The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
Copyright © 2007 ORSC & APORC, pp. 213–221

Large Validation of Anti-learnable Signature in Classification of Response to Chemoradiotherapy in Esophageal Adenocarcinoma Patients

Adam Kowalczyk[1,2] Danielle M. Greenawalt[3,4]
Justin Bedo[1,2] Cuong Duong[3] Garvesh Raskutti[1]
Robert J. S. Thomas[3,5] Wayne A. Phillips[3,4,5]

[1] Life Sciences, NICTA and Department of Electrical and Electronic Engineering,
The University of Melbourne, Parkville, Victoria 3010, Australia
[2] The Research School of Information Sciences and Engineering,
The Australian National University, Canberra, Australia
Divisions of [3]Surgical Oncology and [4]Research, Peter MacCallum Cancer Centre,
St. Andrew's Place, East Melbourne, Victoria 3002, Australia
[5] Department of Surgery, The University of Melbourne, St. Vincent's Hospital, Fitzroy,
Victoria 3065, Australia

Abstract **Motivations.** We present a supervised learning analysis of a cDNA microarray dataset designed for prediction of response to chemoradiotherapy (CRT) in esophageal adenocarcinoma (AC) patients. The dataset has unusual properties: the whole range of supervised learning techniques generates predictive models which classify independent test samples systematically below the accuracy of random guessing (hence the name anti-learning shortened to AL). As this is systematic and can be detected easily by additional cross-validation on the training data, even straightforward, ad hoc reversal of the classifier decision provides a good prediction of patient's response. The main question we tackle here is to what extent this unusual behaviour can be attributed to the real biological processes rather than noise in the data.

Results. The label permutation test shows that the observed AL behaviour has significance level above 99% for a whole range of t-test pre-filtered gene markers. Furthermore, the analysis of multiple random data sets shows that although AL behaviour can be observed on randomly generated data sets, the systematic AL behaviour displayed by AC dataset for a whole range of subsets of pre-filtered gene markers has not been matched even once among one thousand generated data sets. This points towards a specific AL signature in the AC data. We pursue this line further by generating a synthetic dataset, based on a straightforward zero-sum game, matching closely the AL characteristics of the AC dataset. This also makes a formal link to perfect AL studied theoretically in previous publications.

Conclusions. We conclude that non-standard properties of the AC dataset are most likely a signature of a hidden process which is observed indirectly on the level of gene expression. Although this hypothetical mechanism is unknown, it seems possible to achieve our main objective:

the generation of reliable predictors of important response to CRT using a number of supervised learning techniques.

1 Introduction

The ability to detect esophageal cancer (EC) patients with poor response to chemoradiotherapy (CRT) will save them from significant toxicity of the treatment and potential complications with no obvious advantage. In a study evaluating a feasibility of development of such a test using microarray technology, a cohort of forty-six EC patients has been recruited (Greenawalt et al., 2007; Duong et al., 2007). From each patient a tumour biopsy was taken prior to treatment, then profiled using gene expression cDNA microarray with 10,500 probes. The patients were administered the CRT treatment, then followed up and finally classified by experts as "good" or "bad" responders. These cancer patients split into two histological subtypes: 21 squamous cell carcinoma (SCC) and 25 adenocarcinoma (AC). The initial analysis of the dataset shows unequivocally that both histological subtypes display dramatically different behaviour. While SCC dataset has allowed development of predictive models with performance up to 87% as measured by the popular Area under Receiver Operating Characteristic (AROC) in cross validation tests, the models for AC data performed systematically below the level of random guessing (Duong et al., 2007).

