Multivariate gene selection: does it help?

Carmen Lai[1] Marcel Reinders[1]
Lodewyk Wessels[1,2]
[1]Information and Communication Theory Group, Delft University of Technology, Delft, The Netherlands
[2] The Netherlands Cancer Institute, Amsterdam, The Netherlands
{c.lai; m.j.t.reinders; l.f.a.wessels} @ewi.tudelft.nl

Abstract

When building predictors of disease state based on gene expression data, gene selection is performed in order to achieve a good performance and to identify a relevant subset of genes. Although several gene selection algorithms have been proposed, a fair comparison of the available results is very problematic. This mainly stems from two factors. First, the results are often biased, since the test set is in one way or another involved in training the predictor, resulting in optimistically biased performance estimates. Second, the published results are often based on a small number of relatively simple datasets. Therefore, no generally applicable conclusions can be drawn. We therefore adopted an unbiased protocol to perform a fair comparison of state of the art multivariate and univariate gene selection techniques, in combination with a range of classifiers. Our conclusions are based on seven gene expression datasets, across many cancer types. Surprisingly, we could not detect any significant improvement of multivariate feature selection techniques over univariate approaches. We speculate on the possible causes of this finding, ranging from the small sample size problem to the particular nature of the multivariate gene dependencies.

1. Methods and Results

Selection algorithms A set of genes (L) ordered according to their relevance is provided by a gene selection algorithm. We implemented the following gene selection algorithms: 1)*univariate search technique* (U), which estimates the importance of each gene individually, based on the signal-to-noise-ratio (SNR) [6] or t-test[5] as criteria; 2) the *base pair* (BP) approach, which evaluates the relevance of pairs of genes *et al.* [4]; 3) a greedy *forward search* (F) *et al.* [4]; 4) *Recursive Feature Elimination* (RFE) [7], which is an iterative backward selection approach, that employs the Support Vector Machine (SVM) to estimate the feature weights; and 5) the *Liknon* classifier [3], which simultaneously performs relevant gene identification and classification.

Evaluation framework In order to avoid any bias, we perform the selection of the genes and the evaluation of the classification performance in two independent steps, as proposed in [12] and illustrated in Figure 1. In the training phase the optimal gene size $k^\star$ is estimated in a 10-fold cross-validation scheme. The selection algorithm is then applied to the whole training set D_1 in order to obtain the best $k^\star$ genes, i.e. the optimal gene-set, and the final classifier is trained. Finally, the performance of the gene selection strategies together with the corresponding classifiers (Nearest Mean (NMC), Fisher (FLD) or Support Vector (SVM) classifiers) is estimated using a 10-fold cross-validation procedure.

The experimental results are summarized in the Table 1.

2. Conclusions

We have performed a comparison of state of the art multivariate and univariate gene selection algorithms across several cancer diagnostic problems. Surprisingly, we could not detect any significant improvement when employing multivariate gene selection techniques. The univariate selection approach with a simple classifier outperforms or is comparable with the results of the other methods. Therefore information about the gene's correlation, if present, cannot be detected by the statistical analysis of gene expression data. We argue that this is due to the very limited sample size, which prevents the detection of complex patterns in the data.

References

[1] U. Alon et al.. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the Na-*

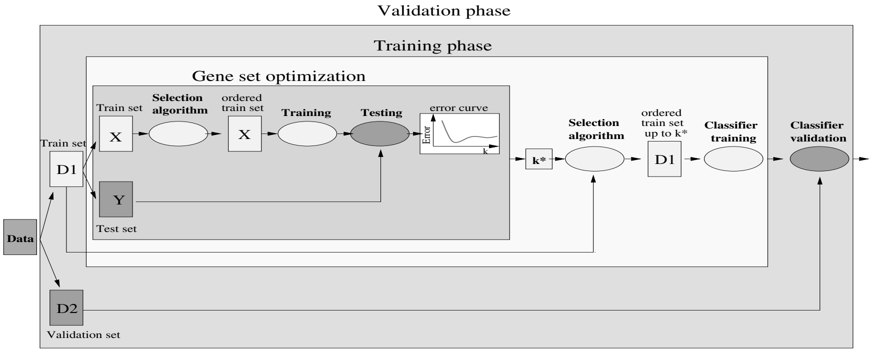

Figure 1. Gene selection and classification framework employed to evaluate the different approaches.

Table 1. The mean and the standard deviation of the 10-fold cross-validation error (in percentage) for the different approaches and datasets employed in the study.

Method	CNS [9]	Colon [1]	DLBCL [2]	HNSSC [10]	Leukemia [6]	Breast [8]	Prostate [11]
gene selection	mean ± std	mean ± std	mean ± std	mean ± std	mean ± std	mean ± std	mean ± std
U, SNR, NMC	30.4 ± 6.5	12.9 ± 4.2	2.5 ± 2.5	21.2 ± 7.1	4.8 ± 2.7	33.0 ± 3.4	9.7 ± 4.2
U, SNR, FLD	42.5 ± 7.3	19.2 ± 5.9	15.8 ± 6.4	33.3 ± 6.6	8.0 ± 3.2	29.9 ± 3.6	10.0 ± 3.0
U, t-test, NMC	32.5 ± 4.9	12.5 ± 4.2	2.5 ± 2.5	21.2 ± 7.3	4.8 ± 2.7	33.5 ± 3.8	10.8 ± 3.4
U, t-test, FLD	35.8 ± 6.5	11.7 ± 3.5	15.8 ± 6.4	36.2 ± 6.2	12.0 ± 4.2	32.6 ± 3.0	8.0 ± 2.5
BP greedy, FLD	43.8 ± 6.2	12.9 ± 3.8	10.0 ± 4.3	36.2 ± 7.0	11.6 ± 3.6	35.8 ± 2.3	9.8 ± 3.3
F, FLD	47.9 ± 5.1	15.4 ± 4.1	10.8 ± 3.7	45.4 ± 8.5	10.2 ± 4.2	35.4 ± 4.2	14.0 ± 3.4
RFE, FLD	34.2 ± 5.0	22.9 ± 4.4	16.7 ± 5.3	35.0 ± 6.3	3.5 ± 2.6	33.8 ± 3.5	10.0 ± 2.6
RFE, Svm	35.4 ± 5.0	22.1 ± 3.5	15.8 ± 5.2	35.4 ± 7.2	4.5 ± 2.6	32.6 ± 3.2	8.0 ± 2.9
Liknon	32.9 ± 6.1	13.3 ± 4.2	13.3 ± 5.3	37.5 ± 7.4	11.8 ± 4.0	34.5 ± 5.2	10.8 ± 3.7
no gene selection	mean ± std	mean ± std	mean ± std	mean ± std	mean ± std	mean ± std	mean ± std
NMC	42.1 ± 5.5	17.9 ± 3.3	6.7 ± 3.5	29.2 ± 7.2	3.5 ± 2.6	36.7 ± 3.2	33.7 ± 3.9
FLD	32.9 ± 6.3	21.7 ± 3.7	14.2 ± 5.4	32.5 ± 6.6	4.5 ± 2.6	35.8 ± 4.1	8.0 ± 2.5
SVM	35.4 ± 7.0	22.1 ± 3.5	9.2 ± 3.8	29.6 ± 5.7	3.5 ± 2.6	34.3 ± 4.2	8.0 ± 2.9

tional Accademy of Siences of the United States of America, 96(12):6745–6750, 1999.

[2] A.A. Alizadeh et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] C. Bhattacharyya et al. Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data. *Signal Processing*, 83(4), 2003.

[4] T. Bo and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome biology*, 3, 2002.

[5] S. Dudoit and J. Fridlyand. *Statistical analysis of gene expression microarray data*, chapter 3. 2003.

[6] T. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[7] I. Guyon, J. Weston, and S. Barnhill. Gene selection for cancer classification using support vector machines. *Machine Learning*, (46):389–422, 2002.

[8] L. J. van 't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[9] S. Pomeroy et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[10] L. Roepman et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nature Genetics*, 37:182–186, 2005.

[11] D. Singh, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell.*, 1:203–209, 2002.

[12] L. Wessels et al.. A protocol for building and evaluating predictors of disease state based on microarray data. *accepted in Bioinformatics*.