Classification of Anti-learnable Biological and Synthetic Data

Adam Kowalczyk

National ICT Australia and
Department of Electrical & Electronic Engineering,
The University of Melbourne,
Parkville, Vic. 3010, Australia

Abstract. We demonstrate a binary classification problem in which standard supervised learning algorithms such as linear and kernel SVM, naive Bayes, ridge regression, k-nearest neighbors, shrunken centroid, multilayer perceptron and decision trees perform in an unusual way. On certain data sets they classify a randomly sampled training subset nearly perfectly, but systematically perform worse than random guessing on cases unseen in training. We demonstrate this phenomenon in classification of a natural data set of cancer genomics microarrays using cross-validation test. Additionally, we generate a range of synthetic datasets, the outcomes of 0-sum games, for which we analyse this phenomenon in the i.i.d. setting.

Furthermore, we propose and evaluate a remedy that yields promising results for classifying such data as well as normal datasets. We simply transform the classifier scores by an additional 1-dimensional linear transformation developed, for instance, to maximize classification accuracy of the outputs of an internal cross-validation on the training set. We also discuss the relevance to other fields such as learning theory, boosting, regularization, sample bias and application of kernels.

1 Introduction

Anti-learning is a non-standard phenomenon involving both dataset and classification algorithms, which has been encountered in some important biological classification tasks. In specific binary classification tasks, a range of standard supervised learning algorithms, such as linear and kernel SVM, naive Bayes, ridge regression, k-nearest neighbors, shrunken centroid, multilayer perceptron and decision trees behave in an unusual way. While they easily learn to classify a randomly sampled training subset nearly perfectly, they *systematically and significantly* perform worse than random guessing if tested on cases unseen in training. Thus reversing the classifier scores can deliver an accurate predictor, far more accurate than the original machine. In such a case we say that the dataset is *anti-learnable* by our classifier.

In this paper we shall demonstrate this phenomenon on a natural data set, a cancer genomics microarray dataset generated for classification of response

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 176–187, 2007.

to treatment in esophageal cancer [1,2] and a synthetic dataset introduced in this paper. For the esophageal dataset, the previous analysis points towards a biological origin of a specific anti-learnable signal in the data [3], although the exact nature of such a mechanism is unclear at this stage.

We start with analysis of synthetic anti-learnable datasets, which are the outcomes of specific 0-sum games (Section 2). For such data we can use analytical methods and prove that anti-learning is the logical consequence of a specific configuration of dataset (Section 2.3). Further, for such datasets we can generate samples of arbitrary size, hence we can use the standard independently identically distributed (*i.i.d.*) setting rather than cross-validation for experimental evaluation. This leads to generation of non-conventional learning curves (Section 2.1) showing a continuum of behavior modes, starting with anti-learning for small size samples to classical, consistent generalization (asymptotic) bounds in the large size training samples limit.

In order to build a bridge to the esophageal data, we have used our synthetic model to generate a dataset of similar size (50 samples, split evenly between two labels and each represented by 10000 features). Then we classified the original and synthetic datasets using a range of classifiers combined with aggressive feature selection (t-test filter). We observe a strong similarity between learning curves for both datasets, which indirectly supports the hypothesis of deterministic origins of an "anti-learnable signal" in the esophageal dataset.

Independently, we demonstrate and evaluate some algorithms, which can successfully classify such non-standard data as well as standard datasets seamlessly. The idea here is to combine the classifiers scores with a module trained to "interpret" them accordingly. In our case, this is exclusively a simple 1-dimensional linear transformation developed to maximize a chosen objective function of the scores from internal cross-validation on the training set (Section 2.2). We show analytically and empirically, that such modified algorithms can perform well in Sections 2.1, 2.3 and 3.

Links to related research. There is a direct link to previous papers on perfect anti-learning [4,5] as follows. A specific cases of WL-game introduced in Section 2 (the magnitude $\mu \equiv$ const and single case per mode) generate the "class symmetric" kernel data studied in those papers. As mentioned before, the paper [3] studied significance of anti-learning in esophageal cancer dataset. A form of anti-learning is in KDD'02, Task 2 data: the anti-learning occurrs for standard SVM and persists for the aggressive feature selection [6,7]. Finally, the existence of anti-learning is compatible with predictions of "No Free Lunch Theorems" [8].

2 Anti-learnable Signature of a 0-Sum Game

An individual outcome of the game is represented by a d_0-dimensional *state vector* $\boldsymbol{s} = (s_1, ..., s_{d_0}) \in \mathbb{R}^{d_0}$, with each dimension corresponding to a "player". The players split into three groups: potential winners, indexed 1 to d_0^+; potential losers, index $d_0^+ + 1$ to $d_0^+ + d_0^-$; and remaining $d_0 - d_0^+ - d_0^- \geq 0$ neutral players.

