Multiclass classification and gene selection with a stochastic algorithm

Kim-Anh Lê Cao [a,b,*] Agnès Bonnet [b] Sébastien Gadat [a]

[a] *Institut Mathématiques de Toulouse, Université de Toulouse and CNRS (UMR 5219)*

[b] *Station d'Amélioration Génétique des Animaux, Institut National de Recherche Agrononomique, 31326 Castanet Tolosan, France*

Abstract

Microarray technology allows for the monitoring of thousands of gene expressions in various biological conditions, but most of these genes are irrelevant for classifying these conditions. Feature selection is consequently needed to help reduce the dimension of the variable space. We start from the application of the stochastic algorithm "Optimal Feature Weighting" (OFW) for selecting features in various classification problems. This method does not depend on the classification method. Gadat and Younes (2007) who established the theoretical part of the model, applied SVM in the framework of binary pattern recognition data sets. The application with CART was performed in Lê Cao et al. (2007) who made a comparative study with other binary wrapper methods in the context of microarray data and emphasized on the biological interpretation.
In this study, we focus on the multiclass problem that wrapper methods rarely handle. From a computational point of view, one of the main difficulties comes from the commonly unbalanced classes situation when dealing with microarray data. From a theoretical point of view, very few methods have been developped to minimize any classification criterion, compared to the 2-class situation (*e.g* SVM, l_0SVM, RFE...). In this paper, we first develop the OFW approach to handle multiclass problems using CART and one-vs-one SVM as classifiers. We then compare our results with those obtained with other multiclass selection algorithm (Random Forests and the filter method F-test), on four public microarray data sets. We assess the statistical relevancy of the results by measuring and comparing their performances . The aim of this study is to heuristically evaluate which method would be the best to select genes distinguishing minority classes. Application and biological interpretation are then given in the case of a real world microarray data set.

Key words: Feature Selection, Stochastic Algorithm, Classification, Unbalanced Multiclass Microarray

Introduction

When dealing with microarray data, one of the most important issues to improve the classification task is to perform feature selection. Thousands of genes can be measured on a single array, most of which are irrelevant or uninformative for discriminative methods and dimensionality thus must be reduced without losing information.
In this context, our objective was to look for predictors (the genes) that would classify the observed cases (the microarrays) into their known classes. The selection of these discriminative variables can be performed in two ways: either explicitly (filter methods) or implicitly (wrapper methods). The filter methods measure the usefulness of a feature by ordering it with statistical tests such as t- or F-tests. These gene-by-gene approaches are robust against overfitting and computationally fast. However, they disregard the interactions between the features and may fail to find the "useful" set of variables: they usually select variables with redundant information. On the other hand, the aim of the wrapper methods is to measure the usefulness of a subset of features in the set of variables. However, when dealing with a large number of variables as it will be the case here, it is computationally impossible to do an exhaustive search among all subsets of features and these methods are proned to overfit. One solution to benefit from the wrapper approach is to perform a search using stochastic approximations that still cover a large portion of the feature space to avoid local minima. The "Optimal Feature Weighting" algorithm (OFW) proposed by Gadat and Younes (2007) allows the selection of an optimal discriminative subset of variables. This meta algorithm can be applied independently with any classifier. Classifiers such as Support Vector Machines (SVM, Vapnik (2000)) and Classification And Regression Trees (CART, Breiman et al. (1984)) were passed up to this stochastic meta-algorithm in Lê Cao et al. (2007)) for 2-class microarray problems. The aim was to make a comparative study of OFW+SVM/CART with other wrapper methods (Recursive Feature Elimination, Guyon et al. (2002), l_0 norm SVM, Weston et al. (2003), Random Forests, Breiman (2001)) and the filter method t-test on public microarray data sets. The relevancy of the results was assessed statistically by measuring the performance of each gene selection, and with a biological expertise in the context of the biological experiment. The results showed that the selections made with OFW were statistically competitive and biologically relevant, even in the case of complex data sets.
From this point, we investigate this stochastic algorithm with multiclass microarray data sets. Multiclass problems are often considered as an extension of 2-class problems. However this extension is not always straightforward as the data sets are often characterized by unbalanced classes with a small number of observations in at least one of the classes. Furthermore, this "rare" class is