In this paper we exclusively concentrate on the analysis of this unusual behaviour of AC dataset. This dataset is linearly separable; perfect classification is possible even with an appropriately selected single feature, so linear algorithms such as SVM, ridge regression or shrunken centroid should have no problem with classifying it. In fact, they seem to have no problem with classifying the training set for a whole range of features selected using the standard t-test. However, the generated models are significantly in error on the independent test set. This extends to a number of non-linear algorithms such as decision trees, multilayer neural, radial basis SVM etc. (see Kowalczyk, 2007). For some of these algorithms, the departure below random guessing level of performance is minimal, but for others such as SVM and centroid it is significant. In the latter case the systematic predictions with high error rates allow a reliable, although non-standard, determination of the right labels. An ad-hoc solution is simply to reverse the decision for the classifier. However, more principled methods can be used as well and will be presented elsewhere (Kowalczyk, 2007). Our main aim here is the presentation of an initial analysis and validation that the observed AL properties of AC dataset are "real" rather than caused by noise.

Interestingly, in these experiments the significant AL is achieved for a significant number of genes ($\approx$1000) selected using the standard t-test. This seems to imply that the signature of CRT response is distributed among many genes. This on its own is not so unusual. For instance, the results of Ein-Dor et al. (2005) show that the signature of survival in breast cancer dataset of van 't Veer et al. (2002) is also distributed and equivalent in performance predictors could be developed using many subsets of genes.

2 Materials and Methods

2.1 Adenocarcinoma Dataset

The dataset contains 25 patients, classified as 14 good and 11 bad responders to CRT. It is a subset of data analysed in Greenawalt et al., 2007 and Duong et al., 2007. The gene expressions were made using cDNA microarray with 10,500 probes representing 9,389 unique cDNAs. Raw array data and protocols are available at `http://www.ebi.ac.uk/arrayexpress`.

2.2 Supervised learning experiment setup

In this paper we use a standardised experimental setup of 5-fold cross validation repeated 20 times, thus the reported averages are for 100 independent tests. We have always used stratified splits of data into 5 folds, preserving the proportions of the both classes in the fold as much as practical.

Throughout this paper for all classifiers excluding PAM (which includes feature selection during training), we have used the t-test. Top scoring features were selected in terms of the absolute value of the statistic $(\mu_+ - \mu_-)(s_+^2/n_+ - s_-^2/n_-)^{-\frac{1}{2}}$ where μ_y and s_y are the mean and variance of the feature, and n_y is the number of instances for the class label y. Feature selection was always applied only to the training set, and repeated each time the training set was changed.

2.3 Performance metrics

We use the Area under ROC (AROC), the plot of the True Positive versus False Negative error rates as our main performance metric. Additionally, we also use Accuracy defined as the average of the True Positive and the True Negative rates. Both metrics are insensitive to the class distribution in the test set. For both, the value of 0.5 represents the performance of trivial classifiers, such as random guessing or allocation of all example to one class; value 1 will be allocated to the perfect classifier; and value 0 to the perfectly wrong one.

2.4 Basic supervised learning algorithms

Our default algorithm is the *centroid*. This is a simple classification technique which produces a linear classifier and requires no tuning parameters. It consists in computation of the means of the two classes in the training set, and then classifying data according to the value of the projection onto the direction of the vector set by these means. Typically the additional additive constant (the bias) is set such that the center of the segment between two means receives score 0. It has been shown to perform very well on microarray datasets (van 't Veer et al., 2002; Ein-Dor et al., 2005) and is known to be the "high regularisation" limit of support vector machine and ridge regression (Bedo et al., 2006). Thus it is indicative of the performance of those classifiers under training involving heavy regularisation.

We have also extensively used *k nearest neighbours* (*k*-NN), always with Euclidean distance in the appropriate feature space.