The outcome is uniquely determined by two parameters, the *magnitude* $\mu_{\boldsymbol{s}} > 0$ and *mode*, $M_{\boldsymbol{s}} \in \{1, ..., d_0^+ + d_0^-\}$, which here is the index of the player, as follows:

$$s_i = \begin{cases} y_{\boldsymbol{s}}\mu_{\boldsymbol{s}}, & \text{for } i = M_{\boldsymbol{s}}; \\ -y_{\boldsymbol{s}}\mu_{\boldsymbol{s}}/(d_0 - 1), & \text{for } M_{\boldsymbol{s}} \neq i \leq d_0^+ + d_0^-; \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

for $i = 1, 2, ..., d_0$, where the *label* $y_{\boldsymbol{s}}$ is defined as 1 if $1 \leq M_{\boldsymbol{s}} \leq d_0^+$ and -1, otherwise. Thus if $y_{\boldsymbol{s}} = +1$, the $M_{\boldsymbol{s}}$th player is a big-time winner, while the reaming, non-neutral players are uniformly worse-off. The opposite holds for $y_{\boldsymbol{s}} = -1$, hence the name *Win-Loss game* or shortly *WL-game*. Note that $\boldsymbol{s}$ as above satisfies the 0-sum constraint:

$$\sum_{i=1}^{d_0} s_i = 0. \tag{2}$$

The subspace $S \subset \mathbb{R}^{d_0}$ of all such possible state vectors is called the *state space*. In general the state vector $\boldsymbol{s}$ is observed indirectly, via the measurement vector $\boldsymbol{x} = (x_1, ..., x_d) \in \mathbb{R}^d$ which is a linear mixture of state variables

$$\boldsymbol{x} = A\boldsymbol{s}, \tag{3}$$

where A is a $d \times d_0$ matrix. If rank(A) = d_0, then the label classes in both $S \subset \mathbb{R}^{d_0}$ and its image

$$X := AS = \{A\boldsymbol{s}\ ;\ \boldsymbol{s} \in S\} \subset \mathbb{R}^d$$

are *linearly separable*. Indeed, any hyperplane defined by the equation $s_i = 0$ for $i > d_0^+ + d_0^-$ always separates these datasets in $\mathbb{R}^{d_0}$, hence its image separates the data in span$(X) \subset \mathbb{R}^d$ and could be easily extended to the whole $\mathbb{R}^d$.

In general we shall consider $d \geq d_0$. In the particular case of $d = d_0$ and $A = I$ being the identity matrix, we say the game is directly observable. Another special case of interest, due to ease of analytical analysis, is *orthogonal mixing* with columns of A are composed of orthogonal vectors of equal length, i.e.

$$A^T A = CI, \tag{4}$$

where $C > 0$. We shall refer to this game as *orthogonal WL-game*. The above condition ensures that the following relations hold for dot-products:

$$C^{-1}\boldsymbol{x} \cdot \boldsymbol{x}' = \boldsymbol{s} \cdot \boldsymbol{s}' = \begin{cases} \mu_{\boldsymbol{s}}\mu_{\boldsymbol{s}'}d_0/(d_0 - 1), & \text{if } M_{\boldsymbol{s}} = M_{\boldsymbol{s}'}; \\ -\mu_{\boldsymbol{s}}\mu_{\boldsymbol{s}'}d_0/(d_0 - 1)^2 < 0, & \text{if } y_{\boldsymbol{s}} = y_{\boldsymbol{s}'} \text{ but } M_{\boldsymbol{s}} \neq M_{\boldsymbol{s}'} \\ \mu_{\boldsymbol{s}}\mu_{\boldsymbol{s}'}d_0/(d_0 - 1)^2 > 0, & \text{otherwise, i.e. if } y_{\boldsymbol{s}} \neq y_{\boldsymbol{s}'}, \end{cases} \tag{5}$$

for any two state vectors $\boldsymbol{s}, \boldsymbol{s}' \in S$, $\boldsymbol{x} = A\boldsymbol{s}$ and $\boldsymbol{x}' = A\boldsymbol{s}'$.

The equation (5) is the crucial relation for the theoretical understanding of anti-learning in this dataset. It states for instances of different modes: any two of the opposite label are more correlated than any two of the same label.

