

# Assessment of SVM Reliability for Microarrays Data Analysis

Andrea Malossini

Department of Information and Communication Technology  
University of Trento  
malossin@dit.unitn.it

Enrico Blanzieri

Department of Information and Communication Technology  
University of Trento  
blanzier@dit.unitn.it

Raymond T. Ng

Department of Computer Science  
University of British Columbia  
rng@cs.ubc.ca

## ABSTRACT

The goal of our research is to provide techniques that can assess and validate the results of SVM-based analysis of microarray data. We present preliminary results of the effect of mislabelled training samples. We conducted several systematic experiments on artificial and real medical data using SVMs. We systematically flipped the labels of a fraction of the training data. We show that a relatively small number of mislabelled examples can dramatically decrease the performance as visualized on the ROC graphs. This phenomenon persists even if the dimensionality of the input space is drastically decreased, by using for example feature selection.

## 1. INTRODUCTION

Gene-expression microarrays, commonly called gene-chips, make it possible to simultaneously measure the rate at which a cell or tissue is *expressing* (translating into a protein) each of its thousands of genes. One can use these comprehensive snapshot of biological activity to infer regulatory pathways in cells, identify novel targets for drug design, and improve the diagnosis, prognosis, and treatment planning for those suffering from disease. The amount of data this new technology produces is more than one can manually analyze. Thus, applying data mining techniques is necessary. However, while data mining techniques are proven successful for business applications, gene expression datasets have characteristics rather different from those of business datasets. We observe three key issues.

- First, the dimensionality of the data,  $p$ , can be very high. In the human genome, there are at least 30,000 genes. And in the human body, there are more than a million proteins. Thus, for one patient, there can quite easily be over 50,000 pieces of data.
- Second, the number of samples,  $n$ , can be small (relative to typical business applications). For many biomedical and pathology studies, 40–80 patients are considered decent-sized. Sample sizes in the order of hundreds are less common. There are a number of reasons why this is the case. First, data acquisition itself may be very expensive. While microarray costs are decreasing, other costs (e.g., wet laboratory cost for micro-dissection of tissues) remain high. For instance, the cost associated with one patient can very easily exceed 10,000 Euros. Money aside, the second reason is that for many diseases, there are simply not enough patients available. One prime example is early stage lung cancer (e.g., carcinoma-in-situ). Because early stage lung cancer is very hard to detect by normal pathological means (e.g., x-rays), we do not know of any medical research centre in the world which has a database of such patients exceeding 100. Finally, the third reason is that even if the patients are there, many of them or their families may not want to participate in research studies.
- Third, biomedical data can be very noisy. One reason is that data may be acquired in laboratory environment, which sometimes can be hard to keep unchanged. Another reason is that making diagnostic decisions (e.g., grading a biopsy) is not completely objective or black-and-white. For the same medical condition, there may be different so called gold-standards, which may lead to different decisions. Thus, robust techniques are very important.

In *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, held in conjunction with ECML/PKDD in Pisa, Italy. 24 September 2004.

Recently a state-of-the-art classification method, Support Vector Machine [3] has been used successfully in microarray data analysis [5, 10, 6]. Unfortunately microarray

datasets are characterized by the huge dimensionality of the input space  $p$  (which comprises thousands of genes) versus the extremely low number  $n$  of training samples (usually of the order of tens) as shown in Table 1. In such cases, a small error in the training set could result in a really poor-performance classifier.

The goal of our work is to assess the reliability of the results obtained by SVM techniques on microarray data. Here we present preliminary work that considers mislabelled training samples as a possible source of unreliability.

## 2. IS THE SVM RELIABLE FOR MICROARRAY DATASETS?

Initially, we started to investigate the problem of mislabelled samples on an artificial dataset and assess the performance of the classifier.

