

# Fringe SVM Settings and Aggressive Feature Reduction \*

Adam Kowalczyk and Bhavani Raskutti  
Telstra, 770 Blackburn Road, Clayton, Victoria, Australia  
{Adam.Kowalczyk,Bhavani.Raskutti}@team.telstra.com

## Abstract

Statistical techniques for aggressive feature reduction are studied on data obtained in a gene knock-out experiment. Essential part of the process is automatic assessment of the quality of various feature selection methods. This is done by comparison of the performance of discriminating models built on candidate subsets of features. Experiments show that typical settings of popular 2-class discriminators, support vector machines (SVM), cannot be used as they produce models of very poor quality. The proposed way around is to use “fringe classifiers” such as SVMs trained on positive class data only or class centroids. Additionally, we also use models generated by such algorithms directly for identification of most discriminating features. We recommend that such simple machine learning techniques should be included into a repertoire of discriminators used on such occasions. We show that such relatively superior performance of fringe SVMs can also be observed on regular text-mining data bases, such as Reuters newswire benchmark, if only the less frequent features (words) are used.

---

\***Keyword:** machine learning, gene knock-out data, support vector machines, feature selection, text mining, aryl hydrocarbon receptor signalling, yeast genome

## 1 Introduction

As our knowledge of genome of various organisms increases in an unprecedented rate, the number of potentially relevant scientific articles proliferates, making it impossible for an individual to keep up. This creates a unique opportunity for various statistical techniques to provide useful tools capable of pre-filtering and identification of most relevant information, prior to human inspection. However, such filtering brings some new challenges for machine learning. These challenges stem from the following: (1) a need to deal with extremely high dimensional spaces with only a handful of “training” examples, (2) “low” relevance of the data sources to any particular problem and (3) the difference in nature between biomining domains and traditional well-understood text mining domains.

These challenges will require the machine learning community to re-evaluate some of their techniques. To that end, we present here an analysis of the data for the second task of the 2002 KDD (Knowledge Discovery and Data mining) Cup [4]. This task was set up on the basis of experiments identifying genes that, when knocked out, cause a significant change in the level of activity of the Aryl Hydrocarbon Receptor (AHR) signalling pathway. Our solution for this task, developed using purely statistical approach, with no domain knowledge whatsoever and a limited effort of a couple of days was unusual. It was a discriminating model trained on data

from the minority class examples only, and it was on par (in fact the winner) with other far more sophisticated approaches. This indicates that statistical data mining techniques indeed can be of great utility to biologists, though tools of general purpose utility are still far away.

In an attempt to develop such tools, we explore the use of support vector machines (SVM) in many different modes, in particular, *fringe classifiers* that correspond to some extreme settings of parameters, e.g. learning from examples of a single class and settings that provide solutions that are equivalent to learning from data centroids. We show that for the 2002 KDD cup data (*AHR-data*) [4], these centroid classifiers provide robust discriminating models, which are far more accurate than typical 2-class SVMs. Additionally, we show that the phenomenon of domination of fringe classifiers is not unique to AHR-data, but is observable in the popular Reuters Newswires text mining benchmark. Hence, fringe classifiers should not be automatically dismissed, but be part of the repertoire of learning algorithms explored.

We investigate the utility of such classifiers along with other discriminating models for feature selection. We use these classifiers, firstly for the evaluation of feature selection methods based on their accuracy of prediction, and secondly, for directly selecting informative features using the generated models. Our investigation shows that such model-based selection can both provide accurate classifiers and select a small subset of features which may be used in a variety of ways: (1) inspection by researchers for identification of biological mechanisms underlying an investigated phenomenon, (2) generation of key words for retrieval of relevant scientific articles, and (3) development of a more refined and simpler to interpret discriminating models.

The paper is organised as follows. Section 2 introduces the machine learning algo-

rithms and the performance measure used in this research. We then present our experiments with AHR-data in Section 3. Section 4 presents results on the Reuters data. In Section 5, we discuss the implications of our results and present some intuitive explanations of the observed phenomena.

## 2 Classifiers and Metrics

In this section we introduce the basic machine learning algorithms used in this paper. We focus on linear classifiers, in particular, on Support Vector Machines (SVM). There is a number of reasons for this focus. Firstly, SVMs have been proven to be top performers in text mining [6, 9, 14] and biomining tasks. Secondly, they have been top performers on AHR-data: 3 out of the top 5 submissions to KDD Cup 2002 were based on SVMs [7, 11, 10]. Most importantly, they are well suited to process sparse, high dimensional data.

Our classification problem is formulated as follows. Given a training sequence  $(x_i, y_i)$  of binary  $n$ -vectors  $x_i \in \{0, 1\}^n \subset \mathbb{R}^n$  and bipolar labels  $y_i \in \{\pm 1\}$  for  $i = 1, \dots, m$ . The case of prime interest here is when the target class, labelled  $+1$ , is much smaller than the background class (labelled  $-1$ ),  $\approx 1\%$  of the data. Our aim is to find a “good” discriminating linear function

$$f(x) := w \cdot x + b \quad (1)$$

that scores the target class instances higher than the background class instances. Here  $x \in \mathbb{R}^n$ , “ $\cdot$ ” denotes the dot product in  $\mathbb{R}^n$  and  $(w, b) \in \mathbb{R}^n \times \mathbb{R}$  is defined by one of the five learning algorithms described below.