2.1 Empirical Learning Curves for Orthonormal WL-Game

Dataset. We have used WL-game to generate finite dataset $(\boldsymbol{x}_j, y_j) \in \mathbb{R}^d \times \{\pm 1\}$, $j = 1, ..., n$ as follows. First, we selected a random sample of states $(\boldsymbol{s}_j) \in S^n$ and generated a mixing $d \times d_o$ matrix A by Gramm-Schmidt orthonormalisation of columns of a random matrix; then we defined $y_j := y_{\boldsymbol{s}_j}$ and $\boldsymbol{x}_j := A\boldsymbol{s}_j$.

Performance metrics. We use the Area under Receiver Operating Characteristics (*AROC* or *AUC*) [9,10], the plot of the True Positive versus False Negative error rates, as our main performance metric. Additionally, we also use Accuracy (*ACC*) 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 performance of trivial classifiers, be it 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.

2.2 The i.i.d. Learning Curves

This experiment has been designed to demonstrate that anti-learning is a phenomenon of learning from a low size sample that disappears in the large size sample limit. We have used a synthetic orthogonal WL-game ($d_0^+ = d_0^- = 100$, $d_0 = 250$ and $d = 300$) to generate 2000 sample data set for re-sampling of a training set from, and then for testing classifiers (on the whole dataset). The results are plotted in Figure 1. We discuss the selected classifiers first.

Centroid. The *centroid* (*Cntr*), our basic (linear) classifier, is defined as follows:

$$f(\boldsymbol{x}) := \frac{2}{\|\boldsymbol{w}\|^2}\boldsymbol{w} \cdot \boldsymbol{x} - \frac{\|\boldsymbol{w}_+\|^2 - \|\boldsymbol{w}_-\|^2}{\|\boldsymbol{w}\|^2}$$

where $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{w}_y := \sum_{i,y_i=y} \boldsymbol{x}_i/n_y$, $y = \pm 1$, and $\boldsymbol{w} = (w_j) := \boldsymbol{w}_+ - \boldsymbol{w}_-$. It is a very simple machine, does not depend on tuning parameters, is the "high regularisation" limit of SVMs and ridge regression [11], and performs well on microarray classification tasks [11]. (The scaling and the bias b are such that the scores of "class" centers are equal to class labels, i.e. $f(\boldsymbol{w}_y) = y$ for $y = \pm 1$.)

Cross-Validation Learners. In Figure 1 we observe that for small size training samples, both *AROC* and *ACC* for primary classifiers such as SVM can reach values close to those for a classifier perfectly misclassifying the data. In such a case, the classifier $-f$ will classify data nearly perfectly. Obviously, for larger training samples, the reverse is true and f is preferred. Can such a decision to reverse the classifier or not be made in a principled way? The obvious way to address this issue is to perform additional cross validation on the training data in order to detect the "mode" of the classifier. A short reflection leads to the conclusion that there are a few possible strategies which can be used to insure that the proper detection of the mode actually happen. Perhaps the most straightforward one is as follows.

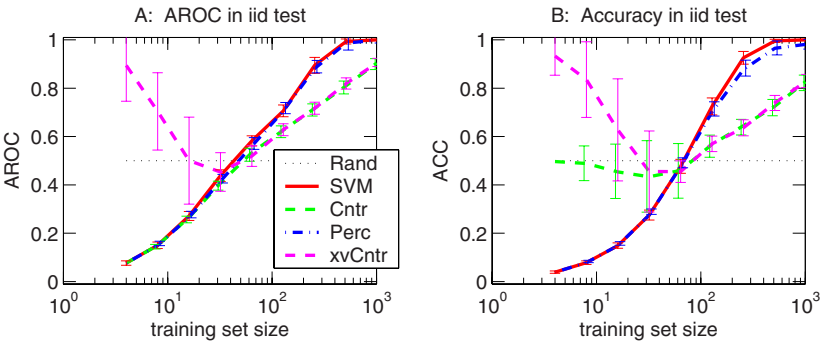

Fig. 1. Area under Reciver Operating Characteristic (AROC) and accuracy (ACC) as functions of increasing random training subset size for synthetic orthogonal WL-game data. We plot the averages of 100 tests on the whole dataset of 2000 instances with standard deviation marked by bars. We have used the following classifiers: Centroid (Cntr), hard margin support vector machine (SVM), Rosenblatt's perceptron [12] and xvCntr generated by Algorithm 1.

Algorithm 1 (xvL_1)**. Given:** a training set $Tr = (\boldsymbol{x}_i, y_i) \in (\mathbb{R}^d \times \{\pm 1\})^n$, an algorithm $f = \mathcal{A}(Tr)$;

1. Calculate cross-validation results, e.g. for LOO: $f_{vx}(i) := f^{\setminus i}(\boldsymbol{x}_i)$ for $f^{\setminus i} := \mathcal{A}(Tr \setminus (\boldsymbol{x}_i, y_i))$, $i = 1, ..., n$;
2. Calculate $aroc_{vx} := AROC((f_{vx}(i), y_i)_{i=1}^n)$;

Output classifier: $f = \phi \circ f' := \mathrm{sgn}(aroc_{xv} - 0.5) \times f'$, where $f' = \mathcal{A}(Tr)$.