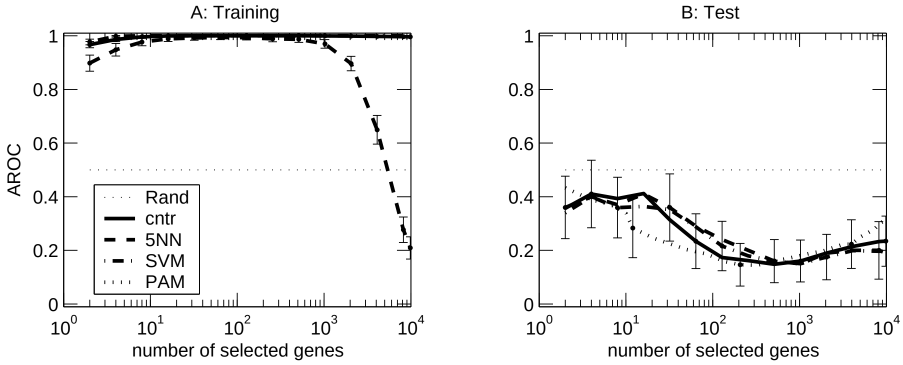

Figure 1: Selected results for AC datasets. We plot average of Area under ROC (AROC) in function of number of features selected with standard deviations marked as error bars. The 5-fold cross validation repeated 20 times was applied, so averages are over 100 trials, and in each trial 80% of data was used for training and the remaining 20% for the independent test. We have used the following classifiers: Centroid (Cntr), hard margin support vector machine (SVM), shrunken centroid (PAM) and 5-nearest neighbours (5-NN). For all classifiers, excluding PAM, the genes were selected using t-test applied to the training subset only. Note that PAM has built-in feature selection.

For the *shrunken centroid*, which is a modified version of the centroid classifier popular in analysis of microarray data, we have used public domain R implementation within the `pamr` package (Tibshirani et al., 2003).

Finally, the *support vector machine (SVM)* here means the hard margin case (Vapnik, 1998). The soft margin SVMs have performed between two extremes, the centroid and the hard margin case (data not shown).

3 Results

3.1 Classification of AC dataset

Figure 1 introduces the problem of AL as the systematic misclassification of the independent test set. Note that with exception of 5-NN classifiers, the algorithms have classified almost perfectly the training set, hence produced classifiers vastly inconsistent between the training and test sets. The 5-NN classifiers, which requires no training or adaptation to the training set other then indirectly, through the feature selection, behave differently. For a larger number of features, when the bias introduced by the selection of features “well” correlated with training set labels fades away, it starts to perform on the training sets similarly as in the independent test.

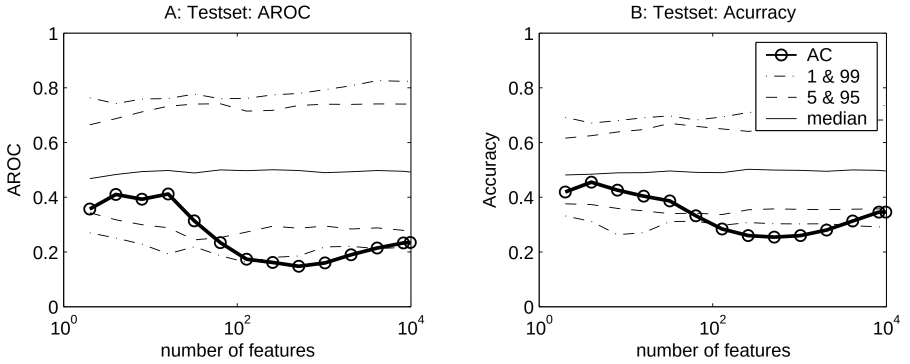

Figure 2: The label permutation test. The background distribution represented by percentile lines was created from the results of classifying the AC dataset with labels permuted 1000 times. Each dataset was evaluated by the centroid algorithm using 5 fold cross validation repeated 20 times. The solid “AC” line represents results of such an analysis for the original AC dataset (un-permuted).

4 Permutation Tests

The results of a standard label permutation test are shown in Figure 2. We observe that for larger numbers of features, the AL performance of AC dataset was outside the 1% percentile ($p < .01$). However, even stronger result holds: none of the randomly permuted datasets scored AROC and Accuracy systematically lower that those for AC dataset.