The first four algorithms, requiring dedicated solvers, are versions of the popular SVMs. For all of them the solution  $(w, b) \in \mathbb{R}^n \times \mathbb{R}$  is defined as a minimiser of the regu-

larised risk functional of the following form.

$$\|w, b\|_*^2 + \sum_{i=1}^m C_i \phi(1 - y_i(w \cdot x_i + b)), \quad (2)$$

where  $\|\cdot\|_*^2$  is a squared “norm” penalising for the “complexity of the classifier”,  $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$  is a convex loss function penalising for deviations of the scores of the machine from allocated labels and the regularisation constants are defined as follows:

$$C_i = \begin{cases} (1+B)C/(2m_+) & \text{if } y_i = +1, \\ (1-B)C/(2m_-) & \text{if } y_i = -1, \end{cases} \quad (3)$$

for  $i = 1, \dots, m$ , where  $C > 0$ ,  $m_+$  and  $m_-$  denote the numbers of examples with labels  $y_i = +1$  and  $y_i = -1$ , respectively. Here  $-1 \leq B \leq 1$  is a *balance parameter* designed to balance the impact of instances from the positive and the negative class.

Now we specify four variations of the regularised risk (2) leading to four different machines to be used in this paper.

**Algorithm 1,  $SVM_{BC}^1$ :** This is the popular *SVM with linear penalty*. Here, we use the norm  $\|w, b\|_*^2 := \|w\|^2 = w \cdot w$  and the “hinge loss”  $\phi(\theta) := \max(0, \theta)$ ,  $\theta \in \mathbb{R}$  [5, 17, 18];

**Algorithm 2,  $hSVM_{BC}^1$ :** Replacing the norm in the above definition by

$$\|w, b\|_*^2 := \|w\|^2 + b^2 \quad (4)$$

we obtain the *homogeneous SVM with linear penalty*;

**Algorithm 3,  $hSVM_{BC}^2$ :** For the (*homogeneous*) *SVM with quadratic penalty* [5] we use norm (4) and the squared hinge loss  $\phi(\theta) := (\max(0, \theta))^2$  for  $\theta \in \mathbb{R}$ ;

**Algorithm 4,  $hRN_{BC}^2$ :** For the *homogeneous regularisation network* [8, 19] or ridge regression [5, 8, 19] we use norm (4) and ordinary square loss  $\phi(\theta) := (\theta)^2$  for  $\theta \in \mathbb{R}$ .

## 2.1 1-class SVMs

Note that  $hSVM_{BC}^1$ ,  $hSVM_{BC}^2$  and  $hRN_{BC}^2$  implement classifiers that correspond to sep-

aration of the data  $(x_i, 1, y_i) \in \mathbb{R}^n \times \mathbb{R} \times \{\pm 1\}$  by a hyperplane  $\langle (w, b), (x, 1) \rangle = 0$  passing through the origin  $(0, 0) \in \mathbb{R}^n \times \mathbb{R}$ . One thing to stress is that (2) provides a non-trivial (i.e.  $\neq \text{constant}$ ), unique classifier (1) in all “regular” cases of interest, in particular, for  $B = \pm 1$  if at least one  $C_i \neq 0$  and  $(0, 0) \in \mathbb{R}^n \times \mathbb{R}$  does not belong to the convex shell spanned by all vectors  $y_i(x_i, 1) \in \mathbb{R}^N \times \mathbb{R}$ . These cases of  $B$  are equivalent to data belonging to a single class label, the target class  $y_i = +1$  for  $B = +1$  and the background class  $y_i = -1$  for  $B = -1$ . We shall call such machines the *1-class SVM’s*.

On the level of the extended feature space  $\mathbb{R}^n \times \mathbb{R}$ , any  $hSVM_{BC}^1$ ,  $hSVM_{BC}^2$  or  $hRN_{BC}^2$  can be reduced to a single class machine. In fact, we can always absorb the signum  $y_i$  by considering the data  $(\tilde{z}_i, \tilde{y}_i) := (y_i x_i, y_i, 1)$  rather than  $(x_i, 1, y_i)$  and then minimising the following functional equivalent to (2):

$$\langle \tilde{w}, \tilde{w} \rangle + \sum_{i=1}^m C_i \phi(1 - \langle \tilde{w}, \tilde{z}_i \rangle) \quad (5)$$

where  $\langle \cdot, \cdot \rangle$  stands for the dot product in  $\mathbb{R}^n \times \mathbb{R}$ . This formally reduces the two class problem (2) to “single class learning”. In the case of  $hSVM^1$ , the solution to (5) can be found using  $SVM^1$  if an extra point, namely  $(0, 0) \in \mathbb{R}^N \times \mathbb{R}$ , with the opposite label  $-1$ , is added to the data. Such a method for one-class learning has been considered previously in [12, 16].

## 2.2 Centroids

Now we introduce the fifth and the simplest of the five algorithms considered here.

**Algorithm 5,  $Cnt_B$ :** We set  $b := 0$  and

$$w := \frac{1+B}{2m_+} \sum_{i, y_i=+1} x_i - \frac{1-B}{2m_-} \sum_{i, y_i=-1} x_i,$$

For  $B = +1$  vector  $w$  is exactly the centroid of the minority (the target) class, for  $B = -1$

it is the centroid of the majority (the background) class while for  $B = 0$  it is half of the difference between the centroids of the two classes.

From Karush-Khun-Tucker conditions for SVM solution, it follows that for low  $C$ , the direction  $w$  of SVM becomes that of the centroid solution. More precisely, we have the following formal result, which as we shall see, is reflected in our experimental results.

**Theorem 1** *If vectors  $x_i \in \mathbb{R}^n$ ,  $i = 1, \dots, m$  are linearly independent, then*

$$\lim_{C \rightarrow 0^+} \frac{w_{hSVM_{BC}^p}}{C} = \lim_{C \rightarrow 0^+} \frac{w_{hRN_{BC}^2}}{C} = w_{Cntr_B}, \quad (6)$$

for  $-1 \leq B \leq 1$  and  $p = 1, 2$ , where  $w$  denotes the solution vector for the appropriate machine. Moreover

$$\lim_{C \rightarrow 0^+} \frac{w_{SVM_{BC}^1}}{C} = w_{Cntr_B}, \quad (7)$$

for  $-1 < B \leq 1$ .

We shall refer to any  $hSVM_{\pm 1, C}^p$ ,  $p = 1, 2$ ,  $hRN_{\pm 1, C}^2$  or  $Cntr_{\pm 1}$  as *fringe SVMs*. We include here  $Cntr_{\pm 1}$  since in view of the above result the direction of  $w$  for such a machine can be at least approximately obtained by minimising regularised risk of the form (2) and our figure of merit, to be introduced below, is fully determined by such a direction.

## 2.3 Performance measures

We have used  $AROC$ , the Area under the Receiver Operating Characteristic (ROC) curve as our main performance measure. In that we follow the steps of KDD 2002 Cup, but also, we see it as the natural metric of general goodness of classifier (as corroborated below) capable of meaningful results even if the target class is a tiny fraction of the data.

We recall that the ROC curve is a plot of the *true positive rate* or precision,  $P(f(x_i) > \theta | y_i = 1)$ , against the *false positive rate*,

$P(f(x_i) > \theta | y_i = -1)$ , as a decision threshold  $\theta$  is varied. The concept of ROC curve originates in the military signal detection but these days it is widely used in many other areas, including data mining, psychophysics and medical diagnosis (cf. review [2]). In the latter case,  $AROC$  is viewed as a measure of general “goodness” of a test, formalised as a predictive model  $f$  in our context, with a clear statistical meaning as follows. According to Bamber interpretation [1],  $AROC(f)$  is equal to the probability of correctly answering the two-alternative-forced-choice problem: given two cases, one  $x_i$  from the negative and the other  $x_j$  from the positive class, allocate scores in the right order, i.e.  $f(x_i) < f(x_j)$ . Additional attraction of  $AROC$  as a figure of merit is its direct link to the well researched area of order statistics via  $U$ -statistics and Wilcoxon-Whitney-Mann test [1].

There are some ambiguities in the case of  $AROC$  estimated from a discrete set in the case of ties, i.e. when multiple instances from different classes receive the same score. Following [1] we implement in this paper the definition

$$AROC(f) = P(f(x_i) < f(x_j) | -y_i = y_j = 1) + 0.5P(f(x_i) = f(x_j) | -y_i = y_j = 1)$$

expressing  $AROC$  in terms of conditional probabilities. In the case of  $f$  being a continuous random variable, the second term above disappears, and  $AROC$  is uniquely determined by the unique order imposed by scores allocated by  $f$ .

Note that trivial uniform random predictor has  $AROC$  of 0.5.

## 3 Analysis of AHR-data

In our main experiments we have used AHR-data set which is the combined training and test data sets used for task 2 of KDD Cup 2002. The data set is based on experiments

by Guang Yao and Chris Bradfield of McArdle Laboratory for Cancer Research, University of Wisconsin. These experiments aimed at identification of yeast genes that, when knocked out, cause a significant change in the level of activity of the Aryl Hydrocarbon Receptor signalling pathway (cf. [4] for more details). In this paper we follow the setting of the “broad task” of the KDD Cup: the discrimination between 127 ‘positive’ genes from the combined class encompassing the labels “change” and “control” and the remaining 4380 genes forming the ‘negative’ class. In our experiments this set has been repeatedly split into 70% for training and 30% for testing. All averages and standard deviations reported are for independent tests on 20 such random splits.

### 3.1 Data representation

Each training and test gene was represented by a vector of binary attributes extracted from the data sources provided. Attributes were extracted from very rich data sources composed of function/localisation annotations, protein-protein interactions and Medline abstracts.

*Hierarchical information about function, protein classes and localisation* was converted to a vector per gene. For instance, the following two entries in the file `function.txt`

```
YGR072W CYTOPLASM | SUBCELLULAR LOCALISATION
```

```
YGR072W NUCLEUS | SUBCELLULAR LOCALISATION
```

yielded three function attributes: “cytoplasm”, “subcellular localisation” and “nucleus” each with a value of 1 for the gene “YGR072W”. This processing created 409 attributes: 213 for gene function, 154 for protein classes and 42 for localisation.

*Textual information from all abstracts* associated with a gene was converted to ‘word token’ presence vectors (‘a bag of words’). A ‘word token’, in this context, is understood as

any string of alphanumeric characters, which may and may not correspond to an ordinary word. Word tokens corresponding to words in a standard list of stop words, such as “the”, “a” and “in”, have been excluded. All ordinary words were stemmed using a standard Porter stemmer. The resulting word token attributes were then checked against the gene alias file, and all aliases were replaced by a single gene name.

The above abstract processing resulted in 48,089 word token attributes. Around 3/4th of these attributes were subsequently eliminated by discarding all those that occurred in only one training gene, and by discarding all those which had a total frequency that was greater than one standard deviation from the mean. After this processing, we were left with 16,474 attributes from the abstracts.

*The gene-gene interaction* file is symmetric. Hence, each entry in the file `interaction.txt` creates two attributes. For instance, the entry, “YFL039C YMR092C” creates two interaction attributes: “YFL039C” and “YMR092C”, and the attribute “YFL039C” is set to 1 for the gene “YMR092C” and vice-versa. Processing of the gene interactions file yielded a total of 1,447 attributes.

Thus, the total number of binary attributes used by the learning algorithm was 18,330 ( $= 409 + 16,474 + 1,447$ ).