Obviously one can use cross-validation schemes other than the leave-one-out (LOO) and can optimize other measures than $AROC$ in designing moderation of the output scores or just train an additional classifier. An example follows.

Algorithm 2 (xvL_2)**.** Use cross validation scores to train an additional 1-dimensional classifier $\phi := \mathcal{A}_2((f_{xv}(i), y_i)_{i=1}^n)$ and then use the superposition $\phi \circ f$ instead of $f = \mathcal{A}(Tr)$.

The function ϕ as in the above two Algorithms will be called a *reverser*.

Discussion of Figure 1. We clearly see disappearance of anti-learning phenomena in the large size sample limit. Note the poor performance of Cntr in Figure 1.B. This is due to the poor selection of the bias and is compatible with results of Theorem 2 and Corollary 1. The large variance for xvCntr is caused by a few cases of small size training samples which had the duplicate examples from the same mode, causing the miss-detection of the anti-learnable mode. Note that such single occurrence in 100 trails could result in std $\approx \sqrt{1/100} = 0.1$.

2.3 IID Anti-learning Theorem

In this section we generalise the analysis of the WL-game in Section 2 to more general kernel machines and prove a formal result on anti-learning for small

size samples in the i.i.d. setting observed in Figure 1. We consider a *kernel function* $k : X \times X \to \mathbb{R}$, on the measurements space $X = AS \subset \mathbb{R}^d$, although we do not need to assume that it is symmetric or positive definite, which are typical assumptions in the machine learning field [12,13,14]. Further, we assume probability distribution Pr on the state space S and consider an i.i.d. training n-sample $(\boldsymbol{s}_i) \in S^n$, with associated n-tuple of measurement-label pairs:

$$Tr := \big((\boldsymbol{x}_i, y_i)\big)_{i=1}^n := \big((A\boldsymbol{s}_i, y_{\boldsymbol{s}_i})\big)_{i=1}^n \subset \big(\mathbb{R}^d \times \{\pm 1\}\big)^n$$

and modes $M_i := M_{\boldsymbol{s}_i}$, for $i = 1, ..., n$. We assume we are given an algorithm that produces a *kernel machine* $f = \mathcal{KM}(k, Tr) : X \to \mathbb{R}$ of the form

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b \not\equiv \text{const}, \tag{6}$$

$$\alpha_i \geq 0 \ \ \& \ \ \sum_{i=1}^{n} y_i \alpha_i = 0. \tag{7}$$

for every $\boldsymbol{x} \in X$. Many algorithms, including popular flavors of SVM [13,12], the centroid (see Section 2.2) and Rosenblatt's perceptron [12], generate solutions satisfying the above conditions. If $b = 0$, we say that f is a *homogenous* machine.

We recall here a re-formulation of our metrics in terms of a probability distribution Pr on S and the order statistic U [10], for convenience:

$$\begin{aligned} AROC(f, S) = {} & Pr\big[f(\boldsymbol{x}_{\boldsymbol{s}}) < f(\boldsymbol{x}_{\boldsymbol{s}'}) \mid y_{\boldsymbol{s}} = -1 \ \& \ y_{\boldsymbol{s}'} = +1\big] \\ & - \frac{1}{2} Pr\big[f(\boldsymbol{x}_{\boldsymbol{s}}) = f(\boldsymbol{x}_{\boldsymbol{s}'}) \mid y_{\boldsymbol{s}} \neq y_{\boldsymbol{s}'}\big], \end{aligned} \tag{8}$$

$$ACC(f, S) = \frac{1}{2} Pr\big[f(\boldsymbol{x}_{\boldsymbol{s}}) < 0 \mid y_{\boldsymbol{s}} = -1\big] + \frac{1}{2} Pr\big[f(\boldsymbol{x}_{\boldsymbol{s}}) > 0 \mid y_{\boldsymbol{s}} = -1\big]. \tag{9}$$

Let $P_{\max} := \max_M Pr[M_{\boldsymbol{s}} = M]$ denote the maximum probability of a mode and by $\pi_y := Pr[y_s = y]$ be the prior probability of label y for $y = \pm 1$.

