Quality of Feature Selection Based on Microarray Gene Expression Data

Henryk Maciejewski

Institute of Computer Engineering, Control and Robotics,
Wroclaw University of Technology,
ul. Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
Henryk.Maciejewski@pwr.wroc.pl

Abstract. This paper is devoted to the problem of feature selection for class prediction based on results of DNA microarray experiments. A method is presented to objectively compare quality of feature sets obtained using different gene-ranking methods. The quality of feature sets is expressed in terms of predictive performance of classification models built using these features. A comparative study is performed involving means comparison, fold difference and rank-test (Wilcoxon statistic) methods. The study shows that best performance can be obtained using the rank-test approach. It is also shown that the means comparison method can be significantly improved by also taking into account fold-change information. Performance of such mixed methods of feature selection can surpass performance of rank-test methods.

1 Introduction

Genome-wide expression profiling using DNA microarray or similar technologies has become an important tool for research communities in academia and industry. Microarray gene expression studies are designed to obtain insights into yet unknown gene functions/interactions, investigate gene-related disease mechanisms or observe relationships between gene profiles and some factors (such as some risk factor or response to some therapy). Microarray studies motivated development of specific data analysis methods, broadly categorized into *class discovery*, *class comparison* and *class prediction* [12]. Class discovery aims to discover groups of co-regulated genes across the samples tested, or, alternatively, to discover groups of samples similar in terms of their gene expression profile, thus discovering new disease taxonomies [2]. The purpose of *class comparison* studies is to identify genes with most different expression profiles across the classes compared. This identifies groups of genes whose expression is significantly related to the classes and which possibly account for the difference between the classes compared. The purpose of *class prediction* is to build a predictive model for determination of the class membership of samples based on their gene expression profiles. Application areas of this seem very promising not only in research, but in medical diagnosis or prediction of response to treatment. Although US Food and Drug Administration recently approved the first microarray chip to help

M. Bubak et al. (Eds.): ICCS 2008, Part III, LNCS 5103, pp. 140–147, 2008.

doctors administer patient dosages of drugs that are metabolized differently by cytochrome P450 enzyme variant, more wide-spread application of microarrays in clinical or regulatory applications requires that several issues related to accuracy and reproducibility of microarray results as well as analysis of microarray data are resolved.

One step on the way to bringing microarrays to clinical applications was recently made by the Microarray Quality Control (MAQC) Consortium [6], [11], who through a comprehensive experimental study showed that microarrays have grown robust enough to produce data that is reproducible and comparable across different microarray platforms and laboratories. As another conclusion from their research, MAQC recommends using fold-change ratio as a measure of gene ranking for the purpose of identification of differently expressed genes in microarray studies. This recommendation is drawn from the observation that this method of gene selection yields best reproducibility of results across different microarray platforms. MAQC also advices against using means-comparison methods, such as the t-test, for gene ranking. This has opened recent discussion related to validity of gene selection methods. E.g., Klebanov et al. [8] argues that fold-change cannot be regarded superior to the t-test, as it realizes smaller power, and in general, reproducibility of results should not be used as an indication about the adequacy of the gene selection method.

This discussion has motivated the study reported in this paper. Our work addresses another perspective (or criterion) for choosing the *right* method of gene selection. It is proposed to judge the quality of a gene selection method by looking at the information value of genes returned by the method and regarded as features for class prediction. In other words, a gene selection method will be deemed superior if it tends to produce features yielding best predictive performance of sample classifiers. In the following section, an approach is explained to arrive at the quality of features obtained using a given gene selection method. Next, this approach is used to compare three widely used gene selection methods (means-comparison, fold-change and a method based on the statistical rank test). The comparative study is based on the data originally published by Golub et al. [5] and Alon et al. [1]. Finally, it is shown that combining different feature selection methods (e.g., enhancing means-comparison method by also including the fold-change criterion) can result in increased performance of class prediction.

