On univariate selection methods in gene expression datasets

Carmen Lai, Marcel J.T. Reinders, Lodewyk Wessels

Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
email:{c.lai}@ewi.tudelft.nl

Keywords: feature selection, greedy search, small sample size problem, microarray datasets

Abstract

Gene expression microarrays enable the measurement of the activity levels of thousands of genes on a single microscope slide. Analysis of these data sets is a recent and challenging manifestation of the small sample size problem in pattern recognition. The primary objective is to build a classifier which classifies a new sample as accurately as possible into one of the diagnostic categories, for example tumor/normal tissue. A secondary objective is to find a small number of genes, i.e. a signature, which the diagnostic classifier employs as input, and which consequently carries the information relevant for the diagnostic task. This process of identifying the genes relevant to the classification task is known as feature selection. A widely used approach evaluates the informativeness of each single gene, based on a criterion such as a signal to noise ratio (SNR), and then employs this univariate ranking to guide the search for an optimal gene set. In this paper we focus on the evaluation of this approach. To achieve this goal we introduce an artificial model to generate an experimental dataset. With this model we investigate the effects of a number of parameters on the classification performance and the quality of the selected gene set. We illustrate the risks and the weaknesses of the univariate selection methods.

1 Introduction

Micro array data has opened new possibilities and challenges in genetic studies. Up to now the studies and the diagnoses have been based on a number of different tests often relying on human experience and subjectivity. In some cases this is still not enough to make a sure statement. A basic assumption of the genetic studies is that the genome carries all the information about the characteristics and the development of an organism. Therefore an understanding of the genome would bring more objectivity in the problem under study.

Gene expression microarrays enable the measurement of the activity levels of thousands of genes on a single microscope slide. An important application of this technology is the prediction of disease state of a patient based on a signature of the gene activities. Such a diagnostic signature is typically derived from a dataset consisting of the gene expression measurements of a series of patients. Since typically hundreds of patients and thousands of gene activities are measured, analysis of these data sets is a recent and challenging manifestation of the small sample size problem in pattern recognition. The primary objective is to build a classifier which classifies a new sample as accurately as possible into one of the diagnostic categories, for example tumor/normal tissue, or benign/malignant. A secondary objective, which is a by-product of the primary objective is to find a small number of genes, i.e. a signature, which the diagnostic classifier employs as input, and which consequently carries the information relevant for the diagnostic task. This process of identifying the genes relevant to the classification task is known as feature selection. Given the small sample size problem, sophisticated search strategies are prone to overtraining. In addition, due to the large number of features, these approaches are also particularly computationally intensive.

Feature selection can either be based on backward or forward selection of genes. The backward selection starts from a complete set of genes re-

moving redundant or uninformative features according to a selection criterion. Examples of this approach use Support vector machines [5, 17, 14] as classifiers.

The forward feature selection is also used often [2]. It starts with one gene and iteratively searches the informative genes between all available ones. These are added into the growing gene subset until a certain performance criterion or a size limit is reached (greedy search). A widely used approach within the area of molecular classification, evaluates the informativeness of each single gene, based on a criterion such as a signal to noise ratio (SNR), and then employs this univariate ranking to guide the search for an optimal gene set [6, 3, 8, 10]. In this paper our attention is on the methodological problems of the rank based forward search. Our aim is not to provide a new strategy to retrieve the relevant genes. Instead we focus on the univariate ranking of genes and investigate how it is affected by the small sample size problem. To achieve this goal we introduce an artificial model to generate an experimental dataset. With this model we investigate the effects of a number of parameters on the classification performance and the quality of the selected gene set. We illustrate the risks and the weaknesses of the univariate selection methods.

The paper is organized as follows. The univariate ranking approach and the artificial dataset are described is Section 2. The experimental setup and the results are discussed in Section 3. Finally, the conclusions are given in Section 4.