* Kim-Anh.Le-Cao@toulouse.inra.fr Tel: +33 561285574; Fax:+33 561285353

often the one of interest for the biologists who would like to diagnose a disease for example. Nevertheless, most algorithms do not perform well for such problems as they aim to minimize the overall error rate instead of focusing on the minority class. Moreover, the classification accuracy appears to degrade very quickly as the number of classes increases. Several methods have been proposed in the recent years. Chen et al. (2004) proposed balanced or weighted random forests, McCarthy et al. (2005) compared sampling methods and cost sensitive learning with however no clear winner in the results. In the context of multiclass microarray data, Li et al. (2004) applied various classifiers with various feature selection methods and conclude that the accuracy is highly dependent on the choice of the classifier, rather than the choice of the selection method, although this would be more natural. Chen et al. (2003) applied four filter methods with low correlation between selected genes, Yeung and Bumgarner (2003) applied uncorrelated or error-weighted Shrunken Centroid.
In this study we compare two ways of handling multiclass data: with or without an internal weighting procedure in OFW. We do not intend to optimize the size of the gene subset. We rather focus on the assessment criteria to measure the performance of the different methods on the first selected genes.
Biological interpretation that is one of the main key to evaluate the relevancy of the biological results will not be given in this paper when analysing the four public data sets, but the reader can refer to Lê Cao et al. (2007) that gives some clues on the biological interpretation importance.
We apply the multicategory classifier CART and the one-vs-one SVM approach with OFW on four public microarray data sets. Numerical comparisons are drawn with Random Forests, known to perform efficiently on such data sets, and one filter method (F-tests), by computing the e.632+ bootstrap error from Efron and Tibshirani (1997) for each feature selection method, by computing the stability of the results (Jaccard Index) and by comparing the different gene lists. Then, the weighted and no weighted approaches are compared in OFW+CART and OFW+SVM with the same tools.
Finally, application and biological analysis are performed on a real world data set.
The first section introduces the theoretical adaptation of the OFW model to the multiclass framework. In next section we consider the computational aspects of the application of CART and SVM in OFW and describe the different tools to assess the performance of the results. Application on the public data sets and on a practical data set follow. The paper ends with further elements of discussion.

1 The model

We introduce our model of feature selection in the framework of multiclass analysis. As we focus here on microarray data, we will mostly refer to "genes" instead of "variables".

1.1 Measure of the classification efficiency

Let $\mathcal{G}$ be a large set of genes numbered from 1 to N that describes a signal $\mathcal{I}$ to belonging to one of the classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$, $k = 1, .., K$. A classification algorithm $\mathbb{A}$ will be chosen according to the problem type (2-class, multiclass), as OFW does not depend on the classification procedure $\mathbb{A}$.
Let us define a positive weight parameter $\mathbb{P}$ on each of the genes in $\mathcal{G}$. After a normalization step, $\mathbb{P}$ can be considered as a discrete probability on the N genes. The goal is to learn a probability that fits the efficiency of each gene for the classification of $\mathcal{I}$ in $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$, so that important weights are given to genes with high discriminative power and lower weights to those that have a poorest influence on the classification task. Denote p any small integer compared to N, a gene subset of size p has to be extracted from $\mathcal{G}$ using $\mathbb{P}$. We then define how to measure the goodness of $\mathbb{P}$ for the set of genes $\mathcal{G}$ and the classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ (*i.e* the objective function).

