Regression Model In The Analysis Of Micro Array Data-Gene Expression Detection

Jamal Fathima .J.I[1] and P.Venkatesan[2]

1. Research Scholar -Department of statistics National Institute For Research In Tuberculosis, Indian Council For Medical Research,Chennai,India,

2.Department of statistics National Institute For Research In Tuberculosis, Indian Council For Medical Research,Chennai,India.

ABSTRACT:

Micro array technology has been increasingly used in medical studies such as cancer research. This technology makes it possible to measure the expression levels of thousands of genes simultaneously under a variety of conditions. Gene expression micro array data has proved an opportunity to measure the expression levels of thousands of genes simultaneously and this kind of high throughput data has a wide application in bio informatics research .Differential gene expression detection and classification of biological samples by micro arrays is a topic of interest in recent studies .In micro array studies the number of samples (n) is relatively small compared to the number of genes (p) per sample. Most classification procedures are more efficient when the sample size is large (greater than p).Several methods like non-parametric approaches have been developed sa far for condensing (shrinking) huge data sets. The SAM statistics proposed by Tusher et al and the nearest shrunken centroid proposed by Tibshirani et al are an ad hoc shrinkage methods. Recently Baolin Wu has proposed a new method using L_1 penalized linear regression model that is the t/F statistics for two class microarray data. In this paper we discuss a method for classification of micro array data into three classes, with an application to public micro array data.

Introduction

Data mining technique is a powerful tool in micro array analysis. It is applied in detecting the molecular variation in cancer data analysis. It helps in classification and detection of differentially expressed genes or in other words the genes which are more significant. In. micro array data the number of samples (n) is relatively small compared to the number of genes'p'(infinitely large). Micro array technology helps in monitoring the expression level of thousands of genes simultaneously. There are various methods in literature to study the estimation of the parameters in linear models.

The general linear model $y = x\beta + \varepsilon$ (1)

Where yi are the repressors, xij are the observations of the i^{th} sample. The estimate of $\hat{\beta}$ is obtained by the ordinary least squares. With huge data sets we have large number of predictors; data interpretation in such situations becomes difficult. Hence we go for selecting a subset of the predictor variables which exhibits a strongest effect. Subset selection leads to shrinkage of the parameters. There are different methods available for shrinkage of parameter-Ridge regression. LASSO regression (Tibshirani 1996). Wu proposed a new method using penalized linear regression model that is the t/F statistics for two class micro array data.As an extension to this we have proposed a model for three class micro array data with t/F statistics.

METHEDOLOGY

In our work we have derived the PLR for K=3 with an application to a micro array data.

Methodology

Let us consider a three class micro array data where the expression levels of 'p' genes are measured from n_1 samples of the first group, n_2 samples of the second group and n_3 samples of the third group.We have taken the log intensity ratios of test sample to normal from a publicly available micro array data. The class indicators are defined as

$$x_{ij} = \begin{cases} 1 & if\ j\ is\ from\ class\ k \\ 0 & otherwise \end{cases} \quad (1)$$

$$\sum_{j=1}^{n} x_{ik} = n_k\,, \quad k = 1,2,3 \quad \sum_1^3 x_{kj} = 1$$

The basic idea in differential gene expression detection is to compare the expression levels across the different classes.This can be done using the general linear model:

$$y_{ij} - \bar{y}_i = \sum_1^3 \beta_{ik} x_{kj} + \epsilon_{ij}\,, \quad j = 1,2,\ldots\ n, \qquad \bar{y}_i = \frac{\sum_{k=1}^3 x_{kj}}{n} \quad (2)$$

Where the response vector has been centered and the intercept is not included in the model. β_{ik} is interpreted as the difference between the mean of expression values across the classes $\bar{y}_i$ and mean expression of class 'k'. To test $H_0: \beta_{i1} = \beta_{i2} = \cdots = \beta_{ik}$,the F-statistics obtained by (Kutner et al ,2004) is

$$F_i = \frac{\sum_{k=1}^K n_k(\bar{y}_{ik} - \bar{y}_i)^2/(K-1)}{\sum\left(y_{ij} - \sum_{k=1}^K x_{ij}\bar{y}_{kj}\right)^2/(n-K)} \quad (3)$$

