A summer with genes: 'Simple' disease classification from microarray data

Jochen Einbeck* and Sam Jackson
Department of Mathematical Sciences,
Durham University

Adetayo Kasim,
Wolfson Research Institute for Health and Wellbeing,
Durham University.

April 14, 2015

In this article we report on the work carried out within the framework of a summer project, part–funded by an IMA small grant, in which an undergraduate student (the second author of this manuscript) developed and implemented methodology for disease classification from gene expression microarray data. While the original motivation for this study was the development of a correlation threshold for gene filtering, a general outcome of this research was that, using very simple statistical techniques (essentially at undergraduate level) but solid state–of–the–art validation routines, good classification accuracies can be obtained using relatively small–sized gene signatures. We applied the techniques on expression data for breast cancer tumour subtype classification, as well as for prediction of the presence or absence of Irritable Bowel Syndrome (IBS).

Introduction

Since the groundbreaking work by Golub et al. (1999), the problem of disease classification through microarray gene expression data has attracted a

*corresponding author, jochen.einbeck@durham.ac.uk, 0191 3343125

tremendous amount of attention in the medical, bioinformatical and statistical research literature, with tens of thousands of published articles over the last 15 years. A typical dataset consists of n subjects, each of which has a pre-regularised measure of mRNA expression for each of the p parameters or genes. Typically, n is in the tens or hundreds, and p in the thousands or tens of thousands, but in any case one will have $n << p$, rendering standard statistical techniques inapplicable. The key task in microarray–based disease classification is the selection of a few significant genes for which the expression values are particularly informative for discriminating the different categories.

As an example, consider the gene expression values of two particular genes from a microarray dataset collected from patients with lymph-node-negative breast cancer (Figure 0.1), which is introduced in full later on. The categories to be discriminated are estrogen–receptor positive (ER+) and negative (ER-). It is obvious that for the left–hand gene, the ER- category tends to correspond to high and the ER+ category to low expression values, rendering this gene very informative for this problem. In contrast, for the right–hand gene the expression values are not separating the groups well and hence do not seem to be a useful indicator for the estrogen–receptor category.

Methodologically, what is needed from this point is a formal procedure for identifying ('filtering') significant genes (or groups of genes), as well as a classifier to carry out the actual classification task. It is also possible to carry out both steps at once, using sufficiently strong regularization in the classification step. Driven by a research environment in which "many journals require methodological innovation" (Boulesteix, 2006) as publication criterion, these statistical techniques have become more and more advanced over recent years. Though these advances have undoubtedly been valuable (with impact far beyond the field of microarray analysis), the complexity of these approaches has left these with a black–box flavour, which may turn out to be detrimental to the acceptance of such methodology in clinical practice.

In this summer project we took a back–to–the–root approach, effectively only using

(a) for the gene selection, a 2–sample–t–test which captures effects of the type depicted in Figure 0.1; as well as
(b) a correlation threshold to eliminate highly correlated genes (this was the original motivation for this work);
(c) for the classification step, diagonal linear discriminant analysis.

We paid careful attention to the question of validation, and we achieved good (or very good) prediction accuracy rates for all considered scenarios.