Figure 3 shows results of a test of a null hypothesis that AC dataset contains no signal and is just a product of noise. To test this we have generated 1000 datasets of the same size as AC data using standard normal distribution. These datasets were subsequently classified using the centroid algorithm combined with t-test feature selection (our standard protocol). We observe that for the larger number of selected features the AL in AC data is stronger than in 1% of the lowest values for the random sets. As before, the performance of none of the generated dataset was systematically lower than AC. This is even more pronounced for Figure 3B. However, this should be taken with caution as the relatively low spread in accuracy for random datasets in the range of larger number of features is due to the systematically inadequate bias of the classifiers (to see this compare Figure 3A). We have also observed a similar pattern of behaviour for other classifiers (e.g. SVM) and other sampling distributions (e.g. uniform distribution).

Figures 2 & 3 provide strong indication that the AC dataset must contain a systematic anti-learnable signal and its strong AL properties are unlikely to be just a result of noise.

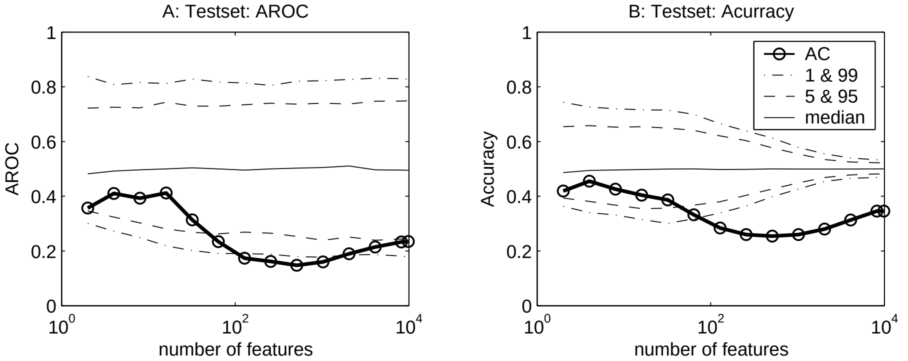

Figure 3: Classification of 1000 random datasets. The background distribution represented by percentile lines was created from the results of classification of 1000 datasets of the size of AC datasets but with measurements drawn from the standard normal distribution. Each dataset was evaluated by the centroid algorithm using our standard 5 fold cross validation repeated 20 times. The solid "AC" line represents results of such an analysis for the "un-permuted" AC dataset.

5 Classification of synthetic dataset

In this section we show that the AL properties of the AC data set can be matched by a synthetic dataset, one incorporating the "anti-learnable signal". It shows that AL can be a reflection of a simple phenomenon, such as a competition for limited resources or mathematically a zero-sum game, which could be obscured in an indirect observation. The model is an adaptation of the theoretical models of "perfect AL" studied in Kowalczyk-Chappelle, 2005 and Kowalczyk-Smola, 2005. Here we outline the gist of the model, leaving aside the formal details and formal analysis (see Kowalczyk, 2007).

We constructed a model dataset in which a large number of species, *N*, share a particular environment which can only support a fixed total number of them. Thus if the numbers of any species increase (or decrease) beyond the normal level, the number of other species must decrease (or increase) accordingly. We use the name CS (for "Competing Species") to refer to this dataset. Our model incorporates two types of pathologies forming the two label classes: Class A, where numbers of one particular species increase significantly and the remaining species adjust by a uniform decrease, and Class B, where a particular species decline significantly and all remaining species uniformly increase their numbers. In the model these changes are not observed directly, but via a pooled signature of all members of community, where each species has its own multidimensional signature. For example, one might imagine a scenario where there are *N*-species of bacteria living in a fermenter, and the signature as mixed waste secreted to the common environment by the members of the community and measured in the output from the fermenter. For our synthetic CS dataset we have set the output dimensions to be 10,000 so as to closely match the dimension of the AC data. A formal implementation of this model follows.