The L_1 Penalized linear regression model for three class micro array data is given by.

$$y_{ij} = \beta_0 + \sum_{k=1}^{2} \beta_{ik} x_{kj} + \epsilon_{ij} + \sum_{k=1}^{2} \lambda_k |\beta_{ik}| \qquad j = 1,2 \ldots .n, \qquad i = 1,2, \ldots .p \quad (4)$$

Minimizing the error sum of squares we get,

$$\sum \epsilon_{ij}^2 = Q_{min} = \sum_{j=1}^{n} \{y_{ij} - \beta_0 - \sum_{k=1}^{2} \beta_{ik} x_{kj}\}^2 + \sum_{k=1}^{2} \lambda_k |\beta_{ik}| \qquad j = 1,2 \ldots .n, \quad i = 1,2, \ldots .p \qquad (5)$$

Where $\sum_{k=1}^{2} \lambda_k |\beta_{ik}|$ is the penalty parameter based on the estimate of β_i's.The least square estimate of β_0 is given as $\acute{\beta_0} = \overline{y_i}$. $\beta_i^{'}s$ are obtained as solution to the equation,

$$Q = \sum_{j=1}^{n} \sum_{k=1}^{2} (\beta_{ij} y_{kj})^2 - 2n_k \left((\bar{y}_{ik} - \bar{y}_i)\beta_{ik} + \sum_{k=1}^{2} \lambda_k |\beta_{ik}| \right) + \sum (y_{ij} - \overline{y_i})^2 \quad (6)$$

$$\frac{\partial Q}{\partial \beta_1} = 2n_1 \beta_{i1} - 2n_1 (\overline{y_{i1}} - \overline{y_i}) = 0 \qquad (7)$$

$$\frac{\partial Q}{\partial \beta_2} = 2n_2 \beta_{i2} - 2n_2 (\overline{y_{i2}} - \overline{y_i}) = 0 \qquad (8)$$

$$\acute{\beta}_{i1} = (\overline{y_{i1}} - \overline{y_i})$$

$$\acute{\beta}_{i2} = (\overline{y_{i2}} - \overline{y_i})$$

$$SSTO = \sum (y_{ij} - \overline{y_i})^2 \qquad (9)$$

$$= (n-k)s_i^2 + \sum_k n_k (y_{ik} - \overline{y_i})^2$$

$$where, s_i^2 = \frac{1}{n-k} \sum_{k=1}^{2} \sum_{j=1}^{n} (y_{ij} - \overline{y_i})^2$$

$$SSE = \sum (y_{ij} - \bar{y} - \sum_{k=1}^{K} (\beta_{ij} y_{kj})^2 \qquad (10)$$

$$(n-k)s_i^2 + \sum_{k=1}^{K} n_k \left\{ |\bar{y}_{ik} - \bar{y}_i| - \left(|\bar{y}_{ik} - \bar{y}_i| - \frac{\lambda_k}{2n_k} \right)_+ \right\}^2 \qquad (11)$$

The test statistic to test H0:$H_o: \beta_o = \beta_{i1} = \beta_{i2}$ is given by

$$F_i^* = \frac{SSTO - SSE}{SSE/(n-3)}, \qquad (12)$$

$$\frac{(SSTO - SSE)}{SSE} = \frac{\sum_{k=1}^{K} n_k \left\{ |\bar{y}_{ik} - \bar{y}_i|^2 - \left\{ |\bar{y}_{ik} - \bar{y}_i| - \left(|\bar{y}_{ik} - \bar{y}_i| - \frac{\lambda_k}{2n_k} \right)_+ \right\}^2 \right.}{(n-k)s_i^2 + \sum_{k=1}^{K} n_k \left\{ |\bar{y}_{ik} - \bar{y}_i| - \left(|\bar{y}_{ik} - \bar{y}_i| - \frac{\lambda_k}{2n_k} \right)_+ \right\}^2} \quad (13)$$

