Classification of arrayCGH data using fused SVM.

Franck Rapaport, Emmanuel Barillot, Jean-Philippe Vert

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–9

Classification of arrayCGH data using a fused SVM

Franck Rapaport [1,2,3]*; Emmanuel Barillot [1,2,3] and Jean-Philippe Vert [1,2,3]

[1]Institut Curie, Centre de Recherche, Paris, F-75248 France.
[2]INSERM, U900, Paris, F-75248 France.
[3]Center for Computational Biology, Ecole des Mines de Paris, 35 rue saint Honore, 77305 Fontainebleau, France.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Array-based comparative genomic hybridization (arrayCGH) has recently become a popular tool to identify DNA copy number variations along the genome. These profiles are starting to be used as markers to improve prognosis or diagnosis of cancer, which implies that methods for automated supervised classification of arrayCGH data are needed. Like gene expression profiles, arrayCGH profiles are characterized by a large number of variables usually measured on a limited number of samples. However, arrayCGH profiles have a particular structure of correlations between variables, due to the spatial organization of BACs along the genome. This suggests that classical classification methods, often based on the selection of a small number of discriminative features, may not be the most accurate methods and may not produce easily interpretable prediction rules.

Results: We propose a new method for supervised classification of arrayCGH data. The method is a variant of support vector machine (SVM) that incorporates the biological specificities of DNA copy number variations along the genome as prior knowledge. The resulting classifier is a sparse linear classifier based on a limited number of regions automatically selected on the chromosomes, leading to easy interpretation and identification of discriminative regions of the genome. We test this method on three classification problems for bladder and uveal cancer, involving both diagnosis and prognosis. We demonstrate that the introduction of the new prior on the classifier leads not only to more accurate predictions, but also to the identification of known and new regions of interest in the genome.

Availability: All data and algorithms are publicly available.

Contact: franck.rapaport@curie.fr

1 INTRODUCTION

Genome integrity is essential to cell life and is ensured in normal cells by a series of checkpoints, which enable DNA repair or trigger cell death to avoid abnormal genome cells to appear. The p53 protein is probably the most prominent protein known to play this role. When these checkpoints are bypassed the genome may evolve and undergo alterations to a point where the cell can become premalignant and further genome alterations lead to invasive cancers.

This genome instability has been shown to be an enabling characteristic of cancer (Hanahan and Weinberg, 2000), and almost all cancers are associated with genome alterations. These alterations may be single mutations, translocations, or copy number variations (CNVs). A CNV can be a deletion or a gain of small or large DNA regions, an amplification, or an aneuploidy (change in chromosome number).

Many cancers present recurrent CNVs of the genome, like for example monoploidy of chromosome 3 in uveal melanoma (Speicher *et al.*, 1994), loss of chromosome 9 and amplification of the region of cyclin D1 (11q13) in bladder carcinomas (Blaveri *et al.*, 2005), loss of 1p and gain of 17q in neuroblastoma (Bown *et al.*, 2001; Van Roy *et al.*, 2002), EGFR amplification and deletion in 1p and 19q in gliomas (Idbaih *et al.*, 2007), or amplifications of 1q, 8q24, 11q13, 17q21-q23, and 20q13 in breast cancer (Yao *et al.*, 2006). Moreover associations of specific alterations with clinical outcome have been described in many pathologies (eg., Lastowska *et al.*, 1997).

Recently array-based comparative genomic hybridization (arrayCGH) has been developed as a technique allowing rapid mapping of CNVs of a tumor sample at a genomic scale (Pinkel *et al.*, 1998). The technique was first based on arrays using a few thousands of large insert clones (like BACs, and with a Mb range resolution) to interrogate the genome, and then improved with oligonucleotide based arrays consisting of several hundreds of thousands features, taking the resolution down to a few kb (Gershon, 2005). Many projects have since been launched to systematically detect genomic aberrations in cancer cells (van Beers and Nederlof, 2006; Chin *et al.*, 2006; Shing *et al.*, 2003).

The etiology of cancer and the advent of arrayCGH make it natural to envisage building classifiers for prognosis or diagnosis based on the genomic profiles of tumors. Building classifiers based on expression profiles is an active field of research, but little attention has been paid yet to genome-based classification. Chin *et al.* (2006) select a small subset of genes and apply a k-nearest neighbor classifier to discriminate between estrogen-positive and estrogen-negative patients, between high-grade patients and low-grade patients and between bad prognosis and good prognosis for breast cancer.

*to whom correspondence should be addressed

Jones *et al.* (2004) reduce the DNA copy number estimates to "gains" and "losses" at the chromosomal arm resolution, before using a nearest centroid method for classifying breast tumors according to their grade. As underlined in Chin *et al.* (2006), the classification accuracy reported in Jones *et al.* (2004) is better than the one reported in Chin *et al.* (2006), but still remains at a fairly high level with as much as 24% of misclassified samples in the balanced problem. This may be related to the higher resolution of the arrays produced by Jones *et al.* (2004). Moreover, the approach used by Jones *et al.* (2004) produces a classifier difficult to interpret as it is unable to detect any deletion or amplification that occur at the local level. O'Hagan *et al.* (2003) used a support vector machine (SVM) classifier using as variables all BAC ratios without any missing values. They were able to identify key CNAs.

The methods developed so far either ignore the particularities of arrayCGH and the inherent correlation structure of the data (O'Hagan *et al.*, 2003), or drastically reduce the complexity of the data at the risk of filtering out useful information (Jones *et al.*, 2004; Chin *et al.*, 2006). In all cases, a reduction of the complexity of the data or a control of the complexity of the predictor estimated is needed to overcome the risk of overfitting the training data, given that the number of probes that form the profile is often several orders of magnitude larger than the number of samples available to train the classifier.

In this paper we propose a new method for supervised classification, specifically designed for the processing of arrayCGH profiles. In order not to miss potentially relevant information that may be lost if the profiles are first processed and reduced to a small number of homogeneous regions, we estimate directly a linear classifier at the level of individual probes. Yet, in order to control the risk of overfitting, we define a prior on the linear classifier to be estimated. This prior encodes the hypothesis that (i) many regions of the genome should not contribute to the classification rule (sparsity of the classifier), and (ii) probes that contribute to the classifier should be grouped in regions on the chromosomes, and be given the same weight within a region. This a priori information helps reducing the search space and produces a classification rule that is easier to interpret. This technique can be seen as an extension of SVM where the complexity of the classifier is controlled by a penalty function similar to the one used in the fused lasso method to enforce sparsity and similarity between successive features (Tibshirani *et al.*, 2005). We therefore call the method a *fused SVM*. It produces a linear classifier that is piecewise constant on the chromosomes, and only involves a small number of loci without any a priori regularisation of the data. From a biological point of view, it avoids the prior choice of recurrent regions of alterations, but produces *a posteriori* a selection of discriminant regions which are then amenable to further investigations.

We test the fused SVM on several public datasets involving diagnosis and prognosis applications in bladder and uveal cancer, and compare it with a more classical method involving feature selection without prior information about the organization of probes on the genome. In a cross-validation setting, we show that the classification rules obtained with the fused SVM are systematically more accurate than the rules obtained with the classical method, and that they are also more easily interpretable.

2 METHODS

In this section we present an algorithm for the supervised classification of arrayCGH data. This algorithm, which we call *fused SVM*, is motivated by the linear ordering of the features along the genome and the high dependancy in behaviour of neighbouring features. The algorithm itself estimates a linear predictor by borrowing ideas from recent methods in regression, in particular the *fused lasso* (Tibshirani *et al.*, 2005). We start by a rapid description of the arrayCGH technology and data, before presenting the fused SVM in the context of regularized linear classification algorithms.

2.1 ArrayCGH data

ArrayCGH is a microarray-based technology that allows the quantification of the DNA copy number of a sample at many positions along the genome in a single experiment. The array contains thousands to millions of spots, each of them consisting of the amplified or synthesized DNA of a particular region of the genome. The array is hybridized with the DNA extracted from a sample of interest, and in most cases with (healthy) reference DNA. Both samples have first been labelled with two different fluorochromes, and the ratio of fluorescence of both fluorochromes is expected to reveal the ratio of DNA copy number at each position of the genome. The log-ratio profiles can then be used to detect the regions with abnormalities (log-ratio significantly different of 0), corresponding to gains (if the log-ratio is significantly superior to 0) or losses (if it is significantly inferior to 0).

The typical density of arrayCGH ranges from 2400 BAC features in the pioneering efforts, corresponding to one approximately 100 kb probe every Mb (Pinkel *et al.*, 1998), up to millions today, corresponding to one 25 to 70bp oligonucleotide probe every few kb, or even tiling arrays (Gershon, 2005).

There are two principal ways to represent arrayCGH data: as a log-ratio collection, or as a collection of status (lost, normal or gained, usually represented as -1, 0 and 1 which correspond to the sign of the log ratio). The status representation has strong advantages over the log-ratio as it reduces the complexity of the data, provides the scientist with a direct identification of abormalities and allows the straightforwad detection of recurrent alterations. However, converting ratios into status is not always obvious and often implies a loss of information which can be detrimental to the study: for several reasons such as heterogeneity of the sample or contamination with healthy tissue (which both result in cells with different copy numbers in the sample), the status may be difficult to infer from the data, whereas the use of the ratio values avoids this problem. Another problem is the low subtelty of statuses. In particular, if we want to use arrayCGH for discriminating between two subtypes of tumors or between tumors with different future evolution, all tumors may share the same important genomic alterations that are easily captured by the status assignment while differences between the types of tumors may be characterized by more subtle signals that would disappear should we transform the log ratio values into statuses. Therefore, we consider below an arrayCGH profile as a vector of log-ratios for all probes in the array.

2.2 Classification of arrayCGH data

While much effort has been devoted to the analysis of single arrayCGH profiles, or populations of arrayCGH profiles in order to detect genomic alterations shared by the samples in the population, we focus on the supervised classification of arrayCGH. The typical problem we want to solve is, given two populations of arrayCGH data corresponding to two populations of samples, to design a classifier that is able to predict

which population any new sample belongs to. This paradigm can be applied for diagnosis or prognosis applications, where the populations are respectively samples of different tumor types, or with different evolution. Although we only focus here on binary classification, the techniques can be easily extended to problems involving more than two classes using, for example, a series of binary classifiers trained to discriminate each class against all others.

While accuracy is certainly the first quality we want the classifier to have in real diagnosis and prognosis application, it is also important to be able to interpret it and understand what the classification is based on. Therefore we focus on linear classifiers, which associate a weight to each probe and produce a rule that is based on a linear combination of the probe log-ratios. The weight of a probe roughly corresponds to its contribution in the final classification rule, and therefore provides evidence about its importance as a marker to discriminate the populations. It should be pointed out, however, that when correlated features are present, the weight of a feature is not directly related to the individual correlation of the feature with the classification, hence some care should be taken for the interpretation of linear classifier.

In most applications of arrayCGH classification, it can be expected that only a limited number of regions on the genome should contribute to the classification, because most parts of the genome may not differ between populations. Moreover, the notion of discriminative regions suggest that a good classifier should detect these regions, and typically be piecewise constant over them. We show below how to introduce these prior hypotheses into the linear classification algorithm.

2.3 Linear supervised classification

Let us denote by p the number of probes hybridized on the arrayCGH. The result of an arrayCGH competitive hybridization is then a vector of p log-ratios, which we represent by a vector x in the vector space $\mathcal{X} = \mathbb{R}^p$ of possible arrayCGH profiles. We assume that the samples to be hybridized can belong to two classes, which we represent by the labels -1 and $+1$. The classes typically correspond to the disease status or the prognosis of the samples. The aim of binary classification is to find a decision function that can predict the class $y \in \{-1, +1\}$ of a data sample $x \in \mathcal{X}$. Supervised classification uses a database of samples $x_1, ..., x_n \in \mathcal{X}$ for which the labels $y_1, ..., y_n \in \{-1, +1\}$ are known in order to construct the prediction function. We focus on *linear* decision functions, which are defined by functions of the form $f(x) = w^\top x$ where $w^\top$ is the transpose of a vector $w \in \mathbb{R}^d$. The class prediction for a profile x is then $+1$ if $f(x) \geq 0$, and -1 otherwise. Training a linear classifier amounts to estimating a vector $w \in \mathbb{R}^d$ from prior knowledge and the observation of the labeled training set.

The training set can be used to assess whether a candidate vector w can correctly predict the labels on the training set; one may expect such a w to correctly predict the classes of unlabeled samples as well. This induction principle, sometimes referred to as *empirical risk minimization*, is however likely to fail in our situation where the dimension of the samples (the number of probes) is typically larger than the number of training points. In such a case, many vectors w can indeed perfectly explain the labels of the training set, without capturing any biological information. These vectors are likely to poorly predict the classes of new samples. A well-known strategy to overcome this overfitting issue, in particular when the dimension of the data is large compared to the number of training points available, is to look for *large-margin* classifiers constrained by *regularization* (Vapnik, 1998). A large-margin classifier is a prediction function $f(x)$ that not only tends to produce the correct sign (positive for labels $+1$, negative for class -1), but also tends to produce large absolute values. This can be formalized by the notion of *margin*, defined as $yf(x)$: large-margin classifiers try to predict the class of a sample with large margin. Note that the prediction is correct if the margin is positive. The margin can be thought of as a measure of confidence in the prediction given by the sign of f, so a large margin is synonymous with a large confidence. Training a large-margin classifier means estimating a function f that takes large margin values on the training set. However, just like for the sign of f, if $p > n$ then it is possible to find vectors w that lead to arbitrarily large margin on all points of the training set. In order to control this overfitting, large-margin classifiers try to maximize the margin of the classifier on the training set under some additional constraint on the classifier f, typically that w is not too "large". In summary, large-margin classifiers find a trade-off between the objective to ensure large margin values on the training set, on the one hand, and that of controlling the complexity of the classifier, on the other hand. The balance in this trade-off is typically controlled by a parameter of the algorithm.

More formally, large-margin classifiers typically require the definition of two ingredients:

- A *loss function* $l(t)$ that is "small" when $t \in \mathbb{R}$ is "large". From the loss function one can deduce the *empirical risk* of a candidate vector w, given by the average loss function applied to the margins of w on the training set:

$$R_{emp}(w) = \frac{1}{n} \sum_{i=1}^{n} l(y_i w_i^\top x) \,. \tag{1}$$

  The smaller the empirical risk, the better w fits the training set in the sense of having a large margin. Typical loss functions are the hinge loss $l(t) = \max(0, 1 - t)$ and the logit loss $l(t) = \log\left(1 + e^{-t}\right)$.
- A *penalty function* $\Omega(w)$ that measures how "large" or how "complex" w is. Typical penalty functions are the L_1 and L_2 norms of w, defined respectively by $||w||_1 = \sum_{i=1}^{p} |w_i|$ and $||w||_2 = \left(\sum_{i=1}^{p} w_i^2\right)^{\frac{1}{2}}$.

Given a loss function l and a penalty function Ω, large-margin classifiers can then be trained on a given training set by solving the following constrained optimization problem:

$$\min_{w \in \mathbb{R}^p} R_{emp}(w) \text{ subject to } \Omega(w) \leq \mu \,, \tag{2}$$

where μ is a parameter that controls the trade-off between fitting the data, i.e., minimizing $R_{emp}(f)$, and monitoring the regularity of the classifier, i.e., monitoring $\Omega(w)$. Examples of large-margin classifiers include the support vector machine (SVM) and kernel logistic regression (KLR) obtained by combining respectively the hinge and logit losses with the L_2 norm penalization function (Cortes and Vapnik, 1995; Boser *et al.*, 1992; Vapnik, 1998), or the 1-norm SVM when the hinge loss is combined with the L_1 loss .

The final classifier depends on both the loss function and the penalty function. In particular, the penalty function is useful to include prior knowledge or intuition about the classifier one expects. For example, the L_1 penalty function is widely used because it tends to produce sparse vectors w, therefore performing an automatic selection of features. This property has been successfully used in the context of regression (Tibshirani, 1996), signal representation (Chen *et al.*, 1998), survival analysis (Tibshirani, 1997), logistic regression (Genkin *et al.*, 2007; Krishnapuram *et al.*, 2004), or multinomial logistic regression (Krishnapuram *et al.*, 2005), where one expects to estimate a sparse vector.

2.4 Fused lasso

Some authors have proposed to design specific penalty functions as a means to encode specific prior informations about the expected form of the final classifier. In the context of regression applied to signal

processing, when the data is a time series, Land and Friedman (1996) propose to encode the expected positive correlation between successive variables by choosing a regularisation term that forces successive variables of the classifier to have similar weights. More precisely, assuming that the variables $w_1, w_2, \ldots, w_p$ are sorted in a natural order where many pairs of successive values are expected to have the same weight, they propose the *variable fusion* penalty function:

$$\Omega_{fusion}(w) = \sum_{i=1}^{n-1} |w_i - w_{i+1}| \,. \tag{3}$$

Plugging this penalty function in the general algorithm (2) enforces a solution w with many successive values equal to each others, that is, tends to produce a piecewise constant weight vector. In order to combine this interesting property with a requirement of sparseness of the solution, Tibshirani *et al.* (2005) proposed to combine the lasso penalty and the variable fusion penalty into a single optimization problem with two constraints, namely:

$$\begin{aligned} & \min_{w \in \mathbb{R}^n} R_{emp}(w) \\ \text{under the constraints} \quad & \sum_{i=1}^{n-1} |w_i - w_{i+1}| \leq \mu \\ & \|w\|_1 \leq \lambda \,, \end{aligned} \tag{4}$$

where λ and μ are two parameters that control the relative trade-offs between fitting the training data (small R_{emp}), enforcing sparsity of the solution (small λ) and enforcing the solution to be piecewise constant (small μ). When the empirical loss is the mean square error in regression, the resulting algorithm is called *fused lasso*. This method was illustrated in Tibshirani *et al.* (2005) with examples taken from gene expression datasets and mass spectrometry. Later, Tibshirani and Wang (2007) proposed a tweak of the fused lasso for the purpose of signal smoothing, and illustrated it for the problem of discretising noisy CGH profiles.

2.5 Fused SVM

Remembering from Section 2.2 that for arrayCGH data classification one typically expects the "true" classifier to be sparse and piecewise constant along the genome, we propose to extend the fused lasso to the context of classification and adapt it to the chromosome structure for arrayCGH data classification. The extension of fused lasso from regression to large-margin classification is obtained simply by plugging the fused lasso penaly constraints into a large-margin empirical risk in (4). In what follows we focus on the empirical risk (1) obtained from the hinge loss, which leads to a simple implementation as a linear program (see Section 2.6 below). The extension to other convex loss functions, in particular the logit loss function, results in convex optimization problems with linear constraints that can be solved with general convex optimization solvers (Boyd and Vandenberghe, 2004).

In the case of arrayCGH data, a minor modification to the variable fusion penalty (3) is necessary to take into account the structure of the genome in chromosomes. Indeed, two successive spots on the same chromosome are prone to be subject to the same amplification and are therefore likely to have similar weights on the classifier; however, this positive correlation is not expected across different chromosomes. Therefore we restrict the pairs of successive features appearing in the function constraint (3) to be consecutive probes on the same chromosome.

We call the resulting algorithm a *fused SVM*, which can be formally written as the solution of the following problem:

$$\begin{aligned} & \min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \max(0, 1 - y_i w^\top x_i) \\ \text{under the constraints} \quad & \sum_{i \sim j} |w_i - w_j| \leq \mu \\ & \sum_{i=1} |w_i| \leq \lambda \,, \end{aligned} \tag{5}$$

where $i \sim j$ if i and j are the indices of succesive spots of the same chromosome. As with fused lasso, this optimisation problem tends to produce classifiers w with similar weights for consecutive features, while maintaining its sparseness. This algorithm depends on two paramters, λ and μ, which are typically chosen via cross-validation on the training set. Decreasing λ tends to increase the sparsity of w, while decreasing μ tends to enforce successive spots to have the same weight.

This classification algorithm can be applied to CGH profiles, taking the ratios as features. Due to the effect of both regularisation terms, we obtain a sparse classification function that attributes similar weights to successive spots.

2.6 Implementation of the fused SVM

Introducing slack variables, the problem described in (5) is equivalent to the following linear program :

$$\begin{aligned} \min_{w,\alpha,\beta,\gamma} \sum_{i=1}^{n} \alpha_i \quad & \text{under the following constraints :} \\ \forall i = 1, ..., n \quad & \alpha_i \geq 0 \\ \forall i = 1, ..., n \quad & \alpha_i \geq 1 - w^\top x_i y_i \\ & \sum_{i=1}^{n} \beta_i \leq \lambda \\ \forall i = 1, ..., p \quad & \beta_i \geq w_i \\ \forall i = 1, ..., p \quad & \beta_i \geq -w_i \\ & \sum_{k=1}^{q} \gamma_k \leq \mu \\ \forall i, j \text{ such that } i \sim j \quad & \gamma_k \geq w_i - w_j \\ \forall i, j \text{ such that } i \sim j \quad & \gamma_k \geq w_j - w_i \end{aligned} \tag{6}$$

In our experiments, we implemented and solved this problem using Matlab and the SeDuMi 1.1R3 optimisation toolbox (Sturm, 1999).

3 DATA

We consider two publicly available arrayCGH datasets for cancer research, from which we deduce three problems of diagnosis and prognosis to test our method.

The first dataset contains arrayCGH profiles of 57 bladder tumor samples (Stransky *et al.*, 2006). Each profile gives the relative quantity of DNA for 2215 spots. We removed the probes corresponding to sexual chromosomes, because the sex mismatch between some patients and the reference used makes the computation of copy number less reliable, giving us a final list of 2143 spots. We considered two types of tumor classification: either by grade, with 12 tumors of grade 1 and

45 tumors of higher grades (2 or 3) or by stage, with 16 tumors of stage Ta and 32 tumors of stage T2+. In the case of stage classification, 9 tumors with intermediary stage T1 were excluded from the classification.

The second dataset contains arrayCGH profiles for 78 melanoma tumors that have been arrayed on 3750 spots (Trolet *et al.*, 2008). As for the bladder cancer dataset, we excluded the sexual chromosomes from the analysis, resulting in a total of 3649 spots. 35 of these tumors lead to the development of liver metastases within 24 months, while 43 did not. We therefore consider the problem of predicting, from an arrayCGH profile, whether or not the tumor will metastasize within 24 months.

In both datasets, we replaced the missing spots log-ratios by 0. In order to assess the performance of a classification method, we performed a cross-validation for each of the three classification problems, following a leave-one-out procedure for the bladder dataset and a 10-fold procedure for the melanoma dataset. We measure the number of misclassified samples for different values of parameters λ and μ.

4 RESULTS

In this section, we present the results obtained with the fused SVM on the datasets described in the previous section. As a baseline method, we consider a L_1-SVM which minimizes the mean empirical hinge loss suject to a constraint on the L_1 norm of the classifier in (2). The L_1-SVM performs automatic feature selection, and a regularization parameter λ controls the amount of regularization. It has been shown to be a competitive classification method for high-dimensional data, such as gene expression data (Zhu *et al.*, 2004). In fact the L_1-SVM is a particular case of our fused SVM, when the μ parameter is chosen large enough to relax the variable fusion constraint (3), typically by taking $\mu > 2\lambda$. Hence by varying μ from a large value to 0, we can see the effect of the variable fusion penalty on the classical L_1-SVM.

4.1 Bladder tumors

The upper plot of Figure 1 show the estimated accuracy (by Leave One Out (LOO)) of the fused SVM as a function of the regularization parameters λ and μ, for the classification by grade of the bladder tumors. The middle plot of Figure 1 represents the best linear classifier found by the L_1-SVM (corresponding to $\lambda = 256$), while the lower plot shows the linear classifier estimated from all samples by the fused SVM when λ and μ are set to values that minimise the LOO error, namely $\lambda = 32$ and $\mu = 1$. Similarly, Figure 2 shows the same results (LOO accuracy, L_1-SVM and fused SVM classifiers) for the classification of bladder tumors according to their stage.

In both cases, when μ is large enough to make the variable fusion inactive in (5), then the classifier only finds a compromise between the empirical risk and the L_1 norm of the classifier. In other words, we recover the classical L_1 SVM with parameter λ. Graphically, the performance of the L_1 SVM for varying λ can be seen on the upper side of each plot of the LOO accuracy in Figures 1 and 2. Interestingly, in both cases we observe that the best performance obtained when both λ and μ can be adjusted is much better than the best performance of the L_1 SVM, when only λ can be adjusted. In the case of grade classification, the number of misclassified samples drops from 12 (21%) to 7 (12%), while in the case of stage classification it drops from 13 (28%) to 7 (15%). This suggests that the additional constraint that translates our prior knowlege about the structure of the spot positions on the genome is beneficial in terms of classifier accuracy.

As expected, there are also important differences in the visual aspects of the classifiers estimated by the L_1-SVM and the fused SVM. The fused SVM produces sparse and piecewise constant classifiers, amenable to further investigations, while it is more difficult to isolate from the L_1-SVM profiles the key features used in the classification, apart from a few strong peaks.

As we can see by looking at the shape of the fused SVM classifier in Figure 1, the grade classification function is characterised by non-null constant values over a few small chromosomal regions and numerous larger regions. Of these regions, a few are already known as being altered in bladder tumors, such as the gain on region 1q (Corson *et al.*, 2005). Moreover some of them have already been shown to be correlated with grade, such as chromosome 7 (Waldman *et al.*, 1991).

On the contrary, the stage classifier is characterised by only a few regions with most of them involving large portions of chromosomes. They concern mainly chromosome 4, 7, 8q, 11p, 14, 15, 17, 20, 21 and 22, with in particular a strong contribution from chromosomes 4, 7 and 20. These results on chromosomes 7, 8q, 11p and 20 are in good agreement with Blaveri *et al.* (2005) who identified the most common alterations according to tumor stage on a set of 98 bladder tumors.

4.2 Melanoma tumors

Similarly to Figures 1 and 2, the three plots in Figure 3 show respectively the accuracy, estimated by 10-fold cross-validation, of the fused SVM as a function of the regularisation parameters λ and μ, the linear classifier estimated by the L_1-SVM when λ is set to the value that minimizes the estimated error ($\lambda = 4$), and the linear classifier estimated by a fused SVM on all samples when λ and μ are set to values that minimise the 10-fold error, namely $\lambda = 64$ and $\mu = 0.5$.

Similarly to the bladder study, the performance of the L_1-SVM without the fusion constraint can be retrieved by looking at the upper part of the plot of Figure 3. The fused classifier offers a slightly improved performance compared to the standard L_1-SVM (17 errors (22%) versus 19 errors (24%)), even though the amelioration seems more marginal compared to the improvement made with bladder tumors and the misclassification rate remains fairly high.

As for the bladder datasets, the L_1-SVM and fused SVM classifiers are markedly different. The L_1-SVM classifier is based only on a few BAC concentrated on chromosome 8, with positive weights on the 8p arm and negative weights on the 8q arm. These features are biologically relevant, and correspond to a known genomic alterations (loss of 8p and gain of 8q in metastatic tumors). The presence of a strong signal concentrated on chromosome 8 for the prediction of metastasis

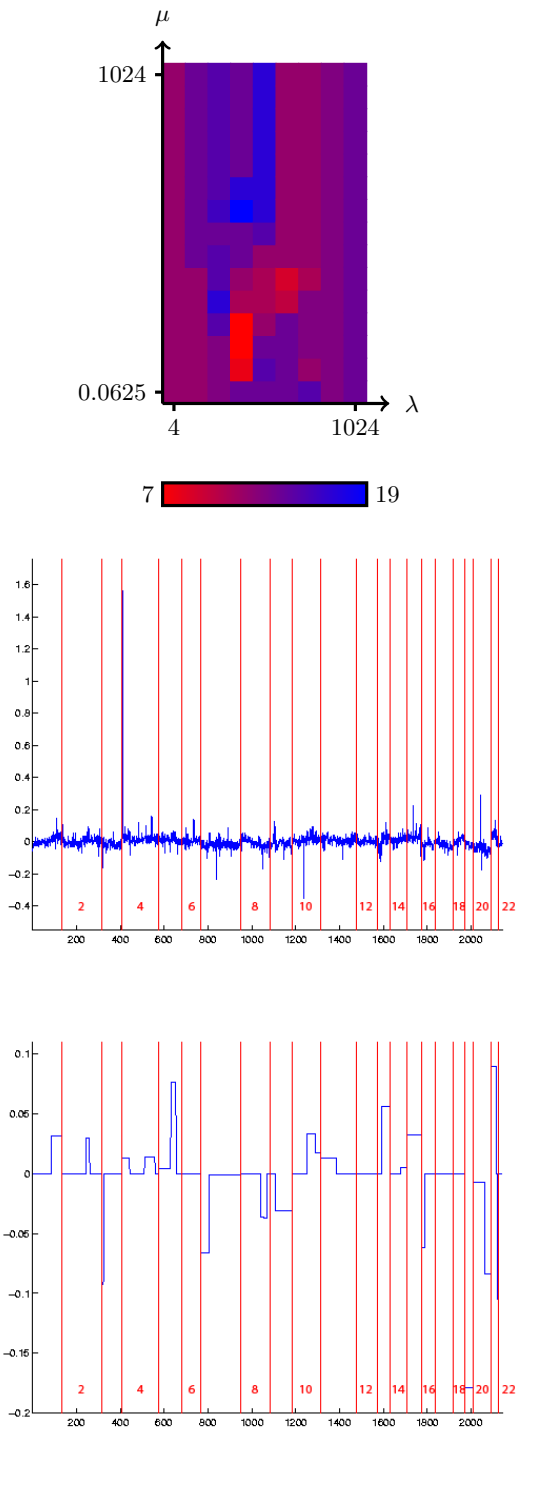

Fig. 1. The figure on the upper side represents the number of misclassified samples in a leave-one-out error loop on the bladder cancer dataset with the grade labelling, with its color scale for different values of the parameters λ and μ which vary logarithmically along the axes. The weights of the best classifier, for classical L_1-SVM (middle) and for fused SVM (lower part) are ordered and represented in a blue line, annotated with the chromosome separation (red line).

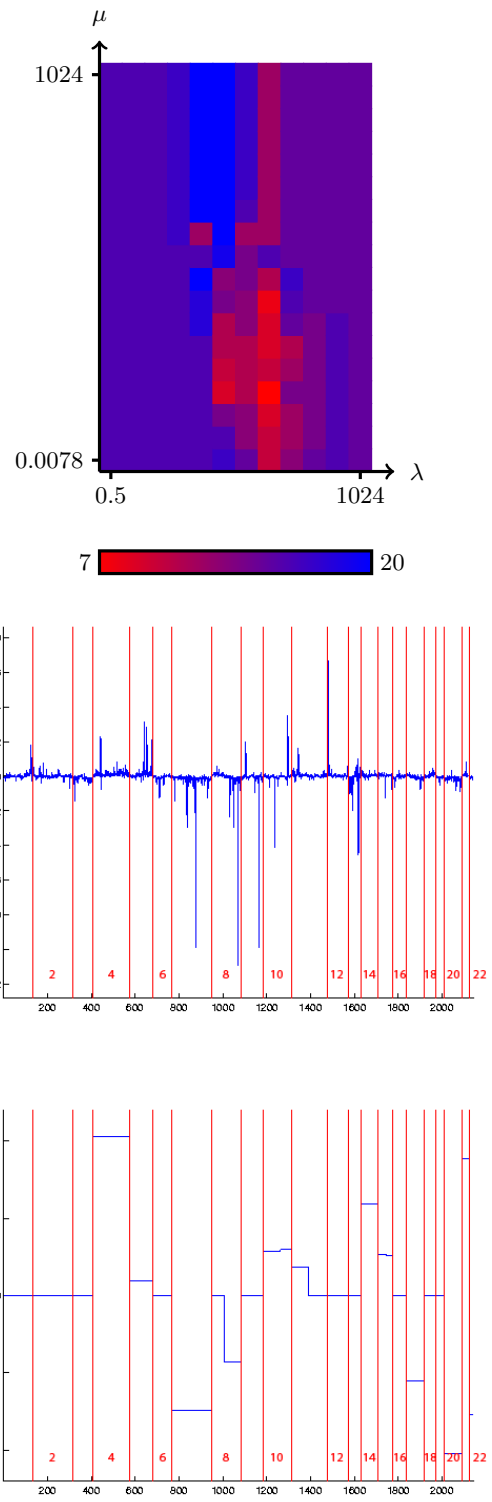

Fig. 2. The figure on the upper side represents the number of misclassified samples in a leave-one-out error loop on the bladder cancer dataset with the stage labelling, with its color scale, for different values of the parameters λ and μ which vary logarithmically along the axes. The weights of the best classifier, for classical L_1-SVM (middle) and for fused-SVM (lower part) are ordered and represented in a blue line, annotated with the chromosome separation (red line).

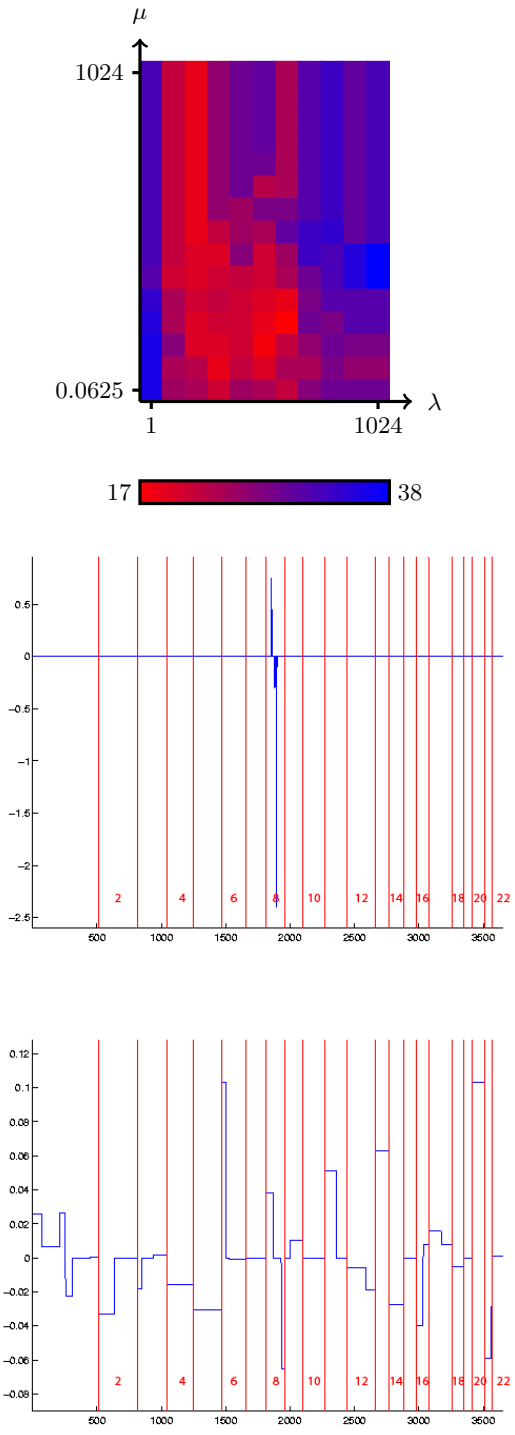

Fig. 3. The figure on the upper part represents the number of misclassified samples in a ten-fold error loop on the melanoma dataset. The weights of the best classifier, for classical L_1-SVM (middle) and for fused SVM (lower part) are ordered and represented in a blue line, annotated with the chromosome separation (red line).

is in this case correctly captured by the sparse L_1-SVM, which explains its relatively good performance.

To the contrary, the fused SVM classifier is characterised by many CNAs, most of them involving large regions of chromosomes. Interestingly, we retrieve the regions whose alteration was already reported as recurrent events of uveal melanoma: chromosomes 3, 1p, 6q, 8p, 8q, 16q. As expected the contributions of 8p and 8q are of opposite sign, in agreement with the common alterations of these regions: loss of 8p and gain of 8q in etastatic tumors. Interestingly the contribution of chromosome 3 is limited to a small region of 3p, and does not involve the whole chromosome as the frequency of chromosome 3 monosomy would have suggested. Note that this is consistent with works by Parrella *et al.* (2003) and Tschentscher *et al.* (2001) who delimited small 3p regions from partial chromosome 3 deletion patients. On the other hand we also observe that large portions of other chromosomes have been assigned significant positive or negative weights, such as chromosomes 1p, 2p, 4, 5, 9q, 11p, 12q, 13, 14, 20, 21. To our knowledge, they do not correspond to previous observations, and may therefore provide interesting starting points for further investigations.

5 DISCUSSION

We have proposed a new method for the supervised classification of arrayCGH data. Thanks to the use of a particular regularization term that translates our prior assumptions into constraints on the classifier, we estimate a linear classifier that is based on a restricted number of spots, and gives as much as possible equal weights to spots located near each other on a chromosome. Results on real data sets show that this classification method is able to discriminate between the different classes with a better performance than classical techniques that do not take into account the specificities of arrayCGH data. Moreover, the learned classifier is piecewise constant and therefore lends itself particularly well to further interpretation, highlighting in particular selected chromosomal regions with particularly highly positive or negative weights.

From the methodological point of view, the use of regularized large-scale classifiers is nowadays widely spread, especially in the SVM form. Regularization is particularly important for "small n large p" problems, i.e., when the number of samples is small compared to the number of dimensions. An alternative interpretation of such classifiers is that they correspond to maximum *a posteriori* classifiers in a Bayesian framework, where the prior over classifier is encoded in our penalty function. It is not surprising, then, that encoding prior knowledge in the penalty function is a mathematically sound strategy that can be strongly beneficial in terms of classifier accuracy, in particular when few training samples are available. The accuracy improvements we observe on all classification datasets confirm this intuition. Besides the particular penalty function investigated in this paper, we believe our results support the general idea that engineering relevant priors for a particular problem can have important effects on the quality of the function estimated and paves the way for further research

on the engineering of such priors in combination with large-margin classifiers. As for the implementation, we solved a linear program for each value couple of the regularization parameters λ and μ, but it would be interesting to generalize the recent works on path following algorithms to be able to follow the solution of the optimization problem when λ and μ vary (Efron *et al.*, 2004).

Another interesting direction of future research concerns the combination of heterogeneous data, in particular of arrayCGH and gene expression data. Gene expression variations contain indeed information complementary to CNV for the genetic aberrations of the dysfunctioning cell (Stransky *et al.*, 2006), and their combination is therefore likely to both improve the accuracy of the classification methods and shed new light on biological phenomena that are characteristic of each class. A possible strategy to combine such datasets would be to train a large-margin classifier with a particular regularization term that should be adequately designed.

ACKNOWLEDGEMENT

We thank Jérôme Couturier, Sophie Piperno-Neumann and Simon Saule, and the uveal melanoma group from Institut Curie. We are also grateful to Philippe Hupé for his help in preparing the data. This project was partly funded by the ACI IMPBIO Kernelchip and the EC contract ESBIC-D (LSHG-CT-2005-518192). FR and EB are members of the team "Systems Biology of Cancer", Equipe labellisée par la Ligue Nationale Contre le Cancer.

REFERENCES

Blaveri, E., Brewer, J. L., Roydasgupta, R., Fridlyand, J., DeVries, S., Koppie, T., Pejavar, S., Mehta, K., Carroll, P., Simko, J. P., and Waldman, F. M. (2005). Bladder Cancer Stage and Outcome by Array-Based Comparative Genomic Hybridization. *Clin Cancer Res*, **11**(19), 7012–7022.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA. ACM.

Bown, N., Lastowska, M., Cotterill, S., O'Neill, S., Ellershaw, C., Roberts, P., Lewis, I., Pearson, A. D., Group, U. C. C., and the U.K. Children's Cancer Study Group (2001). 17q gain in neuroblastoma predicts adverse clinical outcome. u.k. cancer cytogenetics group and the u.k. children's cancer study group. *Med Pediatr Oncol*, **36**(1), 14–19.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Chen, S. S., Donoho, D. L., and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**(1), 33–61.

Chin, S.-F., Wang, Y., Thorne, N. P., Teschendorff, A. E., Pinder, S. E., Vias, M., Naderi, A., Roberts, I., Barbosa-Morais, N. L., Garcia, M. J., Iyer, N. G., Kranjac, T., Robertson, J. F. R., Aparicio, S., Tavare, S., Ellis, I., Brenton, J. D., and Caldas, C. (2006). Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*, **26**(13), 1959–1970.

Corson, T. W., Huang, A., Tsao, M.-S., and Gallie, B. L. (2005). Kif14 is a candidate oncogene in the 1q minimal region of genomic gain in multiple cancers. *Oncogene*, **24**(30), 4741–4753.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.*, **32**(2), 407–499.

Genkin, Alexander, Lewis, David, D., Madigan, and David (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, **49**(3), 291–304.

Gershon, D. (2005). Dna microarrays: More than gene expression. *Nature*, **437**(7062), 1195–1198.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**(1), 57–70.

Idbaih, A., Marie, Y., Lucchesi, C., Pierron, G., Manie, E., Raynal, V., Mosseri, V., Hoang-Xuan, K., Kujas, M., Brito, I., Mokhtari, K., Sanson, M., Barillot, E., Aurias, A., Delattre, J.-Y., and Delattre, O. (2007). Bac array cgh distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. *Int J Cancer*.

Jones, C., Ford, E., Gillett, C., Ryder, K., Merrett, S., Reis-Filho, J. S., Fulford, L. G., Hanby, A., and Lakhani, S. R. (2004). Molecular Cytogenetic Identification of Subgroups of Grade III Invasive Ductal Breast Carcinomas with Different Clinical Outcomes. *Clin Cancer Res*, **10**(18), 5988–5997.

Krishnapuram, B., Hartemink, A. J., Carin, L., and Figueiredo, M. A. T. (2004). A bayesian approach to joint feature selection and classifier design. *IEEE Trans Pattern Anal Mach Intell*, **26**(9), 1105–11.

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(6), 957–968.

Land, S. R. and Friedman, J. H. (1996). Variable fusion: A new adaptive signal regression method. Technical Report.

Lastowska, M., Nacheva, E., McGuckin, A., Curtis, A., Grace, C., Pearson, A., and Bown, N. (1997). Comparative genomic hybridization study of primary neuroblastoma tumors. united kingdom children's cancer study group. *Genes Chromosomes Cancer*, **18**(3), 162–169.

O'Hagan, R. C., Brennan, C. W., Strahs, A., Zhang, X., Kannan, K., Donovan, M., Cauwels, C., Sharpless, N. E., Wong, W. H., and Chin, L. (2003). Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. *Cancer Res*, **63**(17), 5352–5356.

Parrella, P., Fazio, V. M., Gallo, A. P., Sidransky, D., and Merbs, S. L. (2003). Fine mapping of chromosome 3 in uveal melanoma: Identification of a minimal region of deletion on chromosomal arm 3p25.1-p25.2. *Cancer Res*, **63**(23), 8507–8510.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W., and Albertson, D. G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20**(2), 207–211.

Shing, D. C., McMullan, D. J., Roberts, P., Smith, K., Chin, S.-F., Nicholson, J., Tillman, R. M., Ramani, P., Cullinane, C., and Coleman, N. (2003). Fus/erg gene fusions in ewing's tumors. *Cancer Res*, **63**(15), 4568–4576.

Speicher, M., Prescher, G., du Manoir, S., Jauch, A., Horsthemke, B., Bornfeld, N., Becher, R., and Cremer, T. (1994). Chromosomal gains and losses in uveal melanomas detected by comparative genomic hybridization. *Cancer Res.*, **54**(14), 3817–23.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S. G. D., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., Graham, A., Southgate, J., Asselain, B., Allory, Y., Abbou, C. C., Albertson, D. G., Thiery, J. P., Chopin, D. K., Pinkel, D., and Radvanyi, F. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nat Genet*, **38**(12), 1386–1396.

Sturm, J. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, **11–12**, 625–653. Special issue on Interior Point Methods (CD supplement with software).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, **58**(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat Med*, **16**(4), 385–395.

Tibshirani, R. and Wang, P. (2007). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal Of The Royal Statistical Society Series B*, **67**(1), 91–108. available at http://ideas.repec.org/a/bla/jorssb/v67y2005i1p91-108.html.

Trolet, J., Hupe, P., Huon, I., Lebigot, I., Mariani, P., Plancher, C., Asselain, B., Desjardins, L., Delattre, O., Sastre-Garau, X., Thiery, J.-P., Saule, S., Piperno-Neumann, S., Barillot, E., and Couturier, J. (2008). Genomic profiling and identification of high risk tumors in uveal melanoma by array-cgh analysis of

primary tumors and liver metastases. *submitted to Cancer Res*.

Tschentscher, F., Prescher, G., Horsman, D. E., White, V. A., Rieder, H., Anastassiou, G., Schilling, H., Bornfeld, N., Bartz-Schmidt, K. U., Horsthemke, B., Lohmann, D. R., and Zeschnigk, M. (2001). Partial deletions of the long and short arm of chromosome 3 point to two tumor suppressor genes in uveal melanoma. *Cancer Res*, **61**(8), 3439–3442.

van Beers, E. and Nederlof, P. (2006). Array-cgh and breast cancer. *Breast Cancer Research*, **8**(3), 210.

Van Roy, N., Vandesompele, J., Berx, G., Staes, K., Gele, M. V., Smet, E. D., Paepe, A. D., Laureys, G., van der Drift, P., Versteeg, R., Roy, F. V., and Speleman, F. (2002). Localization of the 17q breakpoint of a constitutional 1;17 translocation in a patient with neuroblastoma within a 25-kb segment located between the accn1 and tlk2 genes and near the distal breakpoints of two microdeletions in neurofibromatosis type 1 patients. *Genes, Chromosomes and Cancer*, **35**(2), 113–120.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.

Waldman, F. M., Carroll, P. R., Kerschmann, R., Cohen, M. B., Field, F. G., and Mayall, B. H. (1991). Centromeric copy number of chromosome 7 is strongly correlated with tumor grade and labeling index in human bladder cancer. *Cancer Res*, **51**(14), 3807–3813.

Yao, J., Weremowicz, S., Feng, B., Gentleman, R. C., Marks, J. R., Gelman, R., Brennan, C., and Polyak, K. (2006). Combined cDNA Array Comparative Genomic Hybridization and Serial Analysis of Gene Expression Analysis of Breast Tumor Progression. *Cancer Res*, **66**(8), 4065–4078.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004). 1-norm support vector machines. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Adv. Neural. Inform. Process Syst.*, volume 16, Cambridge, MA. MIT Press.