Definition 1 *Given a probability $\mathbb{P}$ on $\mathcal{G}$ and $\epsilon(\omega)$ the measure of classification efficiency with any p-uple $\omega \in \mathcal{G}^p$, the energy of the system at point $\mathbb{P}$ is the mean classification performance where ω is drawn with respect to $\mathbb{P}^{\otimes p}$ in $\mathcal{G}^p$*

$$\mathcal{E}(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\epsilon] = \sum_{\omega \subset \mathcal{G}^p} \mathbb{P}(w)\epsilon(w) \tag{1}$$

Remark 1 *Remark here that genes selected with respect to $\mathbb{P}$ in (1) are drawn with replacement although it looks more reasonable to use subsets of genes without replacement. This mainly comes from the mathematical derivations to optimize $\mathcal{E}$ that will be described below.*

Note that the energy $\mathcal{E}$ depends on the way we measure the classification efficiency on ω, that we denote $\epsilon(\omega)$. Given any standard classification algorithm $\mathbb{A}$, $\epsilon(\omega)$ will actually be the error rate of $\mathbb{A}$ computed on the training set using the set of extracted features ω. The more $\mathbb{P}$ enables us to hold good genes g for classification (important weight on g and $\epsilon(\omega)$ small each time ω contains this gene g), the less $\mathcal{E}$. Minimizing $\mathcal{E}$ with respect to $\mathbb{P}$ will thus permit to exhibit the most weighted and consequently the most highly discriminative genes. Hence, a natural importance ranking will be read on the weight $\mathbb{P}^\star$ minimizing $\mathcal{E}$.

The energy $\mathcal{E}$ can be minimized with a stochastic version of the standard gradient descent technique. More details about the theoretical derivations can be found in Gadat and Younes (2007).
The function $\mathcal{E}$ has to be minimized up to the constraints defined by a discrete probability measure on $\mathcal{G}$. Thus, the more natural way to optimize (1) is to use a gradient descent of $\mathcal{E}$ projected to the set of constraints. The set of constraints $\mathcal{S}$ is the simplex of probability map on $\mathcal{G}$. We also denote by $\Pi_{\mathcal{S}}$ the Affine projection of any point of $\mathbb{R}^N$ on the simplex $\mathcal{S}$. This natural projection $\Pi_{\mathcal{S}}$ of any point x can be computed in a finite number of steps as mentioned in Gadat and Younes (2007). Using this former projection $\Pi_{\mathcal{S}}$, the Euclidean gradient of $\mathcal{E}$ is

$$\forall g \in \mathcal{G} \qquad \nabla\mathcal{E}(\mathbb{P})(g) = \sum_{\omega \in \mathcal{G}^p} \frac{C(\omega, g)\mathbb{P}(\omega)}{\mathbb{P}(g)}\epsilon(\omega) \tag{2}$$

where $C(\omega, g)$ is the number of occurrences of g in ω. The iterative procedure to update $\mathbb{P}$ is then given by

$$\mathbb{P}_{t+dt} = \mathbb{P}_t - \nabla\mathbb{P}_t dt \tag{3}$$

The main clue is that the Euclidian gradient expression (2) can be seen as an expectation as stated in the next proposition.

Proposition 1 *For any $\mathbb{P}$ probability map on $\mathcal{G}$ and if $\nabla_{\mathcal{S}}$ denotes the gradient of $\mathcal{E}$ with respect to constraints $\mathcal{S}$, $\nabla_{\mathcal{S}}\mathcal{E}$ is given by*

$$\forall g \in \mathcal{G} \qquad \nabla_{\mathcal{S}}\mathcal{E}(\mathbb{P})(g) = \Pi_{\mathcal{S}}\left(\mathbb{E}_\omega\left[\frac{C(\omega, g)}{\mathbb{P}(g)}\epsilon(\omega)\right]\right).$$

This last expression is numerically untractable since it requires the computation of ϵ over all possible p-uple of $\mathcal{G}$. To deal with such gradient, a computable Robbins-Monro algorithm can be used, which gets similar asymptotic behavior as (3) (see for instance Gadat and Younes (2007), Kushner and Clark (1978)). With this stochastic method, the updated formula of $\mathbb{P}_n$ becomes:

$$\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}\left[\mathbb{P}_n - \alpha_n \frac{C(\omega_n, .)\epsilon(\omega_n)}{\mathbb{P}_n(.)}\right] \tag{4}$$

where ω_n is any set of p genes sampled with respect to $\mathbb{P}_n$. Note that the last expression is always defined since when $\mathbb{P}_n(g) = 0$ as we cannot draw this gene in ω_n and the integer $C(\omega_n, g)$ vanishes. The next theorem precisely describes the asymptotic behavior of (4).