Theorem 1. *Assume that the kernel function k satisfies the condition*

$$y_{\boldsymbol{s}} y_{\boldsymbol{s}'} k(\boldsymbol{x}_{\boldsymbol{s}}, \boldsymbol{x}_{\boldsymbol{s}'}) < y_{\boldsymbol{s}} y_{\boldsymbol{s}'} b_0, \tag{10}$$

for every $\boldsymbol{s}, \boldsymbol{s}' \in S$ such that $M_{\boldsymbol{s}'} \neq M_{\boldsymbol{s}}$, where $b_0 \in \mathbb{R}$ is a constant. Let function $\psi : \mathbb{R} \to \mathbb{R}$ be monotonically increasing on the range of k, i.e. for $\xi \in k(X \times X)$.

Then for any kernel machine f trained for kernel $\psi \circ k$ on the n-sample Tr:

$$y_{\boldsymbol{s}} \sum_{i=1}^{n} y_i \alpha_i \ \psi \circ k(\boldsymbol{x}_i, \boldsymbol{x}_{\boldsymbol{s}}) < 0, \tag{11}$$

for every $\boldsymbol{s} \in S$ such that $M_{\boldsymbol{s}} \notin \{M_1, .., M_n\}$. Moreover, there exists B such that

$$AROC(f, S) \leq n P_{\max} / \min_y \pi_y, \tag{12}$$

$$ACC(f + B, S) \leq \frac{n}{2} P_{\max} / \min_y \pi_y. \tag{13}$$

Thus the homogenous kernel machine $\boldsymbol{s} \mapsto \sum_{i=1}^{n} y_i \alpha_i \ \psi \circ k(\boldsymbol{x}_i, \boldsymbol{x}_{\boldsymbol{s}})$ misclassifies every instance with mode unseen in training (see Eqn. 11).

Remark 1. The significance of the monotonic function ψ is that it allows extension of results automatically to many classes of practical kernels which can be represented as a monotonic function of the dot-product kernel. These include the polynomial kernels and, under the additional assumption of fixed magnitude of measurement vectors, the radial basis kernels.

Proof. First, let us note that if assumption (10) holds, then it also holds for the kernel $k \leftarrow \psi \circ k$ and constant $b_0 \leftarrow \psi(b_0)$. This reduces the proof to the special case of $\psi(\xi) = \xi$ for every $\xi \in \mathbb{R}$, assumed from now on.

For a proof of (11) let us assume that (10) holds. Then

$$y_{\boldsymbol{s}} f(\boldsymbol{x}_{\boldsymbol{s}}) = y_{\boldsymbol{s}} \sum_{i=1}^{n} y_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}) < y_{\boldsymbol{s}} b_0 \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Now we proceed to the proof of (12). Denote by $P := Pr[s,\ M_{\boldsymbol{s}} \in \{M_1, ..., M_n\}]$ the probability of an instance $\boldsymbol{s}$ having its mode present in the training set; by P_y the probability of such an instance $\boldsymbol{s}$ with label y. By (11) any two instances with modes not in the training sets are miss-ordered by f, hence

$$\begin{aligned}
AROC(f,S) &\leq 1 - \left(1 - \frac{P_-}{\pi_-}\right)\left(1 - \frac{P_+}{\pi_+}\right) \\
&\leq 1 - \left(1 - \frac{P_-}{\min(\pi_-,\pi_+)}\right)\left(1 - \frac{P_+}{\min(\pi_-,\pi_+)}\right) \\
&\leq \max_{0 \leq x \leq P} 1 - \left(1 - \frac{x}{\min(\pi_-,\pi_+)}\right)\left(1 - \frac{P-x}{\min(\pi_-,\pi_+)}\right) \\
&= \frac{P}{\min(\pi_-,\pi_+)} \leq \frac{nP_{\max}}{\min(\pi_-,\pi_+)}.
\end{aligned}$$

This completes the proof of (12). The proof of (13) follows

$$ACC(f+B,S) \leq \frac{1}{2}\left(\frac{P_+}{\pi_+} + \frac{P_-}{\pi_-}\right) \leq \frac{P_+ + P_-}{2\min(\pi_-,\pi_+)} = \frac{P}{2\min(\pi_-,\pi_+)} \leq \frac{nP_{\max}}{2\min(\pi_-,\pi_+)}.$$

□

Corollary 1. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a reverser generated by either Algorithm 1 or 2 for the homogeneous kernel machines. Then there exists a bias $B \in \mathbb{R}$ such that*

$$AROC(\phi \circ f, S) \geq 1 - n\frac{P_{\max}}{\min_y Pr[y_{\boldsymbol{s}} = y]}, \tag{14}$$

$$ACC(\phi \circ f + B, S) \geq 1 - n\frac{P_{\max}}{2\min_y Pr[y_{\boldsymbol{s}} = y]}, \tag{15}$$

with confidence $> \prod_{i=1}^{n-1}(1 - iP_{\max}) > 1 - \frac{(n-1)n}{2}P_{\max}$.