We generated a two-class classification problem with an input space of  $p = 2000$  features. The first class, labeled with “-1”, is sampled from a multivariate normal distribution with  $\mu = 0$  and  $\Sigma = I$ . The second class, labeled with “+1”, is distributed as the first class except for 7 features where the component of the mean is  $\mu_i = 4$ . This procedure has been adopted in order to simulate the differential expression of a limited number of genes. Sampling from the distributions described above, we generated a series of training sets with  $n = 10, 20, 30, 40, 50, 100, 200$  elements and a test set of 1000 elements. Each training set and test set has half of the elements labeled as “+1” and half labeled as “-1”. We trained a SVM on the training set and then we randomly flipped the labels of a fraction of the training set and trained other SVMs. We performed the flipping on the original training set for percentages of  $\{1, 2, 5, 10, 15, 25, 50\}\%$  (the number of flipping has been truncated to an integer). For each classifier we calculated the confusion matrix and the true positive rate (also called recall)

$$\text{TP rate} = \frac{\text{positives correctly classified}}{\text{total positives}}$$

and the false positive rate (also called false alarm rate)

$$\text{FP rate} = \frac{\text{negatives incorrectly classified}}{\text{total negatives}}$$

For each experiment identified by a value of  $n$  and a flipping percentage, the entire procedure has been repeated 20 times and mean and variance of each classifier have been calculated.

To assess the performance of each classifier obtained we used ROC graphs [7], where in the abscissa we have FP rate and in the ordinate TP rate. In ROC graphs, differently from ROC curves, a single classifier is visualized as a point. Instead of plotting a cloud of 20 points we plotted the mean over the 20 trials and the  $1-\sigma$  errors. In Fig.1(a)–1(b) we show the results of the simulation in a ROC graph for different percentage of label flipping for a training set of 20 elements and 50 elements respectively. It is evident that with only 10% of flipping the performance of the classifier is much lower than the correct classifier (0% flipping) and by incrementing the number of training samples, the variance decreases but the performance of the flipped classifiers are still low. In Fig.2 we show two examples of non-linear SVM classifiers, using a polynomial kernel of degree 2 and a radial basis kernel. The effect of the flipping is still there and in some way it is more accentuated. Usually a linear kernel

Source	$n$	$p$	$p$ after feature selection
West et al. [10]	49	7129	-
Golub et al. [5]	38	6817	-
Vapnik et al. [9]	38	6817	16
Alon et al. [2]	62	2000	-
Alizadeh et al. [1]	96	4026	-
Ramaswamy et al. [8]	76	16063	-
Furlanello et al. [4]	76	16063	315

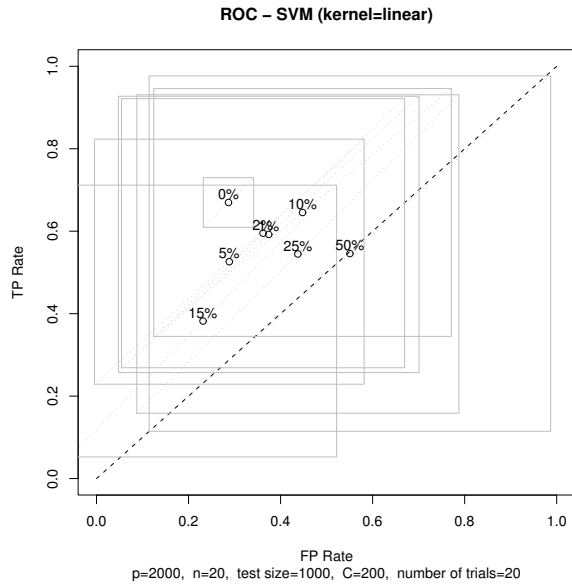
**Table 1: Number of features  $p$  and number of available samples  $n$  in microarray data analysis literature (“-” means no feature selection is performed in the paper).**

successes for microarray data analysis hence we concentrate on the linear kernel.