$$= \begin{cases} \dfrac{\sum_{k=1}^{K} n_k \left\{ |\bar{y}_{ik} - \bar{y}_i|^2 - \left(\frac{\lambda_k}{2n_k} \right)^2 \right\}}{(n-K)s_i^2 + \sum_{k=1}^{K} n_k \left(\frac{\lambda_k}{2n_k} \right)^2} & if\ |\bar{y}_{ik} - \bar{y}_i| > \frac{\lambda_k}{2n_k} \\ 0 & if\ |\bar{y}_{ik} - \bar{y}_i| < \frac{\lambda_k}{2n_k} \end{cases} \quad (14)$$

$$= \frac{\sum_{k=1}^{K} n_k \left\{ |\bar{y}_{ik} - \bar{y}_i|^2 - \left(\frac{\lambda_k}{2n_k} \right)^2 \right\}_+}{(n-K)s_i^2 + \sum_{k=1}^{K} n_k \left(\frac{\lambda_k}{2n_k} \right)^2} \quad (15)$$

$$s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j:y_{kj}} (y_{ij} - \bar{y}_i)^2$$

s_i^2 is the pooled estimate for gene 'i'

In the nearest shrunken PAM(Tibshirani et al(2002,2003) the following ad hoc soft shrunken t-statistics and centroid are used.

$$d_{ik}^{/}(\Delta) = \frac{1}{m_k s_i} sign(\bar{y}_{ik} - \bar{y}_i)(|\bar{y}_{ik} - \bar{y}_i| - m_k s_i \Delta)_+$$

$$\bar{y}_{ik}^{/}(\Delta) = \bar{y}_i + sign(\bar{y}_{ik} - \bar{y}_i)(|\bar{y}_{ik} - \bar{y}_i| - m_k s_i \Delta)_+$$

Where Δ is the shrinkage parameter, $m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}$ $m_k s_i$ equal to the estimated error of the mean difference. Therefore the shrunken centoid $\bar{y}_{ik}^{/}$ is equivalent to the predicted mean response and for the penalized linear model $\lambda_k = 2n_k m_k s_i \Delta$. The sample mean with smaller sample size are shrunken more.

Differential gene expression detection and Classification of samples.

Gene expression detection are based on thresh holding. In this method a value is assigned to each gene. The commonly used score are the p-values and some variants of t/F statistics. We choose a cut off value t_0, the genes whose scores are greater than t_0, are considered to be differentially expressed. In large scale genomics False Discovery rate(FDR) is adopted (Benjamin and Hochberg,1995) to assess the significance of the set of genes.FDR is the expected proportion of false positives and used as a guidance for choosing the cut-off value t_0.

Selection of Penalty parameter (λ).

An optimal value of the parameter (λ), is chosen in such a way it minimizes the t-statistics and s_i (Baolin Wu 2005).In PAM method the tuning parameter (Δ) is chosen to shrink the genes with $d_{ik}^{/}(\Delta)$ to build the classifier using the ordinary sample classification method selects λ_k and differentially expressed genes using FDR simultaneously such that those genes with $F_i > 0$ are considered to be significant.

MA-PLOTS

Dudoit et al (2000) and Yang et al (2002) propose a plot of the log intensity ratio

$$M = \log\left(\frac{w^{(R)}}{w^{(G)}}\right) = logw^{(R)} - logw^{(G)}$$

Against the average intensity;

$$A = log\sqrt{w^{(R)}w^{(G)}} = \frac{1}{2}(logw^{(R)} + logw^{(G)})$$

MA plot is defined as a plot of M versus A. Difference M gives the difference of the log intensities and A represents the average log – intensity for the two cases (tumour and normal).MA Plot allows one to visualize whether the entire distribution of M values are centered with zero mean and that the mean is not influenced by intensity (A). By highlighting the control substances one can see whether controls represent a range of intensities and whether they can provide intensity – dependent corrections.