Theorem 1 *Defining the discretisation time* $\tau_k = \sum_{i=0}^{k} \alpha_i$ *and its associated dual reversion* $I(t) = \sup\{k \quad | \quad \tau_k \leq t\}$, *then the interpolated process* $P^k(t) = \mathbb{P}_{I(\tau_k+t)}$ *is an asymptotic pseudo-trajectory of the ordinary differential equation (3) provided that the sequence of steps* (α_i) *satisfies the two conditions:*

$$\sum_i \alpha_i = \infty \qquad \textit{and} \qquad \exists\, \nu > 0 \quad \sum_i \alpha_i^{1+\nu} < \infty.$$

This last result insures that the stochastic algorithm computing $\mathbb{P}_n$ is asymptotically equivalent to the real gradient descent (3). Several derivations of this theoretical point can be found in Gadat and Younes (2007). In our experiments, we have decided to use a step sequence $\alpha_i = A/(B+i)$ for calibrated constants A and B.

1.3 Detailed algorithm.

We detail the application of the algorithm in the case of a given classifier $\mathbb{A}$:

Let $\mathcal{G} = (\delta_1 \dots \delta_{|\mathcal{G}|})$, $\mu \in \mathbb{N}^*$ and η the stopping criterion.
- For iteration $n = 0$ define $\mathbb{P}_0$ as the uniform distribution on $\mathcal{G}$
- While $|\mathbb{P}_{(n+\mu)} - \mathbb{P}_n|_\infty > \eta$:
  - extract ω_n from $\mathcal{G}^p$ with respect to $\mathbb{P}_{n,p} = \mathbb{P}_n^{\otimes p}$
  - construct $\mathbb{A}_{\omega_n}$ and compute $\epsilon(\omega_n)$
  - compute the drift vector $d_n = C(\omega_n, \cdot)\epsilon(\omega_n)/\mathbb{P}_n(\cdot)$
  - update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n d_n]$
  - $n = n + 1$

2 Application of OFW and performance evaluation

We discuss here the applications in the field of multiclass problems. The application of OFW+CART and the comparisons of OFW+CART/SVM in the binary case can be found in Lê Cao et al. (2007).

2.1 CART and SVM multiclass applied to OFW

CART

OFW is applied with the multiclass classifier CART (Classification And Regression Trees Breiman et al. (1984)) that is well adequate for multiclass problems. CART is constructed via a recursive partitioning routine. It builds

a classification rule to predict the class label of the microarrays based on the feature information following the Gini criterion. To avoid overfitting, trees are then generally pruned using a cross validation procedure. In our special case, the trees were not pruned and a node was declared terminal when all the observations landing in this node belonged to the same class.
Note that CART is unstable by nature: a slight change in the features can lead to a very different construction of the tree. Following the example of Breiman (1996), the trees were aggregated (*bagging*) to overcome this instability. As in Breiman, the trees were unpruned, but there is no overfitting, thanks to the aggregation technique.
To compute the efficiency criterion ϵ at iteration n we launched B trees on B bootstrap samples on different ω_n^b drawn with respect to $\mathbb{P}_n$, where $b = 1, \ldots, B$. We then defined ϵ as the mean classification error rate on the out-of-bag samples. The detailed bagging version of OFW+CART is described in 2.3.
Computations were first run on a cluster using the R [1] packages **Rmpi** and **rpart**. An R package is currently being implemented (written in C for faster computation).

SVM Multiclass

We applied OFW with the one-vs-one SVM approach that is implemented in the **e1071** R package. Other SVM multiclass approaches could have been applied, such as the one-vs-all approach, the approach proposed by Duan and Keerthi (2005), by Lee and Lee (2002) or by Joachims (1999). Unlike CART, SVM is very stable and ϵ was hence computed on only one bootstrap sample ($B = 1$).