### 3.2 Regularisation constant $C$

In order to study the impact of regularisation constant  $C$  on SVM solutions, we plot in Figure 1 mean AROC for four different SVMs as a function  $C$ . We use four different modes: (i) positive 1-class ( $B = +1$ , solid line); (ii) negative 1-class ( $B = -1$ , dotted line); (iii) balanced 2-class ( $B = 0$ , dashed line); (iv) un-balanced 2-class (with  $C_i \equiv C$ ,  $i = 1, \dots, m$ , the dash-dot line). Standard deviations are shown as vertical bars.

An inspection of plots brings a number of interesting observations:

1. The un-balanced 2-class machines and negative 1-class machines have inferior performance relative to either positive 1-class machines or the balanced 2-class SVMs for low  $C$ . Thus only the last two modes will be used in further research in this paper.

2. All positive 1-class and balanced 2-class machines show a very good and roughly equal performance for very low values of  $C$ . Additionally, these values of mean AROC are equal to that for the positive 1-class centroid  $Cntr_{+1}$  (AROC =  $62.4 \pm 3.4$ ) and the 2-class centroid  $Cntr_0$  (AROC =  $61.5 \pm 3.9$ ). This can be inferred from Theorem 1. Indeed, the theorem implies that the directions of the vector of  $w$  of the SVM solutions converge to the directions of the corresponding centroid solutions. Hence values of AROC converge too, since AROC of any linear classifier (1) is uniquely determined by the direction of  $w$ .

3. Note also that unbalanced, homogeneous 2-class SVMs (dash-dot lines) and negative 1-class SVMs (dotted lines) also ‘converge’ in the low  $C$  limit. This is not surprising, as the unbalanced SVM solution is dominated by the negative class instances composing 97.2% of the data. The limit for negative 1-class case is roughly the same as the mean AROC for negative 1-class centroid  $Cntr_{-1}$ : AROC =  $38.2 \pm 3.3$ , which agrees with the Theorem 1 for the case of  $B = -1$ .

4. There are noticeable differences between the performance of different SVMs. For instance, note differences between unbalanced 2-class  $hSVM^1$  and  $SVM^1$  (dash-dot lines in Figures 1A and 1B, respectively).

5. Positive 1-class  $hSVM^2$  is very robust across the whole range of  $C$  values (cf. the solid line in Figure 1C). In particular, for high values of  $C$ , i.e., virtually the hard margin case [10], it performs better than any other SVM tested. This setting was used for the winning submission to KDD Cup 2002.

6. Un-balanced homogeneous 2-class SVMs (dash-dot lines) and negative 1-class SVMs (dot lines) in Figures 1B-1D, perform

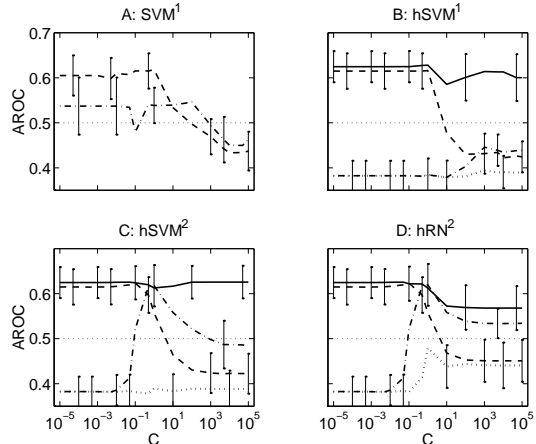


Figure 1: Mean AROC for 1-class and 2-class SVMs as a function of the regularisation constant  $C$  for four linear SVMs and four different modes: positive 1-class ( $B = +1$ , solid line), negative 1-class ( $B = -1$ , dotted line), balanced 2-class ( $B = 0$ , dashed line) and ‘un-balanced’ 2-class (with  $C_i \equiv C$ ,  $i = 1, \dots, m$ , dash-dot line).

very similarly in the low  $C$  limit.

### 3.3 Feature selection

We now explore the utility of the above classifiers for feature selection. We consider their utility as (i) techniques for evaluation of various selected feature sets and (ii) tools for selection of such sets.

We investigate several strategies for scoring features: the first four score features based on their distribution in the training set, while the others are based on the models ( $w$ ) generated. In all cases, the computed score is used to sort the features so that the most informative features may be selected.

**A: DocFreq** (Document frequency thresholding): This method has its origins in information retrieval [15] and is based on the notion that rare features are not informative for predicting classes. In this case the score of a feature is simply the number of instances where it has been equal to 1.

**B: ChiSqua** ( $\chi^2$ ): The  $\chi^2$  measures the

lack of independence between a feature and a class of interest. First, for each feature and each class, i.e.  $y = \pm 1$ , a score is computed on the basis of the two-way contingency table [20]. The final score for a feature is set to the maximum of these class scores.

**C: MutInfo:** (Mutual Information): The following score is allocated to  $j$ th feature:

$$MutInfo(j) = \max_{y=\pm 1} \log \frac{P(x_{i,j} = 1, y_i = y)}{P(x_{i,j} = 1)P(y_i = y)}$$

where the joint and marginal probabilities are estimated from the training set, i.e. with respect to index  $i$ ,  $1 \leq i \leq m$  [20].

**D: InfGain:** (Information gain): This is frequently employed as a term goodness measure in machine learning [13], and measures the number of bits of information obtained for class prediction by knowing the presence or absence of a term in an instance.