Note the "paradoxical" meaning of this result, compatible with experiments in Figure 1. The smaller the sample, the more accurate generalisation, provided the anti-learnable mode is detected and dealt with accordingly.

A simple proof (omitted) uses two observations: (i) that $\prod_{i=1}^{n-1}(1 - iP_{\max})$ is the lower bound on the probability of drawing n-different modes in that many samples, and (ii) that the assumptions insure that the inequality (11) holds for every kernel machine, hence also for f_{xv}, the pooled results of the cross-validation.

Note that for the orthogonal WL-game the dot product kernel $k(\boldsymbol{x}, \boldsymbol{x}') := \boldsymbol{x} \cdot \boldsymbol{x}'$ satisfies the assumption (10) of Theorem 1, see Eqn. 5. Thus we have

Corollary 2. *Corollary 1 holds for the linear kernel and orthogonal WL-game.*

3 Examples of Anti-learning in Natural Data

The esophageal adenocarcinoma dataset (AC) consists of 25 expressions of 9857 genes measured by cDNA microarrays in cancer biopsies collected from esophageal adenocarcinoma patients [1,2], prior to chemo-radio-therapy (CRT) treatment[1]. The binary labels were allocated according to whether the patient responded to the subsequent treatment (11 cases) or not (14 cases). The aim of the experiment was to assess the feasibility of developing a predictor of the response to treatment for clinical usage (an open problem, critical for clinical treatment).

We have also generated another synthetic data set, the output of the WL-game, but with $10,000 * 1000$ mixing matrix A drawn from the standard normal distribution (we have used $d_0^+ = d_0^- = 75$ and $d = 1000$). The data set consisted of 25 instances of each of the two labels. Back-to-back comparison of the classification of these two datasets in Figure 2 shows very similar trends indirectly linking the non-standard properties of AC-data to the anti-learning as understood in Section 2. Here we plot AROC as a function of number of features selected by t-test applied to the training set data only. In Figures 2.A & B we have used the following classifiers: Centroid (Cntr), hard margin support vector machine (SVM), shrunken centroid (PAM) [15] and 5-nearest neighbours (5-NN). In Figures C & D we have used various versions of xv-learner generated by the Algorithm 2 with and $\mathcal{A}_2$ generating the 1-dimensional linear reverser $\psi(\xi) := A\xi + B$ maximizing accuracy of the internal 2-fold cross-validation.

In Figure 3 we plot results for the additional test of 8 supervised learning algorithms on the natural AC-data. We observe that all averages are clearly below random guessing level of 0.5. These results show that the anti-learning persists for a number of standard classifiers, including multilayer algorithms such as decision trees or multilayer neural networks.

[1] Raw array data and protocols used are available at http://www.ebi.ac.uk/arrayexpress/Exp. The processed data used in this paper is available from http://nicta.com.au/people/kowalczyka

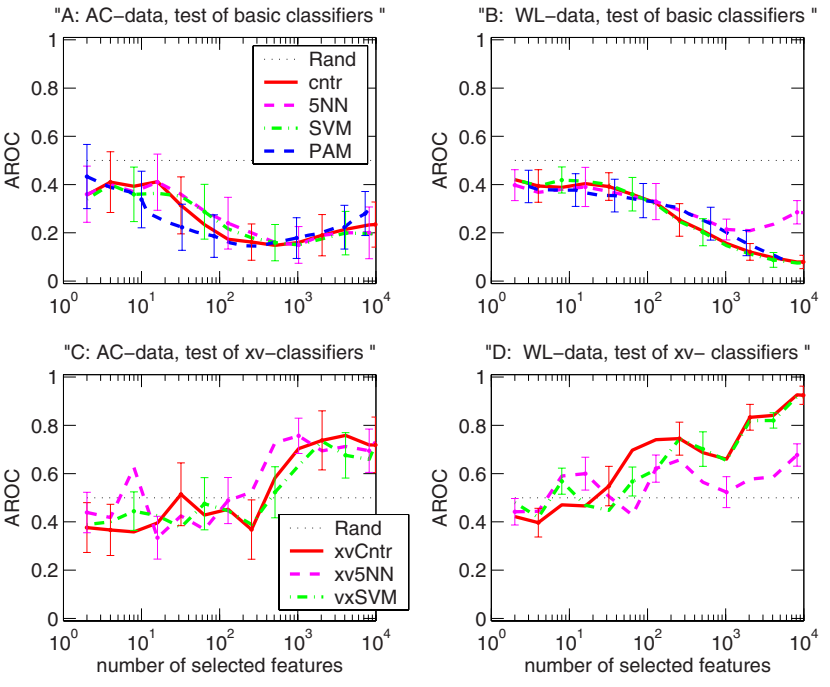

Fig. 2. Comparison of classification of natural adenocarcinoma (AC-data) and synthetic WL-game dataset for selected classifiers. We plot average of 20 repeats of 5-fold cross-validation. For all classifiers but PAM, the genes were selected using t-test applied to the training subset only. Note that PAM has built-in feature selection routine.