2 Quality of Features from Different Gene Selection Methods

Here we present an approach to express quality of a gene selection method in terms of predictive performance of a classifier using the genes regarded as features. In Sect. 2.1 we discuss the challenges that need to be overcome to build and properly validate performance of a sample classifier based on gene expressions. This should be seen as the motivation for the procedure detailed in Sect. 2.2 for judging the quality of gene selection.

2.1 Classification in High Dimensionality Data

Building a class prediction model based on microarray study data is a challenging task due to very high dimensionality of data obtained and relatively small number of samples available. Microarray studies typically produce data vectors with dimensionality $d \sim 10^3$ to 10^4 (the number of genes observed in one DNA chip), while the number of samples tested is at most $n \sim 10^2$. In other words, microarray studies define an ill-formulated problem of classification, where $d \gg n$, while standard approaches to predictive modeling require that $d \ll n$. This implies that significant dimensionality reduction is required. Another challenge related to the analysis of such data concerns proper estimation of expected predictive performance of the classifier for *new* data. Considering the small number of samples available for model building *and* testing requires that a properly tailored data-reuse approach is used. How these issues will be approached in the procedure in Sect. 2.2 is now developed.

The following notation will be used to represent results of a microarray experiment. Let $(x_i, y_i), i = 1, 2, ..., n$ denote data vectors related to the n samples tested in a microarray study, where $x_i \in R^d$ represents gene expressions from the sample i and $y_i \in C = \{c_1, c_2\}$ denotes the class membership associated with the sample i. For the problem of *class prediction* it is assumed the class membership y_i for each x_i is known prior to analysis. Only the binary classification case is considered here, where $y_i \in \{c_1, c_2\}$; this can be extended to the multi class problem by using ANOVA based metrics for gene ranking (such as the F-statistic).

The problem of class prediction is formally stated in the statistical decision theory [7] as looking for a prediction model $f : R^d \mapsto C$, minimizing the expected prediction error:

$$EPE = E\left[L\left(Y, f\left(X\right)\right)\right] \tag{1}$$

where the *loss function* L, used to penalize misclassification events can be defined as e.g.,

$$L\left(Y, f\left(X\right)\right) = \begin{cases} 1 & \text{for } \ Y \neq f(X) \\ 0 & \text{for } \ Y = f(X) \end{cases} . \tag{2}$$

Since in class prediction studies only samples $(x_i, y_i), i = 1, 2, ..., n$ of the random variables X and Y are known, *empirical risk* defined as $\frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f\left(x_i\right)\right)$ is used to estimate the EPE (1). It should be noted that this must be computed based only on the data points *not used* for the purpose of building the model f, and not used in the stage on feature (gene) selection. Considering the small number of data points from a microarray study, EPE can be estimated by repeatedly training and testing the model for different data splits, where a subset of available data is used for feature selection and model building, leaving the remaining (smaller) part for estimation of EPE. This leads to a cross-validation estimate of EPE defined as [7]:

$$CV = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f^{-i}\left(x_i\right)\right) \tag{3}$$

where f^{-i} is the classifier fitted to data with the sample x_i removed. This version of cross-validation realizes smaller bias (at the price of higher variance) as compared with the procedures leaving more samples for testing [7].

It should be noted that estimating EPE based on samples used for feature selection leads to overoptimistic estimates of predictive performance of classifiers, as pointed out in [14], and is not an uncommon error in literature.

Another issue concerning class prediction based on microarray data is related to setting the *right dimensionality* of the feature set. Here the well known fact should be considered [7] pertaining to binary classification that in d dimensions $d + 1$ points can be always perfectly separated by a simplest linear classifier. This implies that for microarray data $((x_i, y_i), i = 1, 2, \dots, n$ where $d \gg n)$, one can always obtain perfect fit of the model to data, providing enough (i.e. $n - 1$) genes are selected. However, such models will not guarantee good predictive performance for new data, as they are prone to *overfitting* meaning small prediction error for training data with high prediction error for test (new) data [12]. This limits the number of genes that should be selected as features for class prediction to no more then the number of data points available in microarray data.