2.2 Different computations of the gradient

In contrary to Gadat and Younes (2007), we made some slight modifications of the gradient descent to improve the speed of the algorithm with OFW+CART. We propose an averaged time version of the initial OFW as follows:

$$D_n = \frac{\sum_{i=1}^{n} \alpha_i \bar{d}_i}{\sum_{i=1}^{n} \alpha_i} \quad \text{with} \quad \bar{d}_i = \sum_{b=1}^{B} \frac{C(\omega_i^b, .)\epsilon(\omega_i^b)}{\mathbb{P}_i(.)},$$

where b is the bootstrap sample on which each CART tree is constructed and $\alpha_i = A/(B + i)$ is the step sequenced refered in section 1.2.

[1] The Comprehensive R Archive Network, http://cran.r-project.org/

This enables the stochastic algorithm to better approximate the mean drift (2) than in the standard case. With CART, the approximation of $\nabla\mathcal{E}$ is actually much more difficult than in the SVM case since the variance of the stochastic algorithm seems higher using CART classifier. This averaging step is hence crucial for the algorithm.

2.3 Detailed OFW+CART algorithm

Here is the detailed version of OFW+CART with bagging.

Let $\mathcal{G} = (\delta_1 \dots \delta_{|\mathcal{G}|})$, $\mu \in \mathbb{N}^*$ and η the stopping criterion. $\mathbb{A}$ is the unpruned classifier CART.

- For iteration $n = 0$ define $\mathbb{P}_0$ as the uniform distribution on $\mathcal{G}$
- While $|\mathbb{P}_{(n+\mu)} - \mathbb{P}_n|_\infty > \eta$:
  - For $b = 1..B$
    - extract ω_n^b from $\mathcal{G}^p$ with respect to $\mathbb{P}_{n,p} = \mathbb{P}_n^{\otimes p}$
    - draw a bootstrap sample b_{samp} and construct $\mathbb{A}_{\omega_n^b}^{b_{samp}}$
    - compute $\epsilon(\omega_n^b)$ on the out-of-bag sample $\bar{b}_{samp}$
  - compute the averaged drift vector D_n as in 2.2
  - update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n d_n]$
  - $n = n + 1$

2.4 Weighting procedure

An efficient way to take into account the unbalanced characteristics of the data set is to weight the internal error rate $\epsilon(\omega)$ according to the number samples of each class in the learning set. This would penalize a classification error made on the minority class and hence put more weight on the variables that help classify this class instead of the majority class.
Let n the total number of cases and m_k, $k = 1..K$ the number of cases in class k. We define the (normalized) weight of *each* observation in class k by $w_k = \frac{1}{m_k \times K}$.
Then for each out-of-bag test observation (the sample not drawn in the bootstrap sample), we note mis_k the number of misclassified observations from class k and the weighted internal error rate is defined as:

$$\epsilon(\omega) = \sum_{k=1}^{K} mis_k \times w_k$$

instead of $\frac{\sum_k^K mis_k}{n}$ in the no weighting case. This weighting procedure also stands for the evaluation step, see following section 2.5.

2.5 Performance measurement

Comparison of the prediction performance

Error rates of all methods on each data set were computed with the e.632+ bootstrap error estimate from Efron and Tibshirani (1997) that is adequate for small sample sizes data sets. Each algorithm will be learned on a bootstrap sample to avoid any overfitting during the gene selection evaluation (Ambroise and McLachlan (2002)). However, note that this performance evaluation does not dictate the optimal number of genes to select. e.632+ only allows to compare the performances of the different selection methods.

Stability

One can define the feature stability as the level of agreement between the set of selected genes chosen in each boostrap sample with the set of selected genes using the full training set. The Jaccard index (Yeung and Bumgarner (2003)) then computed lies between 0 (low level of agreement) and 1 (high level of agreement) and will be used to compare the stability of four all ranking methods.