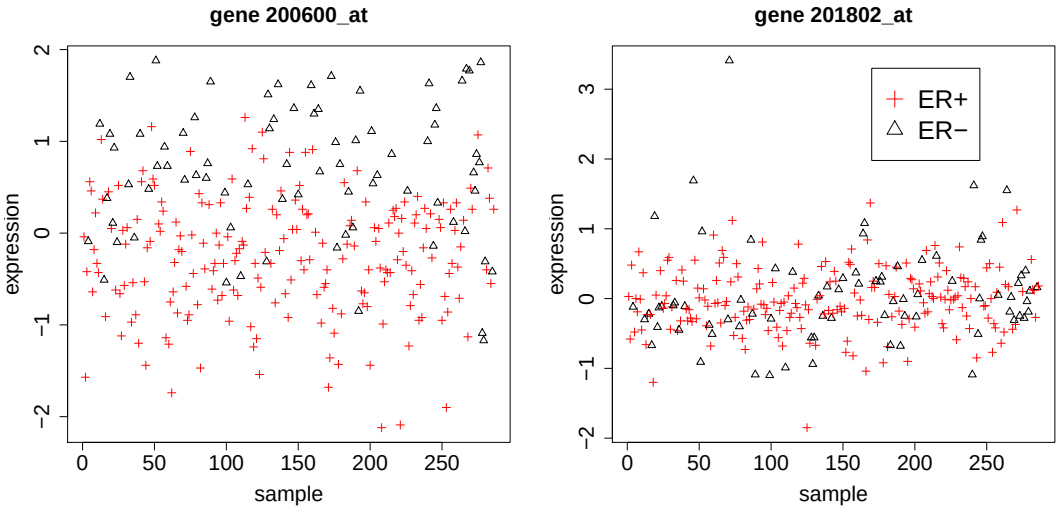

Figure 0.1: Comparing the plots of the expression values of genes which are potentially informative (left) or uninformative (right) for the estrogen-receptor level.

We considered two data sets, where the first one concerns the classification of certain breast cancer tumour subtypes, while the second one concerns the prediction of presence or absence of Irritable Bowel Syndrome.

Feature selection

Many methods of gene selection involve ranking the genes with regard to a certain test statistic. A higher test statistic value corresponds to a gene being seen as more important for differentiating between two or more distinct groups. A simple choice for such a test statistic is the two–sample–t–test, which compares the difference in mean expression values, standardized by a pooled variance estimate, to an appropriate quantile from the t–distribution. Testing and ranking genes individually in this manner ignores the fact that some groups of genes may contribute very similar information, and so inflates the number of required genes. This problem can be dealt with in two different ways. Firstly, rather than carrying out univariate tests marginally on every gene, one can apply multivariate variable selection approaches which try to identify combinations of genes which jointly optimize prediction accuracy — taking the standpoint that "the subset of the variables with the best univariate discrimination power is not necessar-

ily the *best subset of variables*" (Boulesteix et al, 2008). Techniques used in this field include Hotelling's two-sample T^2–statistic for groups of multiple variables as well as the top scoring pair or subset approaches (e.g. Yang and Naiman, 2014). However, these methods are quite complex and computationally intensive, and the publications dealing with them often of rather theoretical nature. Therefore, in this work, we went with the simpler concept of removing 'redundant' genes whose expression shows a high correlation with genes higher up in the list. The incorporation of this correlation threshold into the feature selection process makes the 'wrong' assumption of independent genes more reasonable. Our correlation threshold can be seen as a simplified version of that one suggested by Jäger et al. (2003), who excluded genes with high correlation to previously included genes, rather than genes higher up in the list.

Ordering the genes with regard to decreasing test statistic gives an order of importance. That is, each gene is considered in rank order but is selected only if it does not have a correlation higher than a certain threshold, b, with a gene of higher test statistic (or equivalently, lower p–value). The top k remaining genes can be chosen for use in classification.

The remaining question of *how many* genes to select is of high importance for diagnostic biomakers. This depends on the particular application, but usually a low budget of genes is preferential or necessary because of computational efficiency and the financial cost of processing a single gene expression. In practice a trade–off is required between the additional cost of including extra genes and the expected benefit for the increase in accuracy.

Classification

Once a list of genes has been selected, it is necessary to have a classifier which can take the expression values for these particular genes as inputs, and return a response indicating (say) the presence or absence of a disease as output. The *classification* problem is in this context equivalent to a *prediction* problem: Classification is the prediction of the diagnostic category of a tissue sample from its gene expression values given the availability of similar data from tissues in identified categories (Yeung and Bumgarner, 2003). There do exist many classifiers, including several discriminant techniques, logistic regression, nearest-neighbour classifiers, classification trees, random forests and support vector machines. In this article we will use a very simple classifier, Diagonal Linear Discriminant Analysis (DLDA), which postulates a multivariate Gaussian distribution with *equal and diagonal* covariance matrices for each classification group, and then places the

decision boundary such that, on this boundary, the data are just equally likely to belong to each of the two groups. DLDA is generally quicker and more efficient than Linear Discriminant Analysis since the latter requires the estimation of full covariance matrices which involves many more parameters. Our experience has shown that 'full' LDA, or even extensions such as quadratic discriminant analysis, do not improve the classification accuracy but rather add variability to the classification problem.