2.2 Quality of a Feature Selection Method

Considering the above mentioned challenges, the following procedure is proposed to arrive at the quality measures attributed to a given feature selection method. In the next chapter three specific feature selection methods are compared using this procedure.

1. Select the value d^* of dimensionality of the feature vector from the range $1..n - 2$.
2. Remove the sample (x_i, y_i) from the original data set (the remaining $n - 1$ samples will be referred to as the training data).
3. Using the training data select d^* genes ranked top by the gene selection procedure considered.
4. Reduce dimensionality of the vectors x by leaving only values of expression of the d^* genes selected (the vectors obtained will be denoted x').
5. Build a classification model denoted f^{-i} by fitting it to the $n - 1$ points x', obtained in the previous step.
6. Compute $e_i = L\left(y_i, f^{-i}(x'_i)\right)$.
7. Repeat Steps 2 through 6 for $i = 1, 2, \dots, n$.
8. Compute $CV_{d^*} = \frac{1}{n}\sum_{i=1}^{n} e_i$ (this estimates the EPE - see (3)).
9. Repeat Steps 1 through 8 for a grid of values d^* spaced evenly in the range $1..n - 2$, using approx. 10 values of d^*.
10. Plot the obtained relationship CV_{d^*} versus d^*.

It is proposed that the quality of a feature selection method be judged by observing the minimum values of CV_{d^*} obtained. Also, comparing the plots obtained in Step 10 for different feature selection methods gives an indication about the quality of features produced by competing methods over a range of different dimensionality models. This approach is used in the sample study shown in the following section.

3 Comparing Quality of Commonly Used Feature Selection Methods

Using the approach proposed in Sect. 2.2, the quality of three different methods commonly used for ranking genes is compared:

1. Gene ranking based on the Wilcoxon statistical rank-test,
2. Gene ranking based on the fold difference (used and recommended by Shi et al. [11]),
3. Gene ranking based on the signal to noise measure (which is an example of direct means comparison methods, such as the t-test).

Feature selection based on the first method requires that the Wilcoxon non-parametric group comparison test is performed independently for every gene. This gives a p-value indicating whether expression of this gene for the groups of samples of class c_1 and c_2 can be considered different. Gene selection using this method returns the set of top d^* (Step 4 in Sect. 2.2) genes ranked by increasing p-value of the test.

Feature selection using the fold difference measure requires that for every gene the ratio of mean expressions from samples of class c_1 and c_2 is computed. More specifically, if for a given gene, the mean value of gene expression from samples of class c_1 and c_2 is denoted μ_1 and μ_2, respectively, then the (log) fold difference measure is defined as:

$$fc = |\log(\mu_1) - \log(\mu_2)| \tag{4}$$

which produces high values if either of the means exceeds the other. Gene selection using this method returns the set of top genes ranked by decreasing value of fc.

Feature selection based on the signal to noise uses the measure defined as:

$$sn = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} \tag{5}$$

where σ_1 and σ_2 are standard deviations of expressions of a fixed gene for the samples of class c_1 and c_2, respectively. Gene selection returns the set of top genes ranked by decreasing value of sn.

Fig. 1 compares the EPE vs. model dimensionality for these three gene selection methods (marked in the plot by 'w', 'sn' and 'fc'). As a classifier we used in this study the multilayer perceptron (MLP) model with one hidden layer. We observe that the minimum value of EPE (i.e., the best predictive performance expected for new, independent samples) is realized for the Wilcoxon method, with 15 genes selected. It can be also observed that for the wide range of different dimensionality models (up to 50 features), Wilcoxon feature selection yields significantly fewer prediction errors then signal to noise or fold difference.

Fig. 1 also includes EPE for a model built using a *combined* method of feature selection (marked in the plot as 'snfc'). This approach includes the following steps:

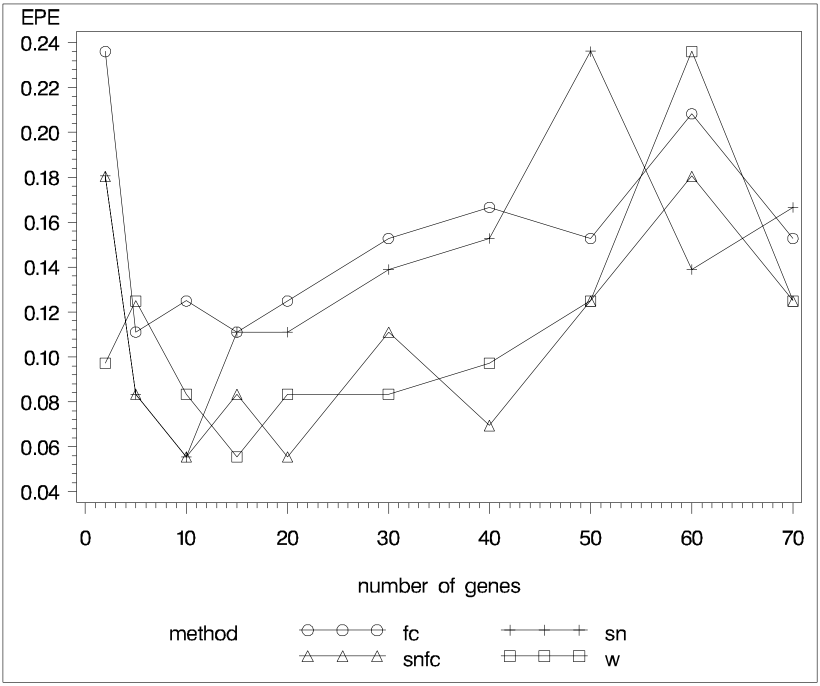

Fig. 1. Expected prediction error of neural network model for different methods of gene selection. (Notation: w=Wilcoxon test, fc=fold difference, sn=signal to noise, snfc=signal to noise with additional fold difference criterion).

1. For each gene, the fc and sn measures are computed according to (4) and (5).
2. Genes are ranked by decreasing values of sn.
3. The required number of top genes is returned, providing a gene realizes at least two-fold difference in expression (i.e., $fc \geq 1$, where in (4) we used the logarithm to the base 2).

The threshold of at least two-fold difference in expression was also used as the feature selection criterion in [11]. Interestingly, features returned from this combined model yield significantly better EPE then features from individual models ns and fc, with minimum values of EPE realized for 10-20 genes. This approach shows similar performance in terms of feature quality to the Wilcoxon method.

The same analysis repeated for a different classification model – logistic regression gives results depicted in Fig. 2.

Basically, the conclusions drawn from Fig. 1 regarding the quality of competing feature selection methods are confirmed: the Wilcoxon method realizes the best predictive performance (for 10-20 features), and the combined method ('snfc') leads to remarkable improvement in feature quality, making this method comparable with the Wilcoxon rank test.

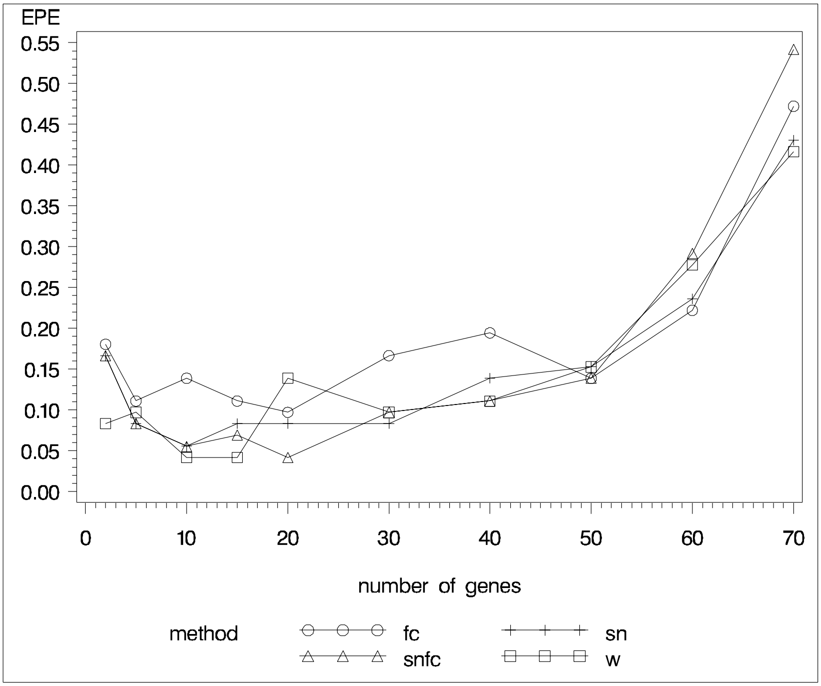

Fig. 2. Expected prediction error of logistic regression model for different methods of gene selection. (Notation: w=Wilcoxon test, fc=fold difference, sn=signal to noise, snfc=signal to noise with additional fold difference criterion).

Figs. 1 and 2 also illustrate model dimensionality related issues: too small a model dimensionality leads to poor prediction performance (due to the model being too simple), while too big a dimensionality leads to overfitting of the model. This compromise should be taken into consideration when setting the *right* dimensionality of a class prediction model.

Similar analysis repeated for the colon data set [1] basically confirms the conclusions drawn from the leukemia study. Again, the Wilcoxon method tends to produce the best predictive performance (the $EPE \approx 0.16$ for 20 genes, using the MLP classifier). Similar results were observed for 10 or 30 features obtained with the combined method.

4 Conclusions

We demonstrated that the quality of gene selection methods can be empirically compared by observing the performance of class prediction models built using features returned by these methods. To obtain a fair picture of the quality, such analysis should not be limited to one pre-fixed number of genes selected, it should rather be made for a representative collection of different dimensionality models, which allows to observe the size of feature vectors yielding good class prediction.

Using this approach, we showed that the Wilcoxon rank test is a superior gene selection method then fold-change or direct means comparison. However, significant improvement can be achieved if different gene selection criteria are used simultaneously. This suggests that feature selection for microarray class prediction probably should comprise information from several different criteria, thus increasing information contents of the feature set. This however requires further research.

Being able to quantitatively rank gene selection methods, as shown in this work, raises another interesting question to what extend the genes selected as best features for sample classification really account for the differences between classes. This open question requires further research in an interdisciplinary team.

References

1. Alon, U., et al.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745–6750 (1999)
2. Bittner, M., Meltzer, P., Chen, Y.: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406, 536–540 (2000)
3. Dudoit, S., Shaffer, J., Boldrick, J.: Multiple Hypothesis Testing in Microarray Experiments. UC Berkeley Division of Biostatistics Working Paper Series, Paper 110 (2002)
4. Ewens, W., Grant, G.: Statistical Methods in Bioinformatics. Springer, New York (2001)
5. Golub, T., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
6. Guo, L., et al.: Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nature Biotechnology 24, 1162–1169 (2006)
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer, New York (2002)
8. Klebanov, L., et al.: Statistical methods and microarray data. Nature Biotechnology 25, 25–26 (2007)
9. Maciejewski, H.: Adaptive selection of feature set dimensionality for classification of DNA microarray samples. In: Computer recognition systems CORES, Springer Advances in Soft Computing, Springer, Heidelberg (2007)
10. Maciejewski, H., Konarski, L.: Building a predictive model from data in high dimensions with application to analysis of microarray experiments. In: DepCoS - RELCOMEX. IEEE Computer Society Press, Los Alamitos (2007)
11. MAQC Consortium [Shi L. et al.]: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology 24, 1151–1161 (2006)
12. Markowetz, F., Spang, R.: Molecular diagnosis. Classification, Model Selection and Performance Evaluation, Methods Inf. Med. 44, 438–443 (2005)
13. Polanski, A., Kimmel, M.: Bioinformatics. Springer, Heidelberg (2007)
14. Simon, R., et al.: Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. Journal of the National Cancer Institute 95, 14–18 (2003)