Definition 2 *Let $S(\Delta)$ be the set of the Δ selected genes from the entire training set and $S(nb, \Delta)$ the set of selected genes from the nb bootstrap sample. The number of true positives (TP) is the number of selected genes that were chosen in both $S(\Delta)$ and $S(nb, \Delta)$:*

$$TP = |S(\Delta) \cap S(nb, \Delta)|.$$

Similarly, we define the false positives (FP) the number of selected genes that were chosen in $S(nb, \Delta)$ but not in $S(\Delta)$:

$$FP = |S(nb, \Delta) \setminus_{S(\Delta)}|,$$

and the number of false negatives (FN) the number of genes that were selected in $S(\Delta)$ but not in $S(nb, \Delta)$:

$$FN = |S(\Delta) \setminus_{S(nb,\Delta)}|.$$

The Jaccard index $J(nb, \Delta)$ is defined as $TP/(TP + FP + FN)$ and is high and close to 1 when there are many true positives and few false positives and false negatives. We then compute the averaged Jaccard index J_Δ over all nb samples for Δ varying between 1 selected gene and Δ_{max} selected genes.

We expect therefore to rank stability of each feature selection procedure with this Jaccard index.

2.6 Ranking methods

Multicategory ranking methods are still rare in the context of classification, especially in microarray data context. A comparative study is performed with the well-known Random Forests (RF, Breiman (2001)). The three wrapper methods (OFW+CART, OFW+SVM and RF) were also compared to the F-test filter method, that is still widely used for selecting genes in the context of microarrays.
Although Random Forests can also be performed with a weighting approach such as Balanced Random Forests (BRF) or Weighted Random Forests (WRF) from Chen et al. (2004), we chose to compare all these methods with no weighting procedure.

3 Results and discussion on public data sets

A short description of the four public data sets is first given. We then compare the results obtained with OFW+CART, OFW+SVM, RF and F-test with no weighting procedure, and the F-test selection was evaluated with a one-vs-one linear SVM.
We finally focus on OFW and compare the weighted *vs.* non-weighted procedure and give some elements of discussion.

3.1 Multiclass data sets

We present the results obtained on four public multiclass data sets.

(1) Lymphoma (Alizadeh et al., 2000) compares 3 classes of cells (42, 9 and 11 observations per class) with 4026 gene expressions.
(2) The 3-class Leukemia version (Golub et al. (1999)) with 7129 genes compares the lymphocytes B and T in ALL (Acute Lymphoblastic Leukemia, 38 and 9 observations) and the AML class (Acute Myeloid Leukemia, 25 observations). The classes AML-B and AML-T are biologically very similar.
(3) The Small Round Blue-Cell Tumor Data of childhood (SRBCT, Khan et al. (2001)) includes 4 different types of tumours with 23, 20, 12 and 8 microarrays per class and 2308 genes.
(4) The Brain data set compares 5 embryonal tumours (Pomeroy et al. (2002)) with 5597 gene expression. Classes 1, 2 and 3 count 10 microarrays each, the remaining classes 4 and 8.

The Brain and the Leukemia data sets were pre-filtered with a very large F-test p-value (0.1 and 0.2, leaving 1963 and 3000 genes). These data sets are succintly described in Table 1.
All these data sets were chosen for their unbalanced characteristics as the minority class represents for each data set a small percentage of the total number of observations (see Table 1). The Brain data set is characterized by a very small number of samples (42) with one class that is extremely under represented compared to the other classes. The data sets are assumed to be correctly normalized.