DLDA is particularly attractive in conjunction with a correlation threshold. From a mathematical point of view, the assumption of diagonal covariance matrices is incorrect as it would imply that expression values for all genes are independent of each other. By selecting genes that are less highly correlated with each other, we can reduce this invalidity and the effect that it has, as much as possible. Thus we have a further motive for removing highly correlated genes.

Validation

Once a classifier has been chosen, it is necessary to check its validity to see how accurately it predicts future subjects. To avoid overfitting, it is essential that the model is validated for other individuals than used to build the classifier. Furthermore, as highlighted by Boulesteix et al (2008), it is important that this validation procedure includes the gene filtering step, not only the classification step. One suitable method for this purpose is the cross-validation method, in which we split the data into a training set and a test set. The training set is used to select significant genes *and* build a classifier. For the test set one works out the proportion of the responses of data items which are correctly predicted by the classifier. This process is repeated many times so that lots of different combinations of training and test sets are considered, and the results are averaged to obtain an average accuracy rate. Throughout this article a training set comprising of 75% of the data is used for feature selection and classification, whilst the remaining 25% are used to test the classifier to obtain a proportion of correctly predicted cases. Our experiments repeated this procedure for 3000 different randomly sampled training and test sets and then averaged the resulting accuracies obtained over all of them.

Breast Cancer

Firstly we consider a microarray data set obtained from $n = 286$ lymph-node-negative primary breast cancer patients (Wang et al. 2005), with expression values available for $p = 17816$ genes. We consider two possible groupings in the context of this data:

- GROUP: non–aggressive (A) vs aggressive (B) cancers;
- ER: estrogen–receptor positive (ER+) vs. negative (ER-).

We begin our analysis with the inclusion of $k = 5$ genes and then increase this number in step sizes of 5, each time allowing for correlation thresholds ranging from $b = 0.6$ to $b = 1.0$. The average cross–validated accuracies over 3000 runs are shown in Figure 0.2, for the GROUP (top) and ER (middle) classification problems, respectively. We can see that, for both classification problems for this dataset, the classification accuracies start to level off from about $n = 35$ genes on, and settle at prediction accuracies slightly above 62% and 89%, respectively. Lowering the correlation threshold below 1 does not have a tremendous impact, though it is clear that it leads to a general improvement in accuracy for the GROUP variable, which however does not turn out to be significant after rigorous statistical testing. For the ER variable, the correlation threshold enables an increase of almost one accuracy point for a small number (≤ 15) of selected genes, which is indeed statistically significant as can be shown through Analysis of Variance. Informally, this is clear by considering the standard errors of the accuracies to the right hand side of Figure 0.2: These are of magnitude 3.5, so the standard errors of the *mean* accuracies are of magnitude $3.5/\sqrt{3000} \approx 0.06$, which is much smaller than the distance between the green and the black curve, for instance. The particularly poor performance for $b = 0.6$ and higher number of genes should be noted — using a very low correlation threshold can lead to the inclusion of noisy genes, which contribute little information to the classification problem. Our accuracy rates of close to 90% for the ER classification is in line with the rates obtained in other studies (Roepman et al, 2009). We are not able to compare our results with Wang et al (2005), who used a larger gene signature of size 76 to solve a different problem.