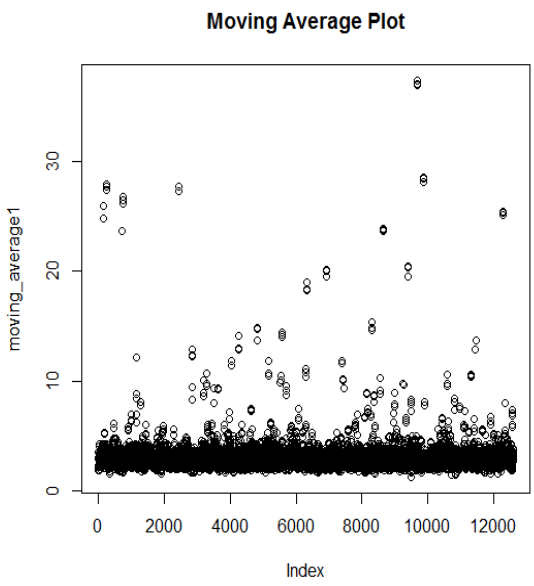
Application to micro array data:

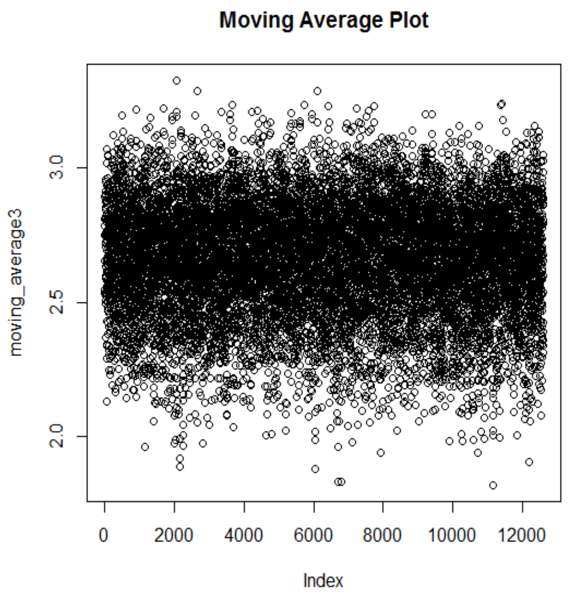
Data studied was obtained from NCBI GENE OMNIBUS .The log intensities of 12607 genes in breast cancer data set. Data filtering is carried out by stem and leaf plot method. The data is from two series consisting of ten samples classified into three group.(Group-I Control Group of four samples Group-II treatment of four samples and group –III normal tissue two samples). By applying the stem and leaf plot we obtained 103 outliers accounting for the differentially expressed genes.

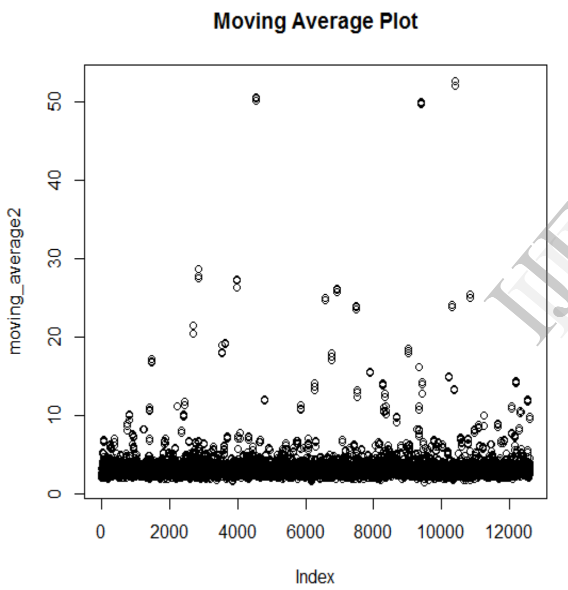
(i) Applying Penalized Linear Regression the F-values are computed .The penalty parameter assumes values between 27-32. The results of the analysis are given below.

```
  [1]  1.56814690 -3.15181788 -3.12537721 -3.30829915  1.23199958 -2.54760547
  [7] -1.83679637 -3.15144714 -2.35344712 -2.83162248 -3.37116858 -2.67160210
 [13] -3.23557261 -2.17760413 -3.30248209 -2.77773665 -3.13812775 -2.89492102
 [19] -2.82177235 -3.19007863 -3.30873299 -3.27180888 -2.25346074  0.04683591
 [25] -2.41863696 -3.22456195 -2.81525033 -3.24441516 -3.14731081 -3.30871023
 [31] -2.66142173 -3.35663912 -3.27932763 -3.08248367 -0.77087633 -3.12266861
 [37] -2.68687711 -2.78975337 -3.14221818 -3.22194199 -3.35056587 -2.60821314
 [43] -3.33134745 -3.13146780 -3.32753214 -2.96624432 -3.33831794 -3.36679320
 [49]  0.16841686 -2.97756091 -2.49976455 -3.26887157 -3.28949430 -3.04602559
 [55] -2.98803676 -1.92307855 -3.15406814 -3.12648960 -3.23488637 -3.22540094
 [61] -3.34295942 -3.25357299 -3.33428391 -2.72423383 -2.81634027 -2.48664388
 [67] -3.32415053  0.49152694 -0.96130543 -3.24241339 -2.19523225 -3.38982990
 [73] -3.24521517 -2.71273605 -0.36754717 -3.02913044 -3.36342970 -3.33855240
 [79] -3.28254737 -3.01205472 -3.27439553 -3.16781854 -3.38345432  0.70096562
 [85] -3.19632021 -3.27914049 -2.54490030 -3.29395519 -3.20748979 -2.96967295
 [91] -2.46364049 -2.79440363 -3.30057738 -3.27718930 -3.16562490 -3.27977569
 [97] -3.10082103 -1.02502409 -3.17844400 -3.32851266 -2.89882525 -3.29791417
[103]  0.11060070 -0.73925021
```


(ii) A moving average plot was constructed on the three groups the results of which are shown in the figures. One could see from the plots that in Group I and Group II most of the genes are under expressed and less than five percent are differentially expressed which account for maximum variation, where as in the plot of Group III the values cluster around the central value showing that the genes intensities come from normal tissues.







DISCUSSION.

In this paper we have studied the identification of differentially expressed genes in a regression frame work using three class micro array data. Generally linear model fitting is done by the method of least squares. As the number of genes is enormously large compared to the sample size ordinary least squares fails as it may lead to over fitting of the data. Hence we go for the penalized L_1 linear regression model for gene expression detection(Wu,2005) and the nearest shrunken centroid classification(Tibshirani et al 2002).

References

1. Alter. O., Brown, P.O. and Botstein.D. (2000) Singular value decomposition for genomewide expression data procession and modeling. Proc. Natl Acad. Sci. USA, 97(18). 10101-10106.
2. Benjamini yaov and Hochberg Yosef. – (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Society, Serie B57/91):289-300.
3. Baggerly, K.A., Coombes, K.R., Hess, K.R., Stivers, D.N., Abruzzo, L.V and Zhang, W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. J. comput. Biol., 8, 639-659.
4. Baldi,P and Brunak,s. (2001) Bioinformatics: The Machine Learning Approach. MIT Press. Cambridge. MT.
5. Botstein, D. and Brown, P.O. (2000). 'Gene shaving' as a method of identifying distinct sets of genes with similar expression patterns. Genome Biology 1, 1-21.
6. Callow. M.J. Dudoit, Gong, E.L. Speed, T.P. and Rubin E.M. – (2000) Micro array expression profiling identifies genes with altered expression in hdi deficient mice. Genome Research 10,2022-2029.
7. Churchill. G.A.(2002). Fundamentals of experimental design for cDNA microarrays, Nature Genetics 32, 490-495.
8. Dudoit. S. Yang W.H. Callow. J and Speed. T.P. (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA micro array experiments, Technical Report – 578.
9. Efron, B., Hasti, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. Annals of statistics, 32, 407-499.
10. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. 1999. Molecular classification of cancer. Class discovery and class prediction by gene expression monitoring. Science 286, 531-537.
11. Hastie. T.J. and Tibshirani. R(1990). Genaralized Additive Models, London Chapman hall.
12. Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation of non-orthogonal problems. Technometrics, 12, 55-67.

13. Kerr, K.K. and Churchill, G.A., (2001). Experimental design for gene expression microarrays. Biostatistics 2, 183-201.
14. Wu.Baolin (2005)-Differential Gene Expression Detection Using penalized Linear Regression models; the improved SAM statistics. Bio informatics, pages 1565-1571.
15. Wu.Baolin (2006) – Differential Gene Expression Detection and sample classification using penalized Linear Regression model, Bio informatics, pages 472-476.