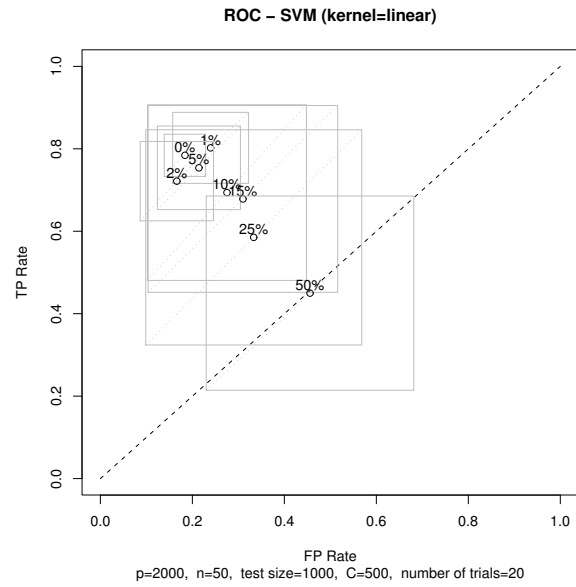
The problem of overfitting which arises when the number of features is much greater than the number of training samples can be lowered by reducing the number of features. Some feature selection techniques are used on SVMs, like RFE (recursive feature elimination) [9] and E-RFE (entropy-based feature elimination) [4], and they can reduce the number of features as shown in Tab. 1. To simulate the feature selection we reduce the number of features from 2000 to 200, where all the 7 expressed features are included in the 200 features. In Fig. 1(c)–1(d) we show the results of the simulation with a dataset of 200 features and training set of cardinality 20 and 50. Notice that even if the number of features is low,  $p = 200$ , given a low number of samples, the performance of the mislabelled classifiers are still poor. In particular, by observing the Fig. 1(d) we note that even if the variance now is much smaller, the discrepancy between the correct classifier (0%) and the 10% flipped classifier is still high. Furthermore if in the training set there are some mislabelled patterns, the error will propagate through the feature selection procedure so we expect, finally, to get the wrong set of important features.

Since a real dataset is far more complex than the synthetic data we generated, we tested the procedure on a real biological dataset, a human breast cancer dataset from [10], which included 49 samples, 24 marked as ER+ and 25 marked as ER-. We built randomly two training sets of 20 and 30 elements and test the SVM classifier on a disjoint random test set of 19 samples. In Fig. 3 we show the results on the Breast Cancer dataset using the SVM and randomly flipping a percentage of the original labels. Again with 10% of flipping the resulting classifier is worse than the correct classifier (0% flipping). This means that with only 2 or 3 wrong labels the classifier that we obtain a sensible decrease of the performance of the classifier.

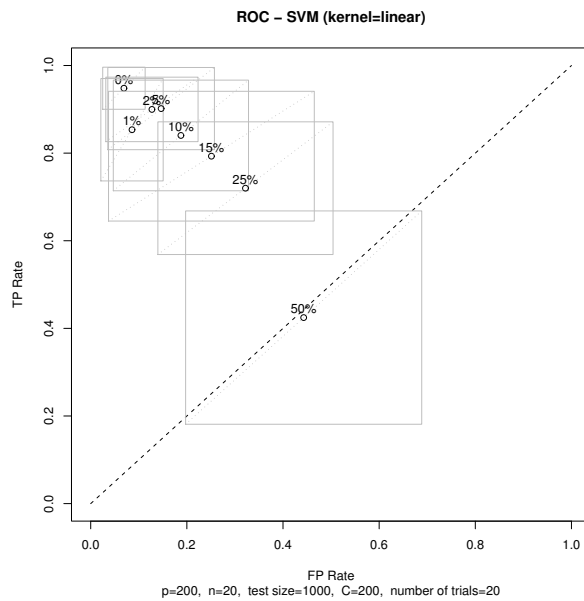
In all experiments with synthetic data the number of features which are expressed is 7. If we raise the number of expressed features to 100, we obtain a better noise-resistant classifier as shown in Fig. 4. A plausible reason for this behavior could be that when we increment the number of feature expressed, the Euclidean distance between the origin (“-1” labelled samples are distributed around the origin) and the point around which “+1” labelled samples are distributed is increased. Thus, the SVM hyperplane classifier has more room to move around the optimal position (0% of flipping) and also large percentage of flipping, for example 10%, does not influence too much the performance