Irritable Bowel Syndrome

Irritable Bowel Syndrome, IBS, is a prevalent disorder affecting between 10% and 20% of people in the Western world (Aerssens et al., 2008). It is characterised by recurrent abdominal pain and an increased frequency to

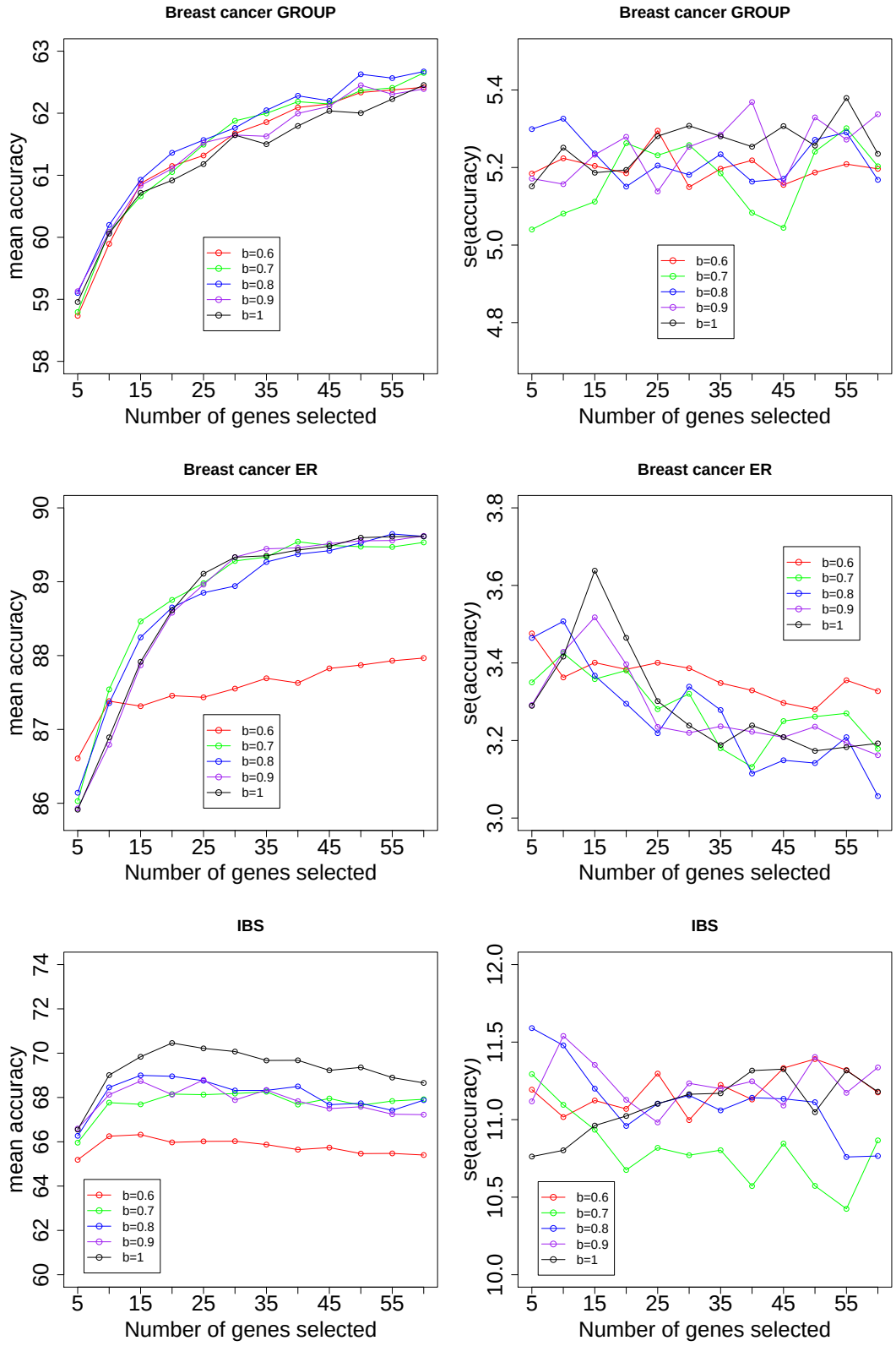

Figure 0.2: Cross–validated mean accuracy rates (left) and associated standard errors (right) for Breast Cancer data (top, middle) and Irritable Bowel Syndrome data (bottom).

need to empty the bowels. The data are given as a pre–regularised set of gene expression readings for 21212 genes from cells for each of a cohort of 34 IBS patients and 24 healthy controls.