The Formal Model: We denote by n_β the number of members of βs species, $\beta = 1, ..., N$. A pathology described above with a particular change D_α in the size of the αth species is defined by the equation

$$n_\beta := n^*_\beta + \Delta n_\beta := \begin{cases} n^*_\alpha + D_\alpha, & \text{for } \beta = \alpha; \\ n^*_\beta - \frac{D_\alpha}{N_1 - 1}, & \text{for } 1 \leq \beta \leq N_1, \beta \neq \alpha; \\ n^*_\beta, & \text{otherwise}, \end{cases} \tag{1}$$

where $n^*_1 \ldots, ..., n^*_N$ are "normal" or default state values and N_1 is a number of subspecies which are "coupled", $1 < N_1 < N$ (the remaining $N - N_1$ species are unaffected by any changes within this sub-community of N_1 species). The pathologies of Class A correspond to the sign $y = sgn(\Delta_\alpha) = +1$, while for Class B we have $y = -1$. Note that the zero-sum game condition, $\sum_\beta \Delta n_\beta = 0$, holds.

As mentioned before, the sizes n_β cannot be observed directly (otherwise, aggressive feature selection would produce a strongly learnable model, which is not what was observed for the AC-data). Instead we observe a d-dimensional pooled signature vector

$$\vec{x} = \sum_{\beta=1}^{N} \frac{n_\beta}{\sum_{\beta'=1}^{N} n_{\beta'}} \vec{s}_\beta - \sum_{\beta=1}^{N} \frac{n^*_\beta}{\sum_{\beta'=1}^{N} n^*_{\beta'}} \vec{s}_\beta + \vec{\varepsilon} \tag{2}$$

$$= C_1 + C_2 \sum_{\beta=1}^{N} (\vec{s}_\beta - \vec{s}^*) \Delta n_\beta + \vec{\varepsilon}, \tag{3}$$

where $\vec{s}_\beta$ is a d-dimensional signature vector of the βth species, $\vec{s}^* := \sum_{\beta=1}^{N} \vec{s}_\beta / N$, $\vec{\varepsilon}$ is random noise and C_1 & C_2 are constants.

In our experiments we have used $d = 10,000$, $N = 300$, $N_1 = 200$, $\Delta_\alpha = \text{ const} = 1$. We have generated 14 and 11 instances of Class A and B, respectively. The entries of signature vectors $\vec{s}_\beta$ were drawn from the normal distribution $N(0,1)$ and the noise $\vec{\varepsilon}$ had normal distribution $N(0,0.1)$; these parameters are not very critical. □

Figure 4 shows the results of classification on this synthetic dataset: note a striking similarity to the plots for the AC dataset in Figure 1.

6 Discussion

The model of competing species is studied more formally in forthcoming paper (Kowalczyk, 2007). In particular, that paper shows formal result on anti-learning and introduces some algorithms which can seamlessly classify both, the anti-learnable and ordinary "learnable" datsets.

The size of the variance of the background distributions in Figure 3 comes as a bit of a surprise. In particular, it implies that AL-datasets are relatively abundant. What this research brings is evidence that it appears in natural datasets. It is rather surprising that not much has been reported to date on the subject. Our guess is that