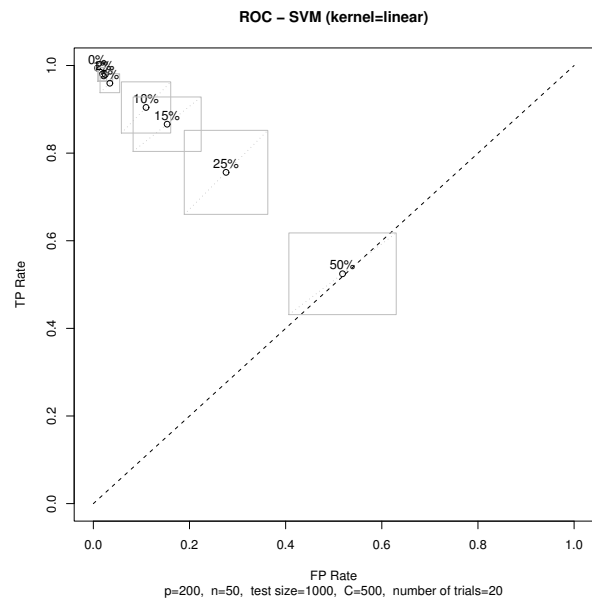
(a) Artificial dataset of 2000 features and 20 training samples. 7 features are expressed.



(b) Artificial dataset of 2000 features and 50 training samples. 7 features are expressed.

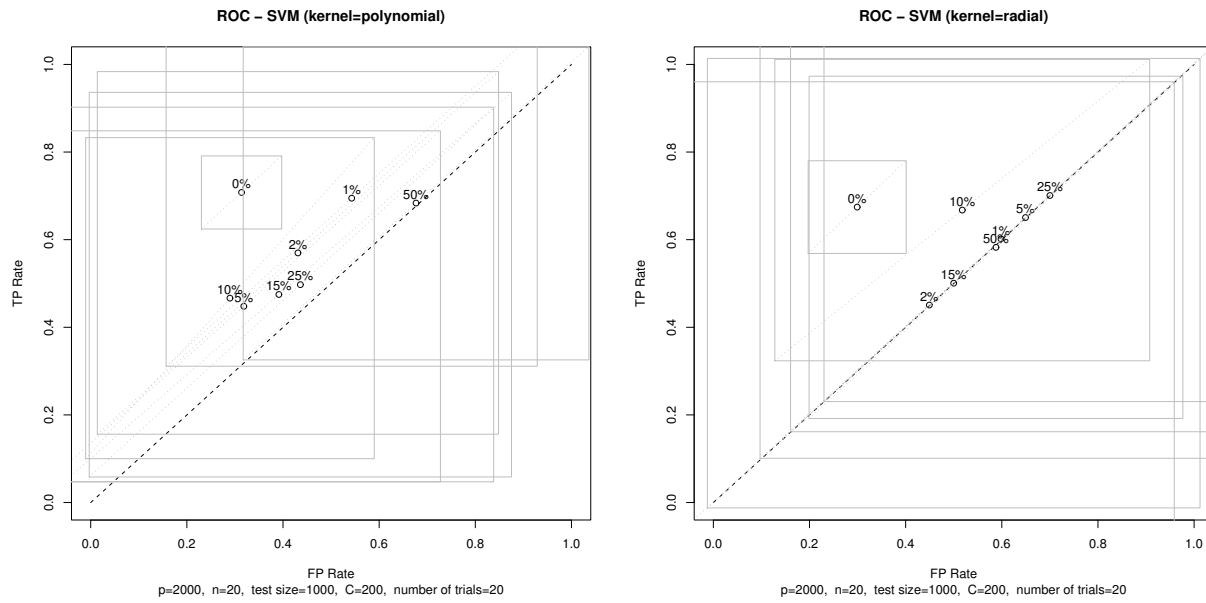


(c) Artificial dataset of 200 features and 20 training samples. 7 features are expressed.



(d) Artificial dataset of 200 features and 50 training samples. 7 features are expressed.

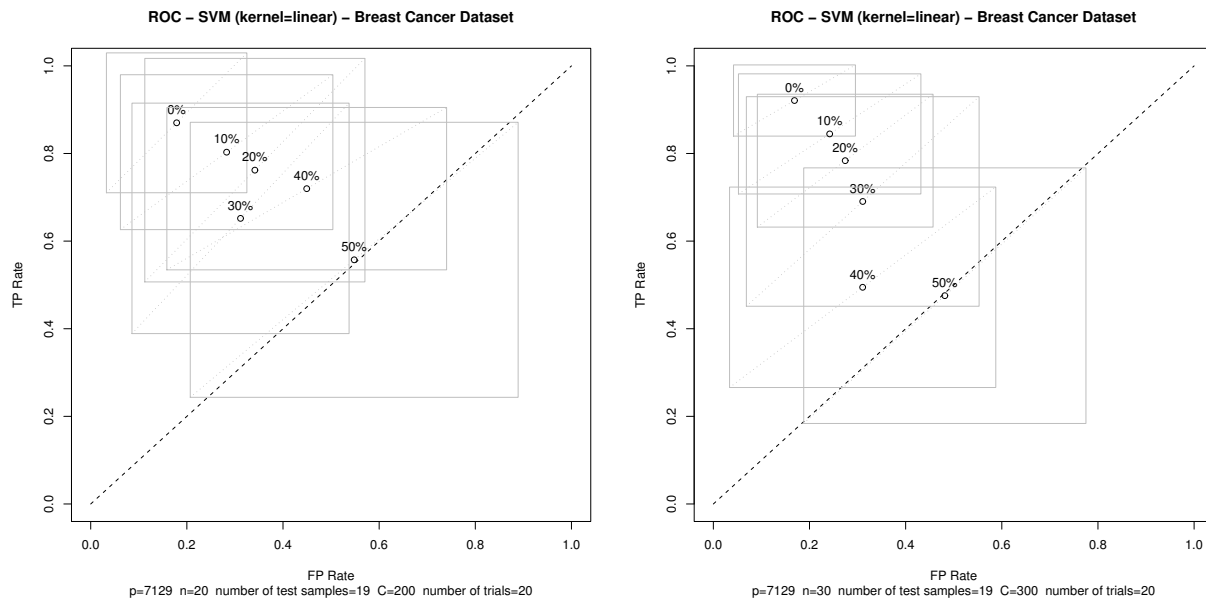
**Figure 1: ROC graphs of SVM classifiers. An incremental percentage of random flipping of the labels is performed and the SVM tested on a 1000-samples unflipped test set. Each experiment is repeated 20 times and mean and  $\sigma$ -error for TP and FP are plotted.**



(a) A polynomial kernel of degree 2 is used in the SVM classification.

(b) . A radial basis kernel is used in the SVM classification

**Figure 2: ROC graphs of SVM classifiers for some non-linear kernels. Artificial dataset of 2000 features and 20 training samples. 7 features are expressed.**



(a) Breast cancer dataset, 7129 features and 20 training samples

(b) Breast cancer dataset, 7129 features and 30 training samples

**Figure 3: ROC graphs of SVM classifiers for the breast cancer dataset. An incremental percentage of random flipping of the labels is performed and the SVM obtained tested on a 16-samples unflipped test set. Each experiment is repeated 20 times and mean and  $\sigma$ -error for TP and FP are plotted.**

