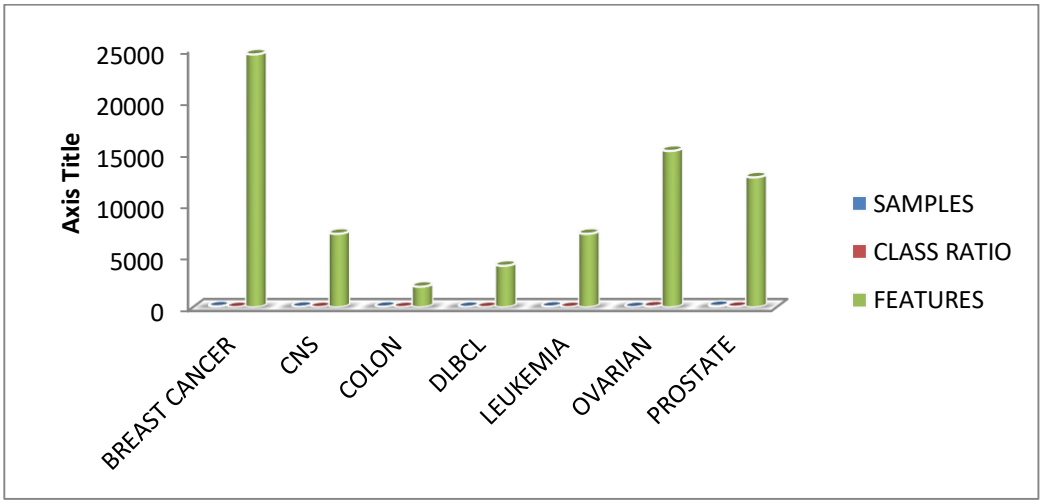

Figure-1: Class ratio of Gene Selection

Breast Cancer used DNA microarray analysis on primary breast tumors and applied supervised classification methods to identify significant genes for the disease. The data contains 97 patient samples, 46 of which are from patients who had developed distance metastases within 5 years (labeled as "relapse"), the rest 51 samples are from patients who remained healthy from the disease after their initial diagnosis for an interval of at least 5 years (labeled as "non-relapse"). The number of genes is 24,481 and the missing values of "NaN" are replaced with 100.

CNS developed a classification system based on DNA microarray gene expression data derived from patient samples of Embryonal tumors of the central nervous system (CNS). The data set used in our study contains 60 patient samples with 7,129 genes, 21 are survivors and 39 are failures.

Colon used Affymetrix oligonucleotide arrays to monitor expressions of over 6,500 human genes with samples of 40 tumor and 22 normal colon tissues. Expression of the 2,000 genes with the highest minimal intensity across the 62 tissues was used in the analysis.

DLBCL used gene expression data to analysis distinct types of diffuse large B-cell lymphoma (DLBCL). DLBCL is the most common subtype of non-Hodgkin's lymphoma. There are 47 samples, 24 of them are from "germinal centre B-like" group and 23 are "activated B-like" group. Each sample is described by 4026 genes. The missing values in the data set are replaced by the corresponding averaged column values.

Leukemia The acute leukemia data set of 72 bone marrow samples with 47 ALL and 25 AML. The gene expression intensities are obtained from Affymetrix high-density oligonucleotide microarrays containing probes for 7,129 genes.

Ovarian identified proteomic patterns in serum to distinguish ovarian cancer from non-cancer. The proteomic spectral data includes 91 controls (Normal) and 162 ovarian cancers, each sample contains the relative amplitude of the intensity at 15,154 molecular mass/charge (M/Z) identities.

Prostate used microarray expression analysis to deter- mine whether global biological differences underlie common pathological features of prostate cancer and to identify genes that might anticipate the clinical behavior of Prostate tumors. The data set contains 77 prostate tumor samples and 59 non- tumor prostate samples with 12,600 genes. The linear version of Support Vector Machine (SVM) with $C = 100$ and the k nearest neighbor (kNN) with one nearest neighbor as the classifiers, which are trained on the training set to predict the label of testing samples. The cross-validation procedure is repeated 10 times, and the mean classification accuracy (ACC) is used to measure the performance.