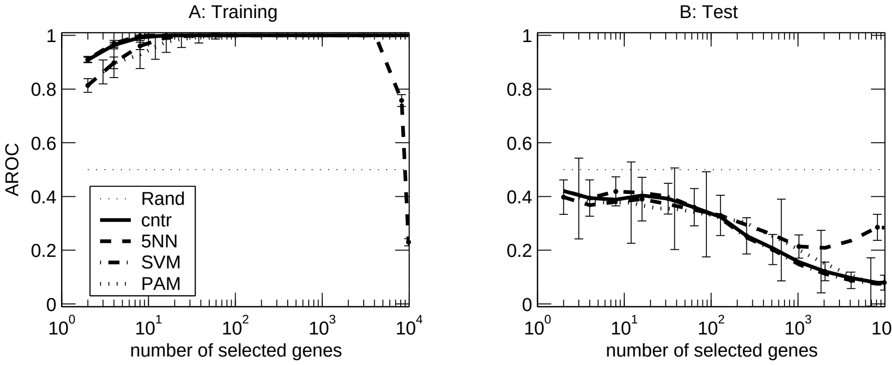

Figure 4: Anti-learnable pattern in the classification of synthetic CS datasets with t-test selected gene subsets. The settings of this experiment are as Figure 1; observe that plots are very close to their counterparts in Figure 1.

AL datasets if encountered are misunderstood, perceived as abnormalities not worth further investigation or reporting, and so are forgotten. At this stage we are aware of the existence of a few other natural datasets with well pronounced anti-learning: some in the area of cancer genomics, others in other areas of bio-medical research, including heart ECG. These results will be reported elsewhere.

The symmetry of the distributions in Figure 2 are consistent with the so called “no free lunch theorems” (see Wolpert, 1996), which roughly state that the averaged test performance of any algorithm over all possible datasets is on the level of random guessing. However, what is of prime interest in Figure 3 is not that their means are ≈ 0.5, but the size of the distribution tails in relation to the observed performance for the AC dataset.

Our results show that the AL-signature in AC dataset is spread over hundreds of genes. This makes it more challenging to design independent wet-lab experiments for validation of our findings. Nevertheless, our bio-informatics driven investigation of the AC data indicates that such an investigation is necessary. Coming to grips with AL-behaviour is a pre-requisite for any efficient design of such follow-up experiments. Further study of a range of synthetic models of anti-learning is required in order to identify the potential biological mechanisms involved and in order to form working hypothesis along which the wet-lab experiments could be designed.

Acknowledgements

This project is funded in part by grant No. 350414 from the NHMRC of Australia. AK, JB and GR acknowledge permission of National ICT Australia to publish this paper. National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Council through Backing Australia's Ability and the ICT Center of Excellence program. This work was supported in part by the IST Programme of the European

Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] J. Bedo, C. Sanderson, and A. Kowalczyk. An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics. Australian Conf. on Artificial Intelligence, pages 170–180, 2006.

[2] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK, 2000.

[3] C.Duong, D.M. Greenawalt, A. Kowalczyk, M. Ciavarella, G. Raskutti, W. Murray, W.A. Phillips and R.J.S. Thomas, Pre-treatment gene expression profiles can be used to predict response to neoadjuvant chemoradiotherapy in esophageal cancer, submitted.

[4] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, pages 171-8, 2005.

[5] D. Greenawalt, C. Duong, G. Smyth, M. Ciavarella, N. Thompson, T. Tiang, W. Murray, R. Thomas, and W. Phillips. Gene Expression Profiling of Esophageal Cancer: Comparative analysis of Barrett's, Adenocarcinoma and Squamous Cell Carcinoma . Int. J. Cancer: 120, 1914-1921 (2007)

[6] A. Kowalczyk and O. Chapelle. An analysis of the anti-learning phenomenon for the class symmetric polyhedron. In S. Jain, H. U. Simon, and E. Tomita, editors, Proceedings of the 16th International Conference on Algorithmic Learning Theory. Springer, 2005.

[7] A. Kowalczyk and A. Smola. Conditions for antilearning. Technical Report HPL-2003-97(R.1), NICTA, Canberra, June 2005.

[8] A. Kowalczyk. Classification of anti-learnable biological and synthetic data, The 18th European Conference on Machine Learning and the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2007, to appear.

[9] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. Stat Sci, 18:104-117, 2003.

[10] L.J. van 't Veer and et al. Gene expression profiling predicts clinical outcome of breast cancer . Nature, 415: 530-6, 2002.

[11] V. Vapnik. Statistical Learning Theory. John Wiley and Sons, New York, 1998.

[12] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7):1341-90, 1996.