2 Univariate ranking approach

The micro array datasets have a huge numbers of features (genes), compared to the samples (patients). In order to find a signature of significant genes, a selection procedure is needed. It should retrieve the genes that are required for accurate classification, i.e. the one informative with respect to the problem under study.

We focus on the evaluation of a gene selection method described in Figure 1. The data is split into two parts X and Y, used respectively for training and testing purposes. Based on a criterion, the informativeness of each gene in the training set X is evaluated individually. The genes are ranked accordingly (preselection step), i.e. from the most to the least informative. A classifier is than trained, starting with the best gene, and is tested on the same genes in the independent test set Y. The procedure is repeated including the genes one by one, in the order established by ranking, until all of them are used. Each time a

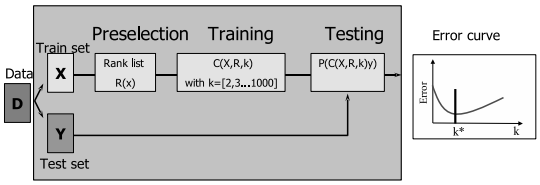


Figure 1: Schema of the evaluation method.

feature is added, the error is computed on the test set. As a result, we can plot the error of the classifier as a function of the number of genes used for classification purposes (error curve). The curve in the right part of Figure 1 illustrates the expected result. Typically this curve will show that a small number of genes gives large error rate, due to insufficient information. The (initial) addition of relevant genes lowers the error, reaching a minimum. Further addition of genes however degrades the classifier performance. Minimizing the error provides a selection of relevant genes (k* in Figure 1), i.e. a signature. Note that the estimation of the information carried by a gene is done in the preselection block, where only the training set X is used. In this way the bias due to the use of testing genes in the training procedure is avoided [2].

Our aim in this paper is not to choose an informative subset of genes but to estimate the discrimination capability of gene subsets of different sizes. Therefore the method must be repeated a number of times. Due to the small sample size a suitable way is given by the cross-validation procedure. The data is divided into N parts. In the first fold the algorithm is trained on $N-1$ splits and tested on the remaining one. The procedure is iterated N times, rotating the splits used for training and testing in such a way that in each fold a different split is used for testing. The cross-validation procedure ensures that the training and test sets are independent. Figure 1 can be seen as the description of one fold of a two-fold cross validation. In the second fold, the same procedure would be repeated using Y as training set and X as testing one. By averaging the two error curves, we obtain an estimate of classifier performance using gene subsets of different sizes.

In order to thoroughly evaluate the described

approach, we need a controlled environment in which the relevant features are known. Therefore we propose the use of an artificial dataset, described in the following section. The dataset is made by informative and non-informative features. Since we know which are the relevant ones, we can repeat the procedure in Figure 1 using the *true* order of the features, and thus excluding the ranking step. Our motivation is to use the resulting error curves to judge the retrieval performance of an investigated approach including ranking. Another advantage of using an artificial dataset is that the small sample size problem can be easily studied compared to a case when a large number of samples is available.

2.1 The artificial dataset

In order to investigate the challenges posed by typical microarray data, we generate a comparable artificial problem. Our goal is not to simulate the real data set, as proposed by [7, 4, 12], but to have a controlled environment with roughly the same complexity, without having to deal with other sources of variations. To study the effect of the small number of training samples on the univariate feature filtering procedure, we generate a dataset for which feature filtering (e.g. ranking) would be able to retrieve the correct feature sets, giving enough data. We extend the simple model proposed by [9, 16, 13]. Our dataset is a matrix $M \times N$ with M samples and N features. Each feature vector is sampled from the following two-class conditional densities:

$$p(X|\omega_1) \backsim N(\mu(i), 1) \qquad p(X|\omega_2) \backsim N(-\mu(i), 1) \tag{1}$$

where $\mu(i)$, is a function of the feature indicator i according to the following:

$$\mu(i) = \begin{cases} -\frac{\mu_0}{I} i + \mu_0, & \text{if } 1 \leqslant i \leqslant I; \\ 0, & \text{if } I < i \leqslant N. \end{cases} \tag{2}$$

The most informative features are the ones with the smallest index value i. The distance between the means of both normal distributions, i.e. the class separation, linearly decreases from $2\mu(1)$ for the first feature towards 0 at I^{th} feature value i. Therefore, the informativeness of a feature is defined by its index value i. All features with an index i larger than I are not informative, since the two normal distributions overlap completely.