**E-J: Model-based feature selection:** In this case, the score of a feature is simply the magnitude of the weight allocated to it by a linear model generated to discriminate two classes of interest, i.e. corresponding entries in the vector  $w = (w_1, \dots, w_n)$ . In our experiments we employ a small variation: we average  $w$  over 20 models generated for 20 random splits of the data into 70% training and 30% test sets. We have used three learning machines,  $hSVM_{B,5000}^1(E,F)$ ,  $hSVM_{B,5000}^2(G,H)$  and  $Cntr_B(I,J)$  in two modes: positive 1-class mode  $B = +1$  (E,G,I) and balanced 2-class mode,  $B = 0$  (F,H,J), thus yielding six additional methods.

Figure 2 shows the results of evaluation of the ten feature selection techniques (columns A-J) by four different algorithms (rows 1-4). Each evaluation technique has been used in two modes: positive 1-class and balanced 2-class. (We have not shown evaluation results for  $SVM^1$  as they are very similar to those for 2-class  $hSVM^1$ .) The results can be summarised as follows:

1. As a general rule, for all SVMs, 1-class models perform much better than 2-

class models when using the same set of features. In addition, these two modes of SVM often give quite opposite evaluation of the utility of selected features (the notable exception being column H). While 1-class finds them informative ( $AROC > 0.5$ ), 2-class finds them detrimental with  $AROC < 0.5$ , i.e. below that of the random classifier.

2. *DocFreq* and *MutInfo* both provide very poor results for low number of features, although they use completely different metrics for scoring. *MutInfo* is strongly influenced by the marginal probability of terms and hence tends to favour rare terms, while *DocFreq* selects the most common terms.

3.  $Cntr_0$  (row 4, dashed line) performs the best of all 2-class algorithms, generally matching 1-class centroid classifier,  $Cntr_{+1}$ .

4. 2-class  $hSVM_{0,5000}^2$  (Columns H) provides very good features for 2-class mode classifiers, allowing them to perform above random  $AROC > 0.5$ . In fact, this feature selection in combination with ridge regression learning,  $hRN_{0,5000}^2$ , provides the best performance for around 2% ( $\approx 300$ ) features.

5. As a general trend, features selected by models allow development of better discriminating models than features selected by evaluated algorithms, provided positive 1-class mode is used for learning.

### 3.3.1 Selected Features

Table 1 lists the top 20 features selected according to different methods. We observe that there is a large overlap between features selected by  $\chi^2$  and positive 1-class methods:  $hSVM_{+1,5000}^p$ ,  $p = 1, 2$  and  $Cntr_{+1}$ . The features selected are primarily from function and localisation data. The 10 common features in the top 20 are: 1:  $F_4$  - subcellular localisation, 2:  $F_7$  - cell cycle and dna processing, 3:  $F_{10}$  - metabolism, 4:  $F_{17}$  - cell fate, 5:  $F_{27}$  - mrna transcription, 6:  $F_{28}$  - transcription, 7:  $F_{29}$  - unclassified proteins, 8:  $F_{32}$  - cellular transport and transport mechanisms, 9:  $F_{58}$

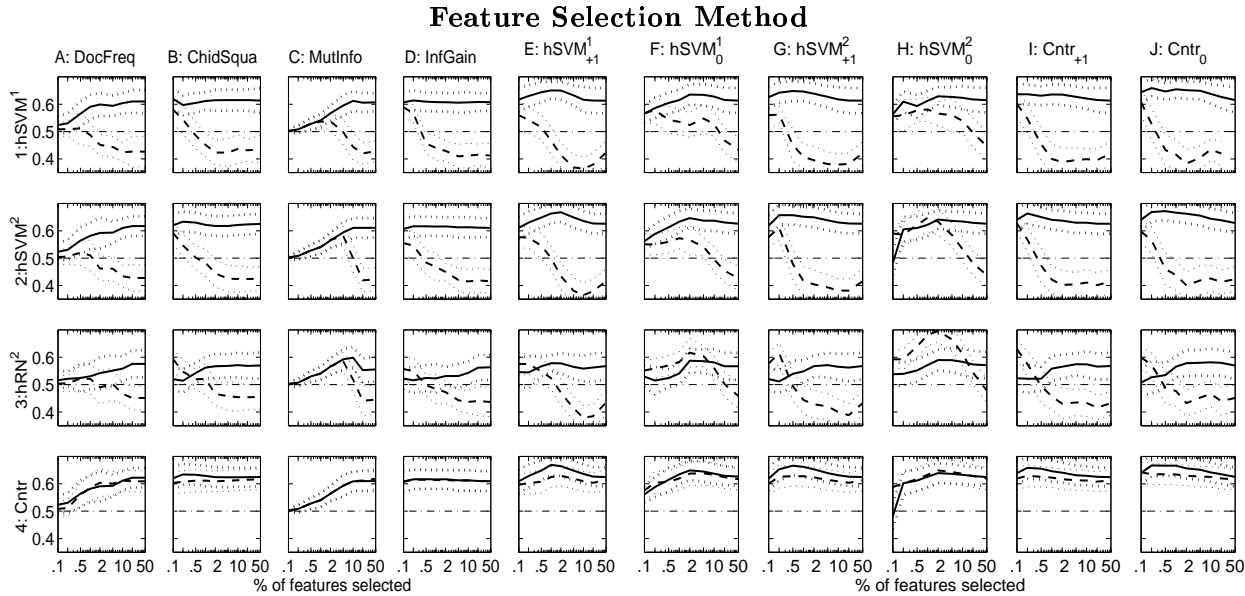


Figure 2: Evaluation of ten feature selection methods (columns A-J) by four classification algorithms (rows 1 - 4). Plots show mean AROC  $\pm Std$  as an envelope as a function of the % of features selected out of the total 18,330. Curves are plotted for two modes: the positive 1-class ( $B = +1$ , solid lines) and balanced 2-class ( $B = 0$ , dashed lines). All SVMs used  $C = 5000$ .

- protein fate (folding, modification, destination), 10:  $L_4$  - cytoplasm.

Similarly, there are overlaps between *InfGain* and the 2-class centroid,  $Cntr_0$ . However, these sets include many features from the abstracts, and thus are different from those selected by the positive 1-class methods. The 15 common features are: 1:  $F_4$  - subcellular localisation, 2:  $F_{22}$  - nucleus, 3:  $L_3$  - nucleus, 4:  $A_{419}$  - redund, 5:  $A_{426}$  - much, 6:  $A_{543}$  - abnorm, 7:  $A_{613}$  - comprom, 8:  $A_{639}$  - despit, 9:  $A_{711}$  - harbor, 10:  $A_{973}$  - surprisingli, 11:  $A_{1002}$  - subset, 12:  $A_{1291}$  - carboxi, 13:  $A_{1609}$  - green, 14:  $A_{2104}$  - taken, 15:  $A_{4290}$  - inviabl.

Interestingly, there are no overlaps in the top 20 features between the 2-class centroid ( $Cntr_0$ ) and the 2-class SVMs ( $B = 0$ ) or between *InfGain* and the 2-class SVMs.

## 4 Tests on Reuters

Experiments reported in the previous section show that fringe SVMs tend to perform better than traditional 2-class SVMs on AHR-data. In this section we report some experiments with popular text mining benchmark, Reuters-21578 news-wires, which show similar tendency. For these experiments we used a collection of 12902 documents (combined test and training sets of so called modApte split available from <http://www.research.att.com/lewis>) which are categorised into 115 overlapping categories. Each document in the collection has been converted to a vector of 20,197 dimensional word-presence feature space (full analogy to the preprocessing of Abstract for AHR-data). Then we gradually removed most frequent features (highest *DocFreq* scores) and trained classifiers on random 5% and then tested on remaining 95% of the data. As usual, the average AROC  $\pm Std$  for 20 such tests is shown in Figure 3. Four

Table 1: Top twenty features selected by ten feature selection methods. We use the convention: the letter stand for the data source (A -abstracts, F - function class, P - protein class, I - gene interactions, and L -localisation) and the subscript is the number of the feature. The last row gives mean AROC  $\pm$  Std of the models used for the model-selection method. We put “-” in front of features with negative weights (2-class SVMs).

Rank	Doc-Freq	Chi-Squa	Mut-Info	Inf-Gain	$hSVM_{B,5000}^1$		$hSVM_{B,5000}^2$		$Cntr_B$	
					B=1	B=0	B=1	B=0	B=1	B=0
1	$F_{95}$	$F_4$	$P_{48}$	$F_4$	$F_{29}$	$F_2$	$F_{29}$	$L_4$	$F_4$	$F_4$
2	$I_{1150}$	$F_{29}$	$P_{92}$	$F_{22}$	$F_4$	$F_{46}$	$F_4$	$I_{953}$	$F_{29}$	$A_{2104}$
3	$A_{1045}$	$F_{10}$	$P_{110}$	$L_3$	$F_{20}$	$F_{150}$	$F_{58}$	$I_{104}$	$L_3$	$F_{22}$
4	$I_{1136}$	$L_3$	$I_{61}$	$A_{2104}$	$P_{17}$	$F_{10}$	$F_{10}$	$-F_{10}$	$F_{22}$	$L_3$
5	$F_{172}$	$F_{22}$	$I_{66}$	$A_{4260}$	$F_{58}$	$F_{58}$	$L_4$	$-F_{46}$	$F_{10}$	$A_{4260}$
6	$F_{39}$	$F_{28}$	$I_{83}$	$A_{543}$	$L_4$	$-I_{104}$	$L_3$	$F_{110}$	$F_{58}$	$A_{426}$
7	$A_{89}$	$F_{58}$	$I_{85}$	$A_{426}$	$F_{21}$	$F_{45}$	$F_7$	$I_{835}$	$F_7$	$A_{639}$
8	$A_{925}$	$F_7$	$I_{95}$	$A_{711}$	$F_{10}$	$F_{38}$	$F_{22}$	$I_{210}$	$F_{17}$	$A_{543}$
9	$A_{970}$	$L_4$	$I_{533}$	$A_{1609}$	$F_{75}$	$F_{28}$	$F_{28}$	$P_{64}$	$A_{426}$	$A_{1002}$
10	$A_{1098}$	$F_{27}$	$I_{534}$	$A_{1002}$	$F_{86}$	$-L_4$	$F_{27}$	$F_{29}$	$A_{2104}$	$A_{1609}$
11	$A_{426}$	$F_{21}$	$I_{535}$	$A_{973}$	$F_{67}$	$-I_{835}$	$F_{17}$	$-F_{28}$	$F_{28}$	$A_{419}$
12	$A_{430}$	$F_{17}$	$I_{645}$	$A_{613}$	$F_{32}$	$-I_{351}$	$F_{67}$	$-F_2$	$F_{27}$	$A_{711}$
13	$A_{448}$	$F_{34}$	$I_{658}$	$A_{1291}$	$F_7$	$P_2$	$F_{32}$	$F_{75}$	$A_{639}$	$F_{58}$
14	$I_{1129}$	$F_{32}$	$I_{686}$	$A_{639}$	$F_{65}$	$F_{40}$	$P_{17}$	$-F_{45}$	$L_4$	$F_7$
15	$A_{669}$	$F_{26}$	$I_{727}$	$A_{562}$	$F_{66}$	$F_{20}$	$F_6$	$-F_{58}$	$A_{543}$	$A_{1714}$
16	$A_{586}$	$F_{15}$	$I_{838}$	$F_{27}$	$F_{64}$	$F_{32}$	$F_{21}$	$-F_{39}$	$F_{32}$	$A_{1291}$
17	$I_{1123}$	$F_{16}$	$I_{870}$	$A_{1699}$	$F_{17}$	$-I_{210}$	$F_{20}$	$A_{1721}$	$A_{1714}$	$F_{17}$
18	$I_{1299}$	$F_6$	$I_{871}$	$A_{849}$	$F_{28}$	$F_{59}$	$F_{65}$	$I_{351}$	$A_{973}$	$A_{613}$
19	$I_{1212}$	$A_{426}$	$I_{1050}$	$A_{1345}$	$F_{27}$	$P_{16}$	$F_{66}$	$P_{48}$	$A_{1282}$	$A_{973}$
20	$P_{17}$	$F_{25}$	$I_{1160}$	$A_{419}$	$F_6$	$F_{39}$	$F_{26}$	$-F_{38}$	$A_{1002}$	$A_{273}$
AROC					.61 $\pm$ .05	.43 $\pm$ .05	.63 $\pm$ .04	.42 $\pm$ .05	.62 $\pm$ .03	.62 $\pm$ .04

different target cases were used: the 3rd, the 6th, the 9th and the combined 11th-15th largest categories. The sizes of target classes are shown in the sub-figure titles.

An inspection of plots highlights a few observations:

1. The accuracy of all classifiers is very high when all features are used. As informative features are removed all SVM models start degenerating, however, the drop in performance for 2-class SVM models is much larger, and 1-class SVM models start outperforming the 2-class models. This behaviour is also present in other categories not shown in Figure 3, so long as the target class is less than 10% of the total data. This trend of better performance with 1-class models is most apparent in  $hSVM_{B,5000}^1$ , although

$hSVM_{B,5000}^2$  also shows similar trends. Thus, when there are many weakly informative features, and the target class is a small fraction of the data set, fringe classifiers outperform traditional 2-class SVM models.

2. Nowhere is the mean AROC  $<$  0.5 indicating the even after feature removal, this data set does not quite have all the properties of AHR-data where 2-class models performed worse than random for many settings of the regularisation constant.

## 5 Discussion

**Related Research.** A possibility of single class learning with support vector machines (SVM) has been noticed previously. In par-

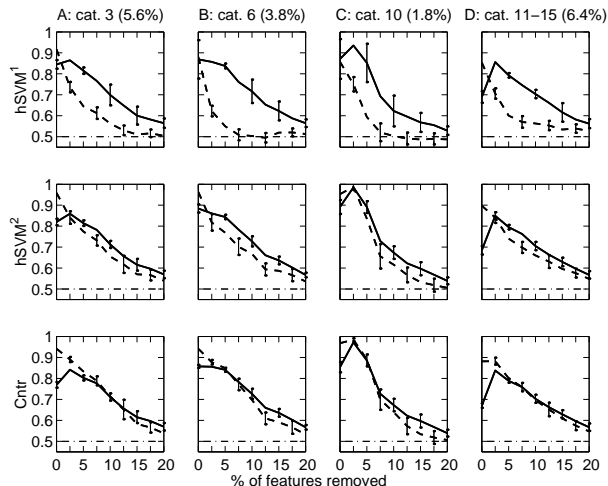


Figure 3: Mean AROC as a function of the % of features removed (with standard deviation envelope). Four different target cases were used: the 3rd, the 6th, the 9th and the combined 11th-15th largest categories. Results are presented for four machines: (1)  $hSVM_{B,5000}^1$ , (2)  $hSVM_{B,5000}^2$  (3)  $RN_{B,5000}^2$  and (4)  $Cntr_B$ . Plots are shown for the positive 1-class ( $B = +1$ ) (solid line) and the balanced 2-class ( $B = 0$ ) (dashed line) modes.

ticular, Schölkopf et al. [16] have suggested a method of adapting the SVM methodology to 1-class learning by treating the origin as the only member of the second class. This methodology has been used for image retrieval [3] and for document classification [12]. In both cases, modelling is performed using examples from the positive class only, and the 1-class models perform reasonably, although much worse than the 2-class models learned using examples from both classes.

In contrast, in this paper, we show that for certain problems such as AHR-data, positive 1-class SVMs significantly outperform models learned using examples from both classes.

**Deterioration of 2-class SVMs.** We have observed that for AHR-data, fringe SVMs tend to have systematically better AROC than the traditional 2-class SVMs. Typically, the latter deteriorate with increase

in the number of features used. In order to gain some insight into this phenomenon, we have compared two 2-class models, a  $hSVM_{0,5000}^2$  (test  $AROC = 0.39$ ) and  $Cntr_0$  (test  $AROC = 0.63$ ) trained on the same data split. For this training set, we found that there were 14,610 features occurring only in the negative class training instances (*NegOnly* features). Both models allocate non-positive (all negative for  $Cntr_0$ ) weights to such features. Our hypothesis is that for many of these features  $hSVM_{0,5000}^2$  allocate excessively low (highly negative) weights, which is an ‘easy way’ to minimise the margin errors. However, when some of these features occur in positive test examples, they push the scores of these examples excessively into negative direction, which causes a deterioration in the overall performance.

Figure 4 shows results corroborating this hypothesis. In Figure 4A we plot weights allocated to the *NegOnly* features, sorted in the reverse order of their magnitude, for each model separately and for each weight vector normalised to the unit length. Figure 4B shows probability of usage of these features in the positive class test examples (the curves are in fact 50-bin histograms). For both models the usage distribution is very similar and the most popular *NegOnly* features have the most negative weights. However, the weights from  $SVM$  for those most popular *NegOnly* features are about twice as large in magnitude as those for  $Cntr_0$  model. These weights result in excessive decreasing of scores of some positive test instances leading to deterioration in the overall performance.