3.2 Comparison of the ranking methods with no weighting procedure

Performance comparison

Figures 1 **(a) (b) (c)** and **(d)** display the e.632+ error rates obtained on Lymphoma, Leukemia, SRBCT and Brain with respect to the number of selected genes with the different ranking methods.
The classification complexity of the data sets is easy to identify as Lymphoma **(a)** and SRBCT **(c)** display an evaluated error rate less than 7% for a selection of 10 genes, whereas for Leukemia **(b)** and Brain **(d)**, the error rates vary between 25 to 35 % for a selection of 10 genes. Note that this high error rate was not due to the prefiltering process in these data sets.
OFW is generally among the best performers, and the error rates of OFW+CART and OFW+SVM are often very close.
RF achieves good results on Leukemia and SRBCT, whereas on Lymphoma and Brain, the performance of the RF selection is the worst. RF might not succeed in selecting genes with information relevant enough, especially in Lymphoma, where all classes are easy to classify with too many informative variables.
On the contrary, the F-test achieves good results on Lymphoma and Brain. This filter method orders genes that are differentially expressed (significant) for at least one of the classes. If genes are differentially expressed for more than one class (or for all classes), the selected genes will be all informative enough and the performance will be good. With Leukemia, the F-test performs the worst. This data set is more difficult to classify as the 2 classes ALLB and ALLT are very similar (Golub et al. (1999)). The difficulty is reinforced as ALLB is the majoritary class while ALLT is the minority one in this 3-class problem. The F-test thus first ordered significant genes that discriminated the easiest class (ALLB), to the detriment of the other classes.
In any case, these results show that one cannot draw general conclusions on the best method to apply. However, on these data sets, OFW+CART and OFW+SVM who performed among the best and seemed to select discriminative genes, could select candidates to answer the biological study.

Remark on the performance assessement with e.632+ bootstrap error rate

The e.632+ error rate was chosen as it is the most adequate to compute the performance of the different methods on small sample data sets (Ambroise and McLachlan (2002)). However we did observe some weaknesses and the interpretation of the results should be done with caution. One would expect the error rate to increase when the number of evaluated variables becomes too big (as more noise enters the selection). This was not the case for any method using the SVM classifier and RF. When more variables entered the selection, the error rate tended to stabilize to a minimum error rate. With this kind of data set, the SVM classifier seems hence to base its classification on the only good variables in the selection. It is the same with RF, which construction is mostly based on the important (discriminative) variables. We did not observe this tendency with OFW+CART, as during the evaluation step, each aggegated tree is constructed on a small variable subset from the selection.
The evaluation error rate should hence be used solely to compare the ranking methods between each others, and not to give an accurate classification error rate of a given variable selection.

Stability

Computation of the Jaccard index with respect to the number of selected genes are displayed on Figures 2 **(a)** **(b)** **(c)** and **(d)**. Maximum stability is obtained on easy data sets (Lymphoma **(a)** and SRBCT **(c)**) with a Jaccard index reaching 0.45 and 0.6. The F-test is undoubtly the most stable method on complex data sets (Leukemia **(b)** and Brain **(d)**) whereas RF is the most stable on the easy data sets. OFW+SVM and OFW+CART are the less stable. The good stability results of the filter method is easy to explain as the F-test selects redundant information usually only on the majority class, whereas the other methods select genes with relevant information on all classes. As the gene selection might be strongly dependent on the observations drawn in the bootstrap sample, especially if one of the classes is minoritary, the methods focusing on the minoritary classes will consequently be less stable.
OFW+SVM and OFW+CART are stochastic methods and are hence less stable for all data sets. When the number of classes becomes large (Brain and SRBCT), the stability results seem largely affected. A compromise needs hence to be taken between information (on all classes) and stability.

Insight in the different selections

Tables 2 and 3 provide more insight of the different 50 gene lists selected with all methods on each data set. For example in Table 2 for the Lymphoma

data set (upper triangle), OFW+SVM and OFW+CART selected 12 common genes among the 50 selected.
The most striking point is the very few number of shared genes between all methods, that highlights the characteristics of each ranking method. Generally, as they are constructed with the same classifier, RF and OFW+CART share a fair amount of genes (22 and 18 on Lymphoma and Leukemia, Table 2). Table 2 also shows that RF selected more significant genes (*i.e* differentially expressed with F-test) than OFW+CART/SVM (30 and 11 on Lymphoma and Leukemia). In Table 3, where the number of classes is bigger than 3 (SRBCT, Brain), the 3 methods RF, OFW+CART and OFW+SVM generally share more genes together than with the F-test. This highlights the poor relevancy of a selection made with an F-test in this context.
On all data sets except SRBCT, OFW+CART and OFW+SVM shared very few genes. This can be explained as the construction of these two classifiers is completely different: CART searches in the feature space the best variable and the best split to divide each node in the tree while SVM looks for the optimal hyperplane between two classes. As for SRBCT where all methods except F-test seem to share numerous genes, it can be explained as they seemed to perform equally with the same relevant genes (see Fig. 1 (**c**)).
Note that the same tendency was observed if we reduced the size of the selection (*e.g* from 50 to 10): the top selected genes were not necessarily the same from one selection to another.