4 Discussion

The crux of anti-learning in our synthetic model is the inequality (5) stating that two examples of the opposite label are more "similar" to each other than two of the same label. This is a direct consequence of the 0-sum game constraints (2) combined with the "winner take all" paradigm. Such a simple "Darwinian" mechanism makes it plausible to argue that anti-learning signatures can arise in the biological datasets. However, there are also many other models generating anti-learnable signature, for instance a model of mimicry, which we shall cover elsewhere.

Anti-learning and esophageal adenocarcinoma. There are at least two reasons why research into anti-learning is currently critical for the project on prediction of CRT response in esophageal adenocarcinoma. Firstly, we need to prove that the measurements of gene expressions contain signal suitable for the prediction, so continuation of this expensive line of research is warranted. Secondly, apart from direct utility of CRT response prediction, there is a secondary, perhaps ultimate goal of this research, which is the determination of biology (say

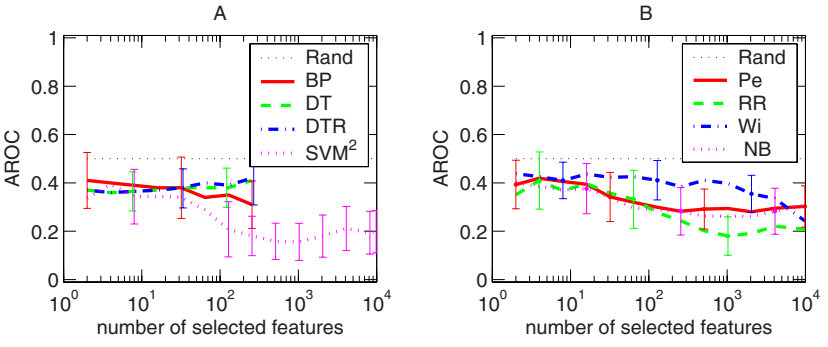

Fig. 3. Anti-learning performance of 8 selected classifiers on the natural AC-dataset (Figures A & C) and synthetic WL-game datasets (Figures B & D). The setting is similar to that in Figure 2, except that in Figure A we have tested for a smaller number of (preselected) features only, ≤ 256, as some of the implementations used did not run for the high dimensional input. For the following three algorithms: BP - Back-Propagation neural network, 5 hidden notes and 1 output, DT - Decision Trees and DTR - Regression Trees, we have used standard Matlab toolbox implementations, newff.m and treefit.m, respectively. For the remaining five algorithms, i.e. NB - Naive Bayes, Pe - Perceptron, RR - Ridge Regression [13], SVM^2 - SVM with the second order polynomial kernel and Wi - Winnow [16], we have used local custom implementations.

pathways) governing the CRT response, which could lead to a new treatment. As supervised learning signature of those precesses is most likely "anti-learnable" in view of our research, its proper interpretation and analysis is possible only from the position of anti-learning, since otherwise the data makes no sense and cannot lead to satisfactory conclusions.

Regularization. High regularization [12,13] is not an answer to the anti-learning challenge. In particular, the centroid, which is a "high regularization" limit of SVM and ridge regression [11], is systematically anti-learning on AC- and WL-datasets. Moreover, according to Theorem 2, for some datasets such as WL-game outcomes, SVM will anti-learn independently of how much regularization is used in its generation.

Kernels. Now let us consider the case of non-linear transformation of data via application of a kernel k [12,14,13]. Our Theorem 2, Remark 2 and Figure 3 argue that for some datasets the anti-learning extends to the popular kernels including the polynomial the radial basis kernels. This is compatible with a common sense observation that the anti-learning is not an issue of the too-poor hypothesis class, which is the main intuitive justification for kernel application.

Boosting. Now we turn to boosting, another heuristic for improving generalization of hard to learn data. The observation is that (ada)boosting [17] weak learners satisfying conditions (6), (7) and (11) outputs a convex combination of them, which again satisfies these conditions, hence the conclusion of Theorem 2

(see [5] for the similar argument line). Thus here the boosting does not change much at all. Intuitively, this is what one should expect: the boosting is effective in some cases where training data is difficult to classify. However, in the case in question, the training data is deceptively easy to deal with, but gives no clues of the performance on an independent test set.