**Persistent dominance of 1-class SVMs.** The above analysis is applicable to a high dimensional feature set. However, we have also observed in Section 3.3 that even in low dimensional spaces, this phenomenon of better performance with one-class learners persists. Our intuitive explanation here is that if the learner uses the minority class examples only, the “corner” (the half space)

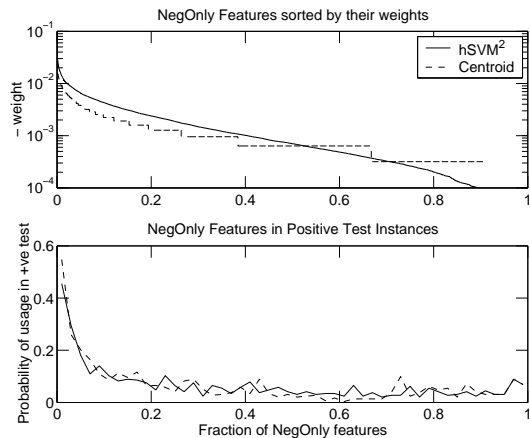


Figure 4: Understanding the influence of sparse high dimensional space on the solutions of 2-class learners. (A) Magnitude of weights for 2-class models for the *NegOnly* (features used only in the negative class in the training set) features in decreasing order of magnitude. (B) Usage of *NegOnly* features in the positive test set. Plots are shown for  $hSVM_{0,5000}^2$  (solid line) and *Centro* (dashed line).

where minority data resides is properly determined. However, when data from both classes is used, the minority class is “swamped” by the background class. In such a case the SVM solver seeking a “maximal margin” separation between classes, chooses a direction which is suboptimal in terms of AROC. The strange thing is that heavy discounting of the majority class by a factor  $B = 10^{-5}$  does not rectify this impact completely [10].

**Weakly informative features.** An alternative explanation for the relatively good performance of fringe classifiers is implied by experiments with Reuters data. We hypothesise that one factor is the relatively “weak” connection between the labels and the features in the case of AHR-data. Since the contrary is true for topic-based classification in Reuters, the superior performance with fringe classifiers is not evident until the most frequent features, which tend to be strongly

indicative of the labels for this dataset, are removed (Figure 3). Thus, we may expect fringe classifiers to work well in other real world applications with weak connection between labels and features.

**Importance of evaluation algorithm for feature selection.** An additional point regarding feature selection is that the performance of a any dedicated statistical system for that purpose is a function of both, the feature selection method and the learning strategy for evaluation of selection. For instance, all 1-class SVM and all centroid learners in Figure 2 perform very well with features selected by *ChiSqua* and *MutInfo*, while all 2-class learners, other than the 2-class centroid, perform poorly with the same features.

## 6 Conclusion

We have shown that SVMs even for a single kernel can split into a number of different modes, with dramatically different performance. Thus this popular class of learning machines cannot be treated as a monolithic black box, but should be viewed as a rich family of classifiers that need to be carefully tuned if top performance is required.

Further, some easy to implement fringe classifiers, such as centroids and positive 1-class SVMs, often outperform complicated 2-class SVMs. The very good performance of the fringe classifiers is related to sparsity of data and weak links between labels and features, and persists even after aggressive feature reduction. Thus, these classifiers could be used as baseline methods for biominning modelling in general and for machine evaluation of utility of various feature selections, in particular.

Finally, model-based feature selection technique can provide better results than dedicated feature pre-selection algorithm, facilitating development of more accurate discriminating models.

## Acknowledgements

The permission of the Managing Director, Telstra Research Laboratories, to publish this paper is gratefully acknowledged.

## References

- [1] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psych.*, 12:387 – 415, 1975.
- [2] R. Centor. The use of ROC curves and their analysis. *Med. Decis. Making*, 11:102 – 106, 1991.
- [3] Y. Chen, X. Zhou, and T. Huang. One-class svm for learning in image retrieval. In *Proceedings of IEEE International Conference on Image Processing (ICIP'01 Oral)*, 2001.
- [4] M. Craven. The Genomics of a Signaling Pathway: A KDD Cup Challenge Task. *SIGKDD Explorations*, 4(2), 2002.
- [5] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, 2000.
- [6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. In *Seventh International Conference on Information and Knowledge Management*, 1998.
- [7] G. Forman. Feature Engineering for a Gene Regulation Prediction Task. *SIGKDD Explorations*, 4(2), 2002.
- [8] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.
- [9] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the Tenth European Conference on Machine Learning ECML98*, 1998.
- [10] A. Kowalczyk and B. Raskutti. One Class SVM for Yeast Regulation Prediction. *SIGKDD Explorations*, 4(2), 2002.
- [11] M. A. Krogel, M. Denecke, M. Landwehr, and T. Scheffer. Combining Data and Text Mining Techniques for Yeast Gene Regulation Prediction: A Case Study. *SIGKDD Explorations*, 4(2), 2002.
- [12] L. M. Manevitz and M. Yousef. One-class SVMs for Document Classification. *Journal of Machine Learning Research*, 2:139–154, 2002.
- [13] J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1), (1986).
- [14] B. Raskutti, H. Ferrá, and A. Kowalczyk. Second Order Features for Maximising Text Classification Performance. In *Proceedings of the Twelfth European Conference on Machine Learning ECML01*, 2001.
- [15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, 1983.
- [16] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution, 1999.
- [17] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2001.
- [18] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [19] G. Whaba. Support vector machines, reproducing Hilbert spaces and the randomised GACV. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods*, pages 69–88, Cambridge, Ma., 1999. MIT Press.
- [20] Y. Yang and J. O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997.