Fjrs.futa.edu.ng

FUTA Journal of Research in Sciences

ISSN: 2315 – 8239 (Print); E-ISSN: 2489 - 0413

FUTA Journal of Research in Sciences, Vol. 15(1), April, 2019: 96-101

THE t - STATISTIC AS A GENE SELECTOR IN CANCER RESEARCH WITH AN APPLICATION TO DISCRIMINANT ANALYSIS AND TEST OF LOCATION

Olusola Samuel Makinde, Sesan Adewunmi Ogundiran and Omolola Olubukola Fadugba

Department of Statistics, Federal University of Technology, P.M. B 704, Akure

Corresponding Author's e-mail: osmakinde@futa.edu.ng

ABSTRACT

Microarray analysis allows scientists to screen thousands of genes and determine the active, hyperactive or silent genes in normal or cancerous tissues. Hence, analytical methods should be developed to distinguish between cancer tissues of gene expression over normal tissues or other cancer tissues type. In this paper, gene selection of small subset of gene using component-wise two sample t statistic is performed on gene expression data. Moreover, we seek to find the effect of gene selection technique on support vector machine (SVM) for classification of cancerous gene level and micro-array data in general. The gene expression datasets for this paper are colon cancer data, leukaemia data and prostate cancer data. Findings from this study showed that discrimination between cancerous patients and non-cancerous patients using SVM when gene selection based on t test is employed tends to be much better than when all the genes are used in terms of probabilities of correct classification. Also, the statistical test of location can only be performed on selected genes rather than full datasets because of sparsity of gene expression data. Using the selected genes, a statistical test of location parameters in $\mathbb{R}^p$ is examined for different classes in gene expression data.

Keywords: Gene expression, t statistic, SVM, test of location parameter, gene selection.

INTRODUCTION

In recent years, the rapid development of DNA microarray technology has made it possible for scientists to monitor the expression level of thousands of genes with a single experiment. With DNA expression microarray technology, researchers are now able to classify different diseases according to different expression levels of normal and tumour cells, to discover the relationship between genes, and identify the critical genes in the development of disease. There are many active research applications of microarray technology, such as cancer classification (Golub *et al.*, 1999; Makinde, 2018), gene function identification, clinical diagnosis (Yeang *et al.*, 2001), and drug discovery studies. Microarray data analysis has been successfully applied in a number of studies over a broad range of biological disciplines including cancer classification by class discovery and prediction (Golub *et al.*, 1999), identification of the unknown effects of a specific therapy, identification of genes relevant to a certain diagnosis or therapy and cancer prognosis.

Gene expression refers to a complex series of processes in which the information encoded in a gene is used to produce a functional product such as a protein that dictates cell function. It involves several different steps through which DNA is converted to an RNA which in turn is converted into a protein or in some cases an

RNA, for example, genes encoding the necessary information for transfer RNAs and ribosomal RNAs (tRNAs and rRNAs).

High dimensionality, that is, a very large number of variables (genes) relative to the small number of observations characterises gene expression data. The gene expression data typically contain expression on 5,000–50,000 genes for less than 100 tumour samples. The t-test statistic is well known to statisticians for testing the difference between the means of two populations. In this paper, the possibility of using it for a two-class gene selection is raised. Effect of gene selection is investigated using a discriminant analysis approach based on support vector machine. Based on this result, a test of location parameters is conducted on classes of some gene expression data.

METHODOLOGY

Feature Selection:

Feature selection is commonly employed to extract features that have high leverage for classification and statistical inference. Feature selection techniques must lead to, but not limited to these, improved performance of statistical methods, better understanding of influential features and efficient computation (Makinde, 2018). Component-wise two sample t test for feature selection is proposed in Tibshirani *et al.* (2002) for high dimensional data.

Under the normality assumption, t-test is employed for testing the hypotheses $H_0: \mu_{Xi} = \mu_{Yi}$ versus $H_1: \mu_{Xi} \neq \mu_{Yi}$ where μ_{Xi} and μ_{Yi} are means of two different populations for gene i. The t-test rejects the null hypothesis H_0 at level of significance α if $|t| > t_{m+n-2}(\alpha/2)$, where

$$t = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \qquad (1)$$

s^2 is the pooled variance, $\bar{X}_i$ and $\bar{Y}_i$ are averages of the two samples for each of the genes, n and m are sample sizes for $\bar{X}_i$ and $\bar{Y}_i$ respectively. In this case, all the genes that exhibit significant differences among the competing groups are selected for further statistical analysis. However, number of contributory genes may still be greater than sample sizes and make classical statistical approach difficult to implement. Another approach is the use of t-statistic as feature selection techniques. That is to consider only k genes with largest t-statistic values. In this paper, t statistic is used to select statistically significant genes to test for equality of location parameters among different classes of gene expression levels. This is achieved by setting a threshold for t- statistic. The selected genes will be used for implementing support vector machines and test of equality of location vectors for gene expression data.

Support vector machines (SVMs):

Vapnik (1982) laid the foundation of SVMs. SVMs have been successfully applied for solving classification problem in different field of study. These include Rossi and Villa (2006), Makinde and Bodunwa (2018). Several versions of SVMs were proposed in Cortes and Vapnik (1995), Park and Liu (2009), Suykens and Vandewalle (1999), among others. We refer readers to Makinde and Bodunwa (2018) for details.

Test of location vector:

Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ is a random sample of gene expression levels of size n from p-dimensional population with mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$. Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_m$ be a random sample of gene expression levels of size m from a p-dimensional population with mean vector $\boldsymbol{\mu}_Y$ and covariance matrix $\boldsymbol{\Sigma}_Y$. It is assumed that $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_Y$. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ are independent of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots,$

$\boldsymbol{Y}_m$. We are interested in testing for equality of mean gene expression levels of both cancerous and non-cancerous groups. The null hypothesis and alternative hypothesis can be formulated as:

$$H_0: \boldsymbol{\mu}_X - \boldsymbol{\mu}_Y = \boldsymbol{\delta} \text{ versus } H_1: \boldsymbol{\mu}_X - \boldsymbol{\mu}_Y \neq \boldsymbol{\delta}$$

for some real vectors $\boldsymbol{\delta}$. Using Hotelling's T^2 test, H_0 is rejected at α level of significance if

$$(\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}} - \boldsymbol{\delta})' \left[a\boldsymbol{S}_{pooled}\right]^{-1} (\overline{\boldsymbol{X}} - \overline{\boldsymbol{Y}} - \boldsymbol{\delta}) > c^2,$$

$$a = \frac{1}{n} + \frac{1}{m} \qquad (2)$$

where $\overline{\boldsymbol{X}}$ and $\overline{\boldsymbol{Y}}$ are sample means of the two groups, $\boldsymbol{S}_{pooled}$ is the pooled dispersion matrix and $c^2 = \frac{(n+m-2)p}{n+m-p-1} F_{p,n+m-p-1}(\alpha)$.

It is observed that the test statistic is impossible to implement for gene expression data because of singularity of S_{pooled}, which is as a result of sparsity of the data. To overcome this problem, gene selection based on t-test is first employed and then followed by test of location.

DATA, RESULTS AND DISCUSSION

Data

Colon cancer information was extracted from DNA microarray data which give 62 tissues x 2000 gene expression values (Alon *et al*., 1990). Colon data set is available on R package *rda*. The 62 tissues include 22 normal and 40 colon cancer tissues and the matrix contains 2000 gene expression with high minimal intensity across the 62 tissues. Prostate cancer is a disease which only affects men. The prostate cancer data, available on R package *SPLS* were normalized, log transformed and has a standard mean and variance of zero and one respectively across genes as described in Dettling and Beuhlmann (2002). Leukaemia is a cancer which starts in blood-forming tissue, usually the bone marrow. It leads to the over-production of abnormal white blood cells, the part of the immune system which defends the body against infection. The leukaemia dataset, described in Golub *et al*. (1999), is available in R package *plsgenomics*. Table 1 presents the summary of gene expression data considered in this paper.

Discriminant analysis

Figure 1 presents the plot of computed absolute value of t statistic of the first 50 genes from two groups for each of leukaemia, colon and prostate cancer data. Using the gene selection methods described above, the informative genes are selected. The reduced gene expression data are split into training set and test set. The training set is used to train the classifier while the test set is used to estimate the performance of the developed system. Support vector machine (SVM) reads in the training data with their class labels and generates a classification model that assigns each gene expression level in the test set into one of the groups.

Table 2 below presents a comparison of classification performance on gene expression data with and without gene selection based on t statistic. It is observed that gene selection has greater importance in the analysis and helps to eliminate non-contributory genes in the data set and thus increase the accuracy in the analysis of gene expression data. Other gene selection methods in literature include shrunken centroid discriminant analysis (SCRDA) (Guo *et al*., 2007). However, Ogundiran (2017) has shown that the classifier performance does not significantly improve for very high dimensional data classification in some cases. In order to demonstrate this, support vector machine with feature selection based on SCRDA was performed on the data. Results in Table 2 confirm this claim.

Table 1: Summary of gene expression datasets

Data	Class labels	sample sizes	pooled sample size	Dimension (No of genes)
Colon cancer	Tumor	40	62	2000
	Normal	22		
Prostate cancer	Tumor	52	102	6033
	Normal	50		
Leukaemia	ALL	27	38	3051
	AML	11		

* ALL denotes acute lymphoblastic leukaemia; AML denotes acute myeloid leukaemia

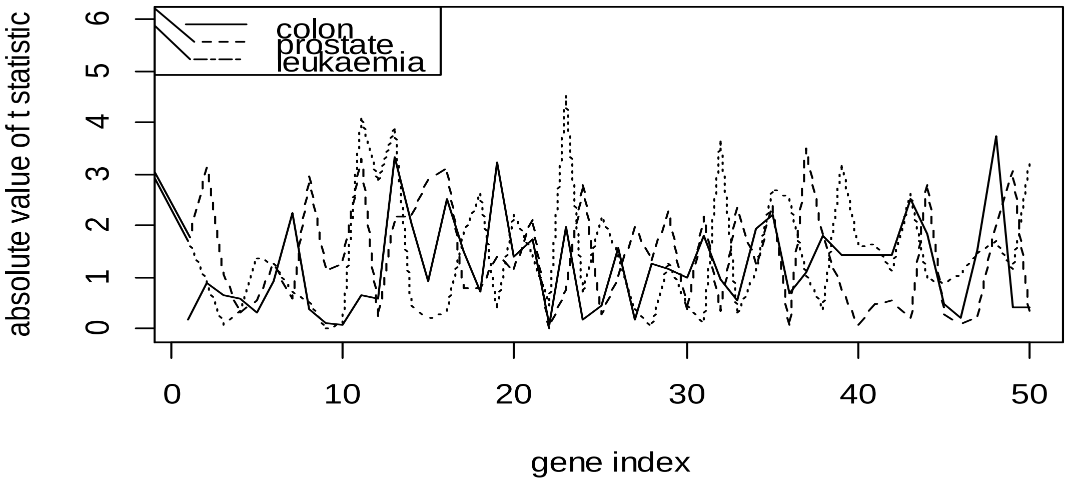

Figure 1: Plot of absolute value of t statistic for the first 50 genes in colon cancer, prostate cancer and leukaemia data.

Table 2: Comparison of classification performance on gene expression data with and without feature selection in terms of probabilities of correct classification.

Dataset	SVM	SVM + t test	SVM+SCRDA
Colon cancer	0.6452	0.8387	0.6774
Prostate cancer	0.7647	0.9216	0.9216
Leukaemia	0.6842	1.0000	1.0000

Test of equality of location vectors

Table 3 presents the values of Hoteling's T^2 statistic and the critical values for each of the three datasets. It is observed from Table 3 that the Hoteling's T^2 statistic value for each of the datasets exceeds the critical value. This shows evidence of significance of difference between class means in colon cancer, prostate cancer and leukaemia data. The relationship between this test of location and discriminant analysis can be explained by the fact that as the difference between the mean vectors widens; the more cleanly the separating hyperplane shows the distinction among the competing classes. That is, as $\mu_X - \mu_Y$ increases in magnitude, the distinction between the cancer types becomes clearer and the discriminant power of any classification method on such data increases, leading to formation of efficient separating hyperplane between the cancer types.

Table 3: Summary of result of test of location vectors for some gene expression data.

Dataset	Gene selection threshold	Number of selected genes	Hotelling T^2 statistic value	Critical value	Remark
Colon	3.0	16	63.7211	2.4986	Reject H_0
Prostate	5.5	32	391.388	2.3346	Reject H_0
Leukaemia	6.5	19	370.037	4.4066	Reject H_0

CONCLUSION

The gene expression data typically contain expression on 5,000–50,000 genes for less than 100 tumour samples. In this paper, gene selection using component-wise two sample t statistic was presented with the aim of investigating its effect on some statistical methodologies for gene expression data and removing noisy genes in the analysis of gene expression data. The effect of gene selection on classification based on support vector machine was measured using proportion of correct classification. Based on this result, equality of mean vector of gene expression levels was tested for two classes of cancerous tissues in gene expression data. Findings from this study showed that discrimination between cancerous patients and non-cancerous patients using SVM when gene selection based on t test is employed tends to be much better than when all the genes were involved in terms of probabilities of correct classification. Also, the statistical test of location using classical approach can only be performed on selected genes rather than full datasets because of sparsity of gene expression data. Using the selected genes, a statistical test of location vector based on Hotelling T^2 test showed significantly different class means for each of gene expression data. Optimal threshold for t statistic in order to determine the number of selected genes can be chosen by leave – one out cross validation of the error rates.

REFERENCE

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Mack, S.Y.D. and Levine, J. (1999). Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America , 96(12):6745-6750.

Dettling, M. and Beuhlmann, P. (2002) Supervised clustering of genes, Genome Biology, 3(12):1-15

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov,

J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, D.D. and Lander, E.S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 286(15):531-537

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays, Biostatistics, 8(1), 86-100

Makinde O.S. (2018). On rank distribution classifier in high dimension. Submitted.

Makinde O.S. and Bodunwa O.K. (2018) Classification of gene expression data: Distance based method. Kuwait Journal of Science. To appear.

Ogundiran S.A. (2017) Effect of gene selection on gene expression data with application to Support Vector Machine. B.Tech Thesis. The Federal University of Technology, Akure, Nigeria.

Park S.Y. and Liu Y. (2009) From the support vector machine to the bounded constraint machine. *Statistics and its Interface*, **2**, 285-298.

Rossi F. and Villa N. (2006) Support vector machine for functional data classification. *Neurocomputing*, **69**, 730 - 742.

Suykens J.A.K. and Vandewalle J. (1999) Least squares support vector machine classifiers. *Neural Processing Letters*, **9**(3), 293-300.

Yeung, K.Y., Haynor, D.R., and Ruzzo, W.L. (2001) Validating clustering for gene expression data, *Bioinformatics*, vol. 17, pp. 309-318

Vapnik V.N. (1982). Estimation of dependences based on empirical data. Addendum 1, New York: Springer-Verlag.