Note that each feature vector is generated independently, therefore the univariate ranking is a proper evaluation criterion (provided that there are enough training samples).

3 Experiments

First the set up of the experiments is presented, with particular emphasis on the parameters chosen for the artificial model. Later the evaluation of the gene preselection based on individual ranking is discussed.

3.1 Experimental set up

As described in Figure 1, the first step in the training procedure is to estimate the informativeness of the genes individually. Several criteria may serve this purpose, such as Pearson correlation, Fisher criterion, or signal-to-noise ratio (SNR). Since for each feature both classes are normally distributed, we chose the SNR because it captures the difference between two normal distribution. Besides the SNR is simple and popular [6, 8]. The signal to noise ratio is defined as follows:

$$SNR = \frac{|m_1 - m_2|}{\sqrt{s_1^2 + s_2^2}}, \tag{3}$$

where m_1 and m_2 are the estimated means of the two classes and s_1 and s_2 are the estimates of the respective standard deviations. The higher the SNR the more informative the corresponding genes.

Since the artificial dataset is generated from independent features which have normal-based class conditional densities with equal variance (i.e. $\sum_{\omega_1} = \sum_{\omega_2} = I$), the nearest mean classifier is an optimal Bayes classifier. We therefore chose to use the nearest mean classifier as classifier in Figure 1. Thus we may expect that the classifier will not hamper the evaluation procedure. Additionally, the nearest mean classifier is a stable classifier that behaves favorably in a small sample size problems [15].

As discussed in Section 2 the cross-validation is a suitable procedure to estimate the classification error. For classification purpose it is important to have a training set as big as possible. The number of folds in cross-validation determines the sizes of the train and test set. On the other hand we would like to have a test set large enough to be representative of the data. As a compromise we choose to use 10 fold cross-validation. This choice is also suggested by Ambroise *et all* [2] and Kohavi *et all* [11]. In order to avoid the possible biases caused by a single draw of the dataset, we repeat the experiment 10 times, using as datasets 10 different draws from the same model.

3.1.1 Setup of the artificial dataset

Let us now discuss more in detail the setup of the artificial dataset. First we shortly describe the

characteristics of real microarray datasets that are relevant for our study.

The *Leukemia* dataset [6] consists of two types of leukemia: acute mieloid leukemia (AML), present with 25 samples, and the acute lymphoblastic leukemia (ALL), with 47 samples. The feature space is reduced to 3571 genes, based on the protocol described in [**?**]. The *Colon* dataset [1] is composed of 40 normal healthy samples and 22 tumor samples in a 1908 feature space. Both datasets are public and widely used in the literature. The *Breast cancer* dataset [8] consists of 145 lymph node negative breast carcinomas, 99 from patients that didn't present metastasis within five years and 46 from patients that where affected by a second tumor within five years. The size of the feature space is 4919.

For generating the artificial dataset, the first choice that has to be made concerns the dimensionality. We want to simulate real complexity conditions, such as the small sample size. Therefore the number of the samples M is set to 100 which is comparable to real datasets. The number of features is set to 1000 (N) mainly for computational reasons. Additionally a large dataset is generated that will be used to estimate the true error of the build classifiers. The dimensions of this dataset is set to 1000 samples $\times$1000 features.

The gene expression datasets are often unbalanced due to the different availability of the samples, e.g. the benign tissue is more common than the malignant one. In the above mentioned datasets, one class is roughly 30% of the number of samples (the other 70%). Therefore we choose to preserve this unbalance also in the artificial dataset.

Two more parameters remain to be set: the starting value μ_0, i.e. the class separability of the best features, and index I, that limits the amount of the informative features in the data. As previously described, we chose the SNR as a criterion to rank the features to be used in the classification procedure. In order to set the parameters μ_0 and I, we first investigate how their values affect the calculated SNR values for the individual genes.

Figure 2 shows the ranked value of the SNR for all features for different datasets. Note that the ranking of the SNR values is necessary in order to have monotonically decreasing (readable) curves. In the three sub plots the artificial datasets have 0.15, 0.25, 0.35 as values for $|\mu_0|$. As reference we use an uninformative dataset ($I = 0$) of size 100x1000, the *Breast cancer*, the *Colon*, and the *Leukemia* datasets. Since the number of features of the real datasets are larger than 1000, the plots show a uniform resampling of their SNR values. The artificial datasets are generated with a varying number of relevant features, i.e. $I = 100, 250, 500, 1000$ respectively.

The SNR curves of the *Colon* and *Leukemia* datasets are much higher than that of the *Breast cancer* dataset. This suggests that more information is present in this first two datasets, reflecting the complexity of each dataset aims at a different problem. The *Colon* dataset, for example, was collected to distinguish between tumor and normal tissues. The *Breast cancer* dataset, on the other hand, aims at distinguishing between different clinical developments of the same type of cancer, which is a far more challenging diagnostic problem.

The figure shows, as one might expect, that the smaller the μ_0 and the smaller the I, the closer the SNR curves approach the one of the uninformative dataset.

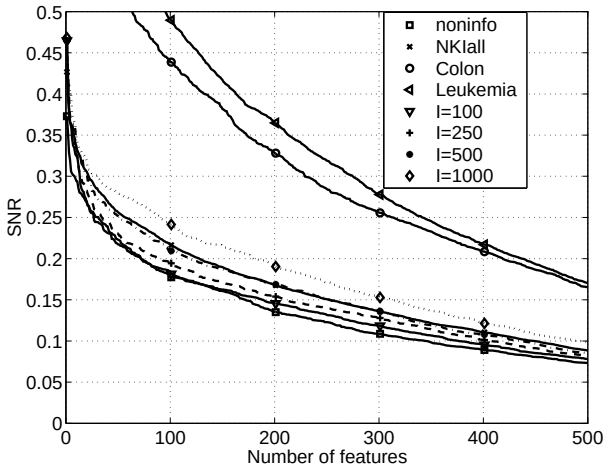
Note that while the uninformative dataset has the lowest curve, some of the uninformative features reach a value, comparable to the informative features of the other datasets. This is related to the small sample effect. The more uninformative features are tested with a limited small training set, the higher the probability of having a large SNR due to sampling effects. This misleads the evaluation of the feature relevance.

We want to focus on difficult conditions, in which the information appears not to be easily retrievable, as in the *Breast cancer* case. Figure 2 shows that the setup that results in a dataset whose SNR curve matches the one of the *Breast cancer* are the settings $|\mu_0| = 0.25$ and $I = 250$.

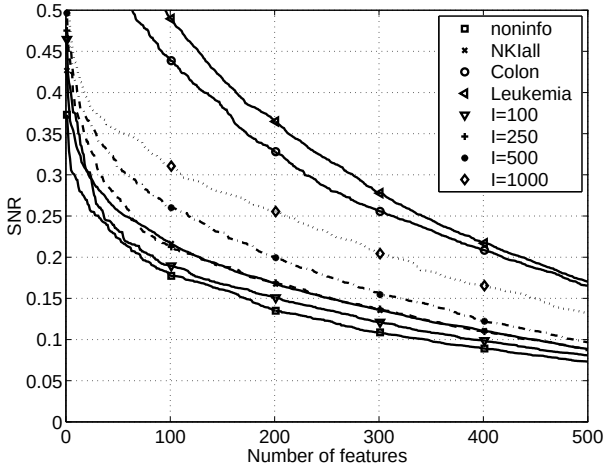
3.2 Experimental results

In this section, we evaluate the methodology described in Section 2 using the artificial dataset. Figure 3 summarizes the results. The classification error is calculated in two ways. On one hand the classification error, averaged across the 10 folds and the 10 artificially generated datasets, is plotted as a function of the number of features used to train the classifier. On the other hand, the classification error is calculated by testing the classifier on the large independent test set of 1000 samples. Due to the larger sample size, this test set allows the estimation of the true error of the classifier.

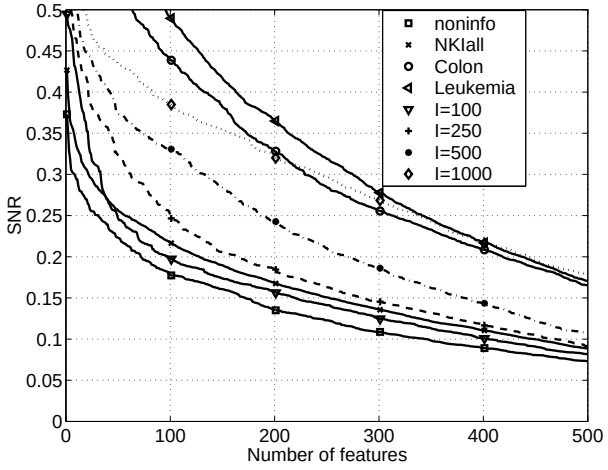
Evaluation of the methodology. In the artificial dataset the original index of the features corresponds to their amount of informativeness. We can test the efficiency of the proposed methodology, directly using the original feature order, thus excluding the preselection step in Figure 1. The



(a) Artificial datasets with $\mu_0 = 0.15$



(b) Artificial datasets with $\mu_0 = 0.25$



(c) Artificial datasets with $\mu_0 = 0.35$

Figure 2: Calculated SNR for feature j where $SNR(f_j) < SNR(f_{j+1})$.

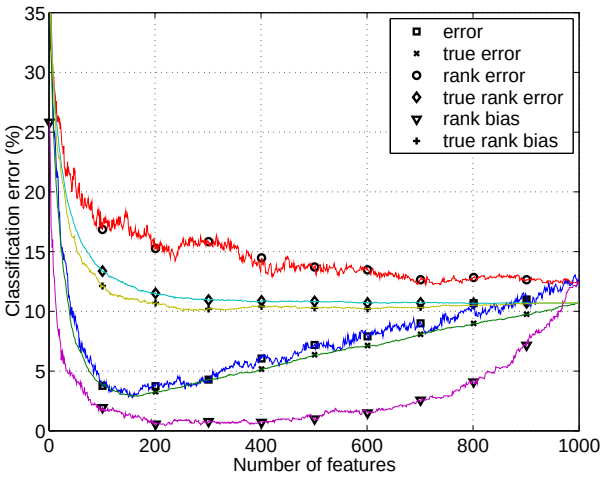


Figure 3: Average classification error as a function of the number of features used to train the classifier.

results while testing on the cross-validation test set and on the large test set, are plotted with the lines with x and square respectively. Since the cross-validation error estimate of the method is closed to the *true* error, we can conclude that the cross-validation methodology is a good evaluation tool. The 10 fold cross-validation however has a larger variation on the error, as can be observed by the presence of the local minima.

Necessity of a correct test procedure. As pointed out by Ambroise *et al.* [2], a bias is introduced if the estimation of feature relevance is made using all the data, since the test set in not independent anymore. The line called *rank bias* (see line with triangles in Figure 3) shows the cross-validation error while the features of the complete dataset were first ordered according to SNR and then the cross-validation procedure was run. This error is apparently very low, while the true error computed on the larger dataset (line with plus) is much higher instead. From this we conclude that all the steps taken to derive a classifier, i.e. gene ranking, selection and classifier training, must be performed only on the training set, keeping an independent test set aside.

Evaluation of the ranking approach. The line with circles in Figure 3, which we call *the rank error*, represents the cross-validation error while applying the method described in Figure 1. When comparing this error curve with the true error (i.e. features in original order) one can conclude that the ranking according to SNR is not able to identify the relevant features. Due to the small sample size, uninformative features have high SNR and

consequently a high rank. The size of the feature set should increase to include the necessary informative features, but including more features also degrades the classifier. As a results no minimum is detected anymore. Clearly this does not fulfil the original target of deriving a small good signature. We can conclude that the estimate of the gene informativeness is very poor. This is due to the small sample effect since if we apply the same methodology to the large dataset of 10 000 samples (experiment not shown), the rank and true errors overlap. Clearly the SNR estimates are accurate now.

4 Conclusions and future work

We discussed the effects of the estimate of individual gene relevance on the gene selection procedure in a classifier design system. For this purpose we generated a controlled environment. On one hand we simulated real conditions, e.g. by choosing comparable dimensions of the dataset, and forcing the artificial dataset to have a SNR behavior similar to the one of the *Breast cancer* dataset. On the other hand, we tuned the data model in order to fit the ranking criterion and the classifier. For that reason, a simple artificial model was generated based on independent class conditional normal distributions, and the nearest mean classifier was chosen as classifier, since it approximates well the distributions.

We have also emphasized the importance of a correct test procedure. Each step taken in the training phase, i.e. gene preselection and the derivation of a classifier itself, must be performed only on the training set keeping an independent test set aside to estimate the performance of the classifier.

Our work illustrates that the ranking step introduces the largest degradation in the classifier performance, and that this stems from the small sample size limitations: relevant and irrelevant genes cannot be distinguished from each other on an individual basis. As a consequences relevance of the gene subsets selected accordingly is diminished. Future work will be performed on extending the model and studying other gene selection approaches.

References

[1] U Alon, N Barkai, D A Notterman, K Gish, S Ybarra, D Mack, and A J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, 1999.

[2] C Ambroise and McLachlan G J. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, 99(10):6562–6566, 2002.

[3] A Ben-Dor, L Bruhn, N Friedman, I Nachman, M Schummer, and Z Yakhini. Tissue classification with gene expression profiles. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 54–64. ACM Press, 2000.

[4] A Chilingaryan, N Gevorgyan, A Vardanyan, D Jones, and A Szabo. A multivariate approach for selecting sets of differentially expressed genes. *Mathematical Biosciences*, 2002.

[5] TS Furey, N Christianini, N Duffy, DW Bednarski, M Schummer, and D Hauessler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[6] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[7] W Huber, A von Heydebreck, H Sueltmann, A Poustka, and M Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2003.

[8] Laura J van 't Veer, H Dai, M J van de Vijver, D He Yudong, A A M Hart, M Mao, H L Peterse, K van der Kooy, M J Marton, A T Witteveen, G J Schreiber, R M Kerkhoven, C Roberts, P S Linsley, R Bernards, and S H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[9] A Jain and D Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEEPAMI*, 1997.

[10] J Khan, JS Wei, M Ringner, LH Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, CR Antonescu, C Peterson, and PS Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 2001.

[11] Ron Kohavi. The power of decision tables. In *Proceedings of the European Conference on Machine Learning*.

[12] MA Newton, CM Kendziorski, CS Richmond, FR Blattner, and KW Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 2001.

[13] P Pudil, J Novovicova, and J Kittler. Floating search methods in feature selection. *PRL*, 1994.

[14] A Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research, Special Issue on Variable Selection*, 3:1357–1370, 2003.

[15] Marina Skurichina. *Stabilizing weak classifiers*. PhD thesis, Delft,Technical University, 2001.

[16] G V Trunk. A problem af dimensionality: a simple example. *IEEEPAMI*, 1979.

[17] J Weston, S Mukherjee, O Chapelle, M Pontil, T Poggio, and V Vapnik. Feature selection for svms. In *Proc of NIPS*, pages 668–674, 2000.