4. RESULTS AND DISCUSSION

Gene selection to PLSDR, we used t-statistic gene ranking methods to select top l genes, where $l = 20, 50, 100, 200, 500, 1,000, 1,500$ and $2,000$ respectively. The comparative classification results on eight microarray data sets by using SVM and kNN gene selection to PLSDR is not definitely positive. Gene selection improved the performance of dimension reduction on the data set of DLBCL, but dramatically decreased the classification accuracy on the data set of Prostate. Meanwhile, PLSDR is insensitive to the preliminary gene elimination on some data sets, such

as the data set of Ovarian. Even for one certain performance improved data sets, the optimal gene number is not same with different classifiers, *i.e*. on the data set of DLBCL, top 20 is optimal for SVM and it is top 100 for *k*NN. In conclusion, the effect of preliminary gene selection heavily relies on the data set and the applied classifiers. Without validation, any attempt to dramatically reduce the size of feature set has the danger to harm the final generalization performance. So, we consider only reducing irrelevant genes from the original set are a wise alternative. It is also not necessary to select a tight gene subset due to the stage of dimension reduction which will project the data into a much smaller subspace. The ACC results of our proposed algorithm Compared with results of different size of subsets, our method achieves satisfactory results. Figure-2 shows the features of Gene Selection.

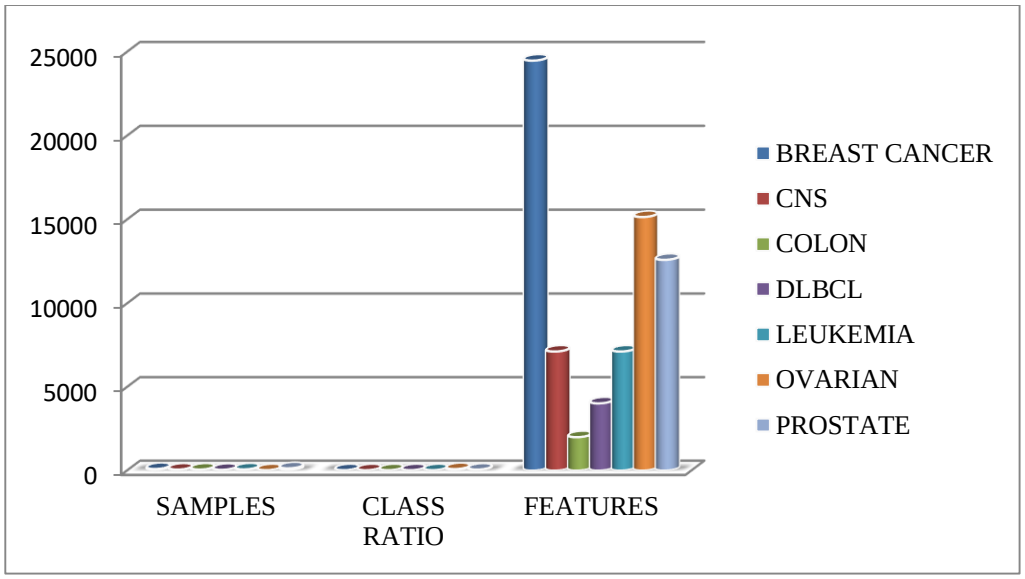

Figure-2: Features of Gene Selection

5. CONCLUSION

In this method, the bioinformatics and related fields whether a preliminarily gene selection would be applied before PLSDR is an interesting problem, which often was neglected. In this paper, we examined the influence of preliminarily gene selection by the t-statistic gene ranking method to PLSDR. The gene selection greatly rely on data sets and classifiers, furthermore, simply selecting some top ranking genes is not a good choice for the application of PLSDR. Based on the notion that irrelevant genes are always not useful for modeling, we proposed an efficient and effective gene elimination method by the indication of t-statistic scores of random features, which improves the prediction accuracy of learning machines for PLSDR.

REFERENCES:

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction

by gene expression. *Bioinformatics & Computational Biology*, 286(5439):531–537, 1999.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 6745–6750, 1999.

[3] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19(5):563–570, 2003.

[4] D. V. Nguyen, D. M. David, and M. Rocke. On partial least squares dimension reduction for microarray-based classification: a simulation study. *Computational Statistics & Data Analysis*, 46(3):407–425, 2004.

[5] J. J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):Article 6, 2006.