Gene filtering with optimal threshold selection

Josep Bau-Macià[1], Jordi Solé-Casals[1], Cesar F. Caiafa[2,3] and Sergio Lew[4]

[1] University of Vic, Sagrada Família 7, 08500, Vic, SPAIN
[2] IAR-CONICET, C.C.5, (1894) Villa Elisa, Buenos Aires, ARGENTINA
[3] FIUBA, Av. Paseo Colón 850 (1063), Capital Federal, ARGENTINA
[4] IIBM-FIUBA, Av. Paseo Colón 850 (1063), Capital Federal, ARGENTINA

Abstract: Gene filtering is a useful preprocessing technique often applied to microarray datasets. However, it is no common practice because clear guidelines are lacking and it bears the risk of excluding some potentially relevant genes. In this work, we propose to model microarray data as a mixture of two Gaussian distributions that will allow us to obtain an optimal filter threshold in terms of the gene expression level.

Keywords: Gene expression; gene filtering; microarray data; MOG model.

1 Introduction

Non-specific gene filtering is a very useful technique as it increases the sensitivity of the microarray data analysis and reduces the dimensionality of the dataset. Thus, correct and stringent filtering will substantially reduce the problem of overfitting in classification problems. Several filtering approaches exist, some of them often used in combination. The most used are based on (i) filtering by expression level and (ii) filtering by gene variance across samples. These techniques involve the use of more or less subjective thresholds. The drawback of data-independent thresholds is that gene expression distributions are very variable between different microarray datasets and can result in too stringent or too loose filtering conditions. In this work we develop a data-driven selection of a threshold based on the minimization of the classification error.

2 Materials and Methods

ALL Dataset: The Acute Lymphoblastic Leukemia (ALL) data were reported by Chiaretti et al [Chiaretti et al, S. 2004]. We consider the comparison of the 37 samples from patients with the BCR/ABL fusion gene resulting from a chromosomal translocation (9;22) with the 42 samples from

the NEG group. They are available in the R package ALL. The comparisons conducted in this work maintain the criteria of removing the genes with inter-quartile range (IQR) below 0.5 used by Scholtens and Heydebreck [Scholtens D. and Von Heydebreck A., 2005] but using an optimized threshold for intensity filtering instead of $6.64 = \log_2(100)$.

Mixture Of Gaussian (MOG) model: We assume that each gene belongs to one of the following two classes: class C_1 (un-expressed genes), class C_2 (expressed genes), and each one of these classes can be well modeled using a Gaussian distribution with specific means μ_1, μ_2, and standard deviations σ_1, σ_2. Then, the probability density functions (pdfs) for gene expression values conditioned to a particular class is: $p(x|C_p) = \Phi\left(\frac{x-\mu_p}{\sigma_p}\right)$, $(p = 1, 2)$ with $\Phi(x) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}x^2\right)$ being the zero-mean and unit-norm Gaussian pdf. Then, the pdf for the variable x (gene expression value) is: $p(x) = \alpha_1 p(x|C_1) + \alpha_2 p(x|C_2)$, where $\alpha_p = P(C_p)$ $(p = 1, 2)$ are the probabilities of each class.

Fitting the MOG model: We use the Maximum Likelihood (ML) criterion to fit the model. Since it is difficult to obtain a closed form of the likelihood for a MOG model, a well-known solution is to use the Expectation-Maximization (EM) algorithm. **Optimal threshold selection:** Once the MOG model is fitted to the available data we need to determine the optimal threshold h to classify samples as belonging to C_1 $(x < h)$ or C_2 $(x \geq h)$. Our objective is to choose h such that the error of classification is minimized. It is easy to show that such a value of h must satisfy $\alpha_1\Phi\left(\frac{h-\mu_1}{\sigma_1}\right) = \alpha_2\Phi\left(\frac{h-\mu_2}{\sigma_2}\right)$, which can be explicitly solved.

Differential expression analysis: NEG samples were compared to BCR/ABL samples applying a Welch t-test for equality of the mean expression levels in the two groups in order to obtain the differential expression p-value for each gene.

3 Results

For ALL dataset, after 151 iterations the MOG model converged to an optimal intensity threshold (OIT) value of 4.17 on a log2 scale (Figure 1). This value is clearly lower than the 6.64 arbitrary intensity threshold (AIT) used by Scholtens and Heydebreck. Table 1 shows how the selection of the threshold affects to the number of significant genes ($p_{val} < 0.05$) discarded by the filtering process. Using an arbitrary threshold set up at 6.64 the total number of discarded significant genes is 101, which represents the 61.6% of significant genes of the whole dataset. On the other hand, using our optimal threshold at 4.17 determined by the MOG model, the total number of discarded significant genes is 37 which represents the 22.6% of significant genes of the whole dataset. Clearly our method increases the number of significant genes not discarded by the filtering process.

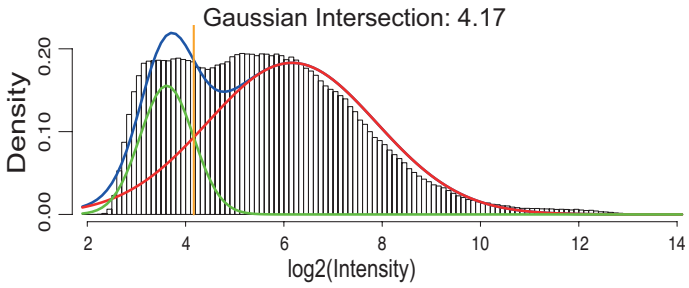

FIGURE 1. Gaussian mix model of the ALL data.

TABLE 1. Discarded genes depending on the selected threshold.

	#Genes	% Genes	#Genes $p_{val} < 0.05$	%Genes $p_{val} < 0.05$
Total	12625	100%	164	100%
Discarded with AIT ($h = 6.64$)	10231	81%	101	61.6%
Discarded with OIT ($h = 4.17$)	8599	68.1%	37	22.6%

4 Conclusions

A new method for the automatic selection of a threshold for filtering genes in microarray datasets has been proposed and compared to classical filtering techniques. Our experimental results on the ALL dataset demonstrates the advantage of using the proposed technique.

Acknowledgments: This work has been in part supported by the MINCYT-MICINN Research Program 2010-2011 (Ref. AR2009-0010) and by the University of Vic under the grants R0904 and R0901.

References

Scholtens D. and Von Heydebreck A. (2005). *Analysis of Differential Gene Expression Studies. Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health.* **Part III**, 229-248.

Chiaretti S., Li X., Gentleman R., Vitale A., Vignetti M., Mandelli F., Ritz J., Foa R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood.* **103**, 2771-2778.