Anti-learning and Overfitting. Overfitting is a deficiency of an algorithm with excessive capacity [12] which fits a model to idiosyncracies and noise of the training data. However, the anti-learning we are concerned with here is essentially different issue. Firstly, we prove that it is possible for a predictor to operate well below the accuracy of random guessing and still be a reliable forecaster. Secondly, we have shown that the anti-learning can be a signature of a deterministic phenomena (see the WL-game definition in Section 2).

The large sample limit and VC bounds. It follows clearly from Figure 1 that there is no contradiction between anti-learning and predictions of the learning theory such as VC-bounds [12,13]. Anti-learning occurs for a small size training set, where the asymptotic predictions of VC-theory are vacuous, and disappear in the large size sample limit, where VC-bounds hold.

5 Conclusions

We have demonstrated the existence of strong anti-learning behavior by a number of supervised learning algorithms on natural and synthetic data. Moreover, we have shown that a simple addition of an extra decision step, a reverser, can exploit this systematic tendency and lead to accurate predictor. Thus anti-learning is not a manifestation of over-fitting classifiers to the noise, but a systematic though usual, mode of operation of a range of supervised learning algorithms exposed to a non-standard dataset. Such a phenomenon, whenever encountered, should be systematically investigated rather than labelled as failure and forgotten. On a level of datamining we can offer a rough explanation of anti-learning by a specific geometry in the dataset, though this surely does not account for all of the phenomena encountered in nature. More research is needed into handling such datasets in practice as well as into the natural processes capable of generating such signatures.

Acknowledgements

We thank Justin Bedo and Garvesh Raskutti of NICTA, and Danielle Greenawalt and Wayne Phillips of Peter MacCallum Cancer Centre for help in preparation of this paper.

National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

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. Greenawalt, D., Duong, C., Smyth, G., Ciavarella, M., Thompson, N., Tiang, T., Murray, W., Thomas, R., Phillips, W.: Gene Expression Profiling of Esophageal Cancer: Comparative analysis of Barrett's, Adenocarcinoma and Squamous Cell Carcinoma. Int J. Cancer 120, 1914–1921 (2007)
2. Duong, C., Greenawalt, D., Kowalczyk, A., Ciavarella, M., Raskutti, G., Murray, W., Phillips, W., Thomas, R.: Pre-treatment gene expression profiles can be used to predict response to neoadjuvant chemoradiotherapy in esophageal cancer. Ann Surg Oncol (accepted, 2007)
3. Kowalczyk, A., Greenawalt, D., Bedo, J., Duong, C., Raskutti, G., Thomas, R., Phillips, W.: Validation of Anti-learnable Signature in Classification of Response to Chemoradiotherapy in Esophageal Adenocarcinoma Patients. Proc. Intern. Symp. on Optimization and Systems Biology, OSB (to appear, 2007
4. Kowalczyk, A., Chapelle, O.: An analysis of the anti-learning phenomenon for the class symmetric polyhedron. In: Jain, S., Simon, H.U., Tomita, E. (eds.) Proceedings of the 16th International Conference on Algorithmic Learning Theory, Springer, Heidelberg (2005)
5. Kowalczyk, A., Smola, A.: Conditions for antilearning. Technical Report HPL-2003-97(R.1), NICTA, NICTA, Canberra (2005)
6. Kowalczyk, A., Raskutti, B.: One Class SVM for Yeast Regulation Prediction. SIGKDD Explorations 4(2) (2002)
7. Raskutti, B., Kowalczyk, A.: Extreme re-balancing for svms: a case study. SIGKDD Explorations 6(1), 60–69 (2004)
8. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Computation 8(7), 1341–1390 (1996)
9. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning 42(3), 203–231 (2001)
10. Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psych. 12, 387–415 (1975)
11. Bedo, J., Sanderson, C., Kowalczyk, A.: An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics. In: Australian Conf. on Artificial Intelligence, pp. 170–180 (2006)
12. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons, New York (1998)
13. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
14. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge, MA (2002)
15. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Class prediction by nearest shrunken centroids, with applicaitons to dna microarrays. Stat. Sci. 18, 104–117 (2003)
16. Kivinen, J., Warmuth, M.K.: Additive versus exponentiated gradient updates for linear prediction. In: Proc. 27th Annual ACM Symposium on Theory of Computing, pp. 209–218. ACM Press, New York (1995)
17. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)