Means and standard errors of accuracies for 3000 cross-validated runs of filtering and classification are shown in Figure 0.2 (bottom). Unlike for the previous data set, the mean accuracy curves take now clearly identifiable maxima when 15–20 genes are selected. The correlation threshold does appear to lower the accuracies in this example, though it should be noted that it still has the positive effect of shifting the maxima towards lower numbers of genes. The cross–validated accuracies of approximately 70% are comparable to those found in Aerssens et al (2008) using a complex combination of techniques.

Conclusion

For the prediction of absence or presence of Irritable Bowel Syndrome, as well as for breast cancer subtype classification, we have used a simple trilogy of basic statistical methods (t–test, correlation threshold, DLDA) to achieve accuracy rates which appear comparable to those provided in the literature using more complex methods.

A particular focus of this work has been the investigation of the correlation threshold. Yeung and Bumgarner (2003) considered a similar notion of a correlation threshold in connection with a shrinkage threshold for removing genes to increase the feature stability. Sensible choices for the correlation threshold are $0.7 \leq b \leq 1$. We found that threshold values $b < 1$ are particularly beneficial if the target number of genes is small, say 5-15, in the sense that it either increases the accuracy, or shifts the accuracy maximum to the left (see the left column of Figure 0.2). Both the 'optimal' choice of the threshold and the number of features for a particular microarray dataset is specific to the dataset, and could be selected through an inner cross–validation loop within a nested–loop cross–validation procedure (Göhlmann & Talloen, 2009).

In conclusion, this article has shown that, at least for the examples under study, very good classification rates can be achieved using simple statistical methods and relatively small–sized gene signatures. The accuracy curves presented in this article appear smoother compared to those found in the literature. The application of a correlation threshold does, in some cases, lead to an improvement in accuracy, but even in situations where it doesn't, it has other conceptual advantages which may still justify its use. The methods can in principle be extended to classification problems involving

more than two groups though we did not find the correlation threshold to be beneficial in this case.

Acknowledgements

This work was supported by an EPSRC vacation bursary as well as IMA Small Grant SGS27/14. The authors wish to thank Janssen Pharmaceutica N. V., Beerse, Belgium, for providing the IBS data.

References

1. Golub T.R. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286, 531–37.
2. Boulesteix A.-L. (2006). Reader's reaction to "Dimension reduction for classification with gene expression microarray Data" by Dai et al (2006), *Statistical Applications in Genetics and Molecular Biology*, 5, 1–7.
3. Boulesteix A.-L. et al. (2008). Evaluating microarray-based classifiers: an overview, *Cancer Informatics* 6, 77–97.
4. Yang S. and Naiman D.Q. (2014). Multiclass cancer classification based on gene expression comparison, *Statistical Applications in Genetics and Molecular Biology*, 13, 477–496.
5. Jäger J. et al. (2003). Improved gene selection for classification of microarrays, *Proceedings from Pacific Symposium on Biocomputing*, 2003, Lihue, Hawaii, USA.
6. Yeung, K.Y. and Bumgarner, R.E. (2003). Multiclass classification of microarray data with repeated measurements: application to cancer, *Genome Biology*, 4, R83.
7. Wang Y. et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *Lancet* 365, 671–679.
8. Roepman P. et al. (2009). Microarray–based determination of estrogen receptor, progesteron receptor, and HER2 receptor status in breast cancer, *Clinical Cancer Research*, 15, 7003–7011.
9. Aerssens J. et al. (2008). Alterations in mucosal immunity identified in the colon of patients with irritable bowel syndrome, *Clinical Gastroenterology and Hepatology*, 6, 194–205
10. Göhlmann H. and Talloen, W. (2009). *Gene Expression Studies Using Affymetrix Microarrays*, Boca Raton, Chapman & Hall, CRC.