3.3 Comparisons of the weighted and non-weighted procedures of OFW

The aim of this section is to compare the weighted and non-weighted versions of OFW only, as the other ranking methods do not share the same weighting procedure (especially WRF/BRF for RF, Chen et al. (2004)), the F-test having no weighting procedure).

Performance comparison

In order to compare the internal weighting procedure in OFW+CART or SVM, we computed the e.632+ error rate for both approaches: weighted (wOFW) or non-weighted (OFW). We remind that the weighted procedure implies an internal weighted error rate in the gradient.
For the e.632+ computations, the learning of the *nb* bootstrap samples of wOFW or OFW for each classifier was performed. Then, during the testing phase, both types of learning were evaluated with a *weighted* e.632+. This was necessary in order to compare the improvement of the performance with the weighting approach. A non-weighting approach in e.632+ would indeed favour the majority class to the detriment of the minority class and would still give

a (wrongly) low error rate.
On Figures 3 **(a)** **(b)** **(c)** and **(d)** the weighted e.632+ error rate of OFW and wOFW are displayed, with the application of either CART or SVM for the four data sets.
There is often a strong difference between the performances of OFW+CART and wOFW+CART, showing that CART seems affected by unbalanced classes, whereas there is no difference between the two variants of OFW+SVM. The one-vs-one SVM approach seems hence extremely well adequate for unbalanced classes. wOFW+CART seems to improve the error rate compared to OFW+CART on the easy data set Lymphoma. And for SRBCT, all methods perform similarly.
These graphs show that the weighting procedure in OFW+SVM seems not necessary in the multiclass case as the one-vs-one SVM aims to classify each class, even minoritary. On the contrary, for OFW+CART, the weighting procedure might be needed as by construction, CART tends to favour the majoritary classes.

Stability

The comparisons of the Jaccard index for both versions of the algorithm is displayed on Figures 4. wOFW+SVM seems to improve the stability of the results of the 3-class data sets Lymphoma and Leukemia. When the number of classes is larger, the non-weighted versions are the most stable.
These Jaccard indexes are very low as the proportion of the minoritary observations is often diminished during the bootstrap sampling and the selected variables discriminating the minority classes must stronlgy depend on each bootstrap sample.

Comparisons of the lists (weighted vs. non-weighted)

We compared the lists given by the weighted *vs.* the non-weighted procedures in OFW+CART or SVM in Table 4. There is a difference in the gene selections between the weighted and non-weighted version of OFW. For example on Lymphoma, OFW+SVM and wOFW+SVM shared 13 genes out of the 50 selected. This is surprising as section 3.3 showed that there was not a strong difference in the performance of both methods (Fig. 3 **(a)**). However, with SRBCT, where all performances of the four tested version were similar (Fig. 3 **(c)**), the number of shared genes was quite close and high compared to the other data sets (from 24 to 31 in Table 4).
This table shows that the less the number of genes that are shared between OFW and wOFW, the better the improvement of the selection in terms of relevancy might be (as wOFW aims to favour minoritary classes). For example the selections of wOFW+SVM in Lymphoma might be more informative than the

OFW+SVM selection, the same stands for wOFW+CART *vs.* OFW+CART in Leukemia and Brain.

4 Application on a real world microarray data set

4.1 The pig folliculogenesis data set

This experiment was designed to compare different sizes of healthy follicles granulosa cells during the last stages of antral phase. Large (L), Medium-sized (M) and Small (S) follicles from three different sows per size category were used. After extraction, the RNA isolated from these cells was used to hybridise 42 microarrays that includes duplicates, resulting in 20 Large, 14 Medium-sized and 8 Small follicle cases (GEO accession number: GSE5798). After a normalizing and a filtrering steps, the expression of 1564 clones remain on each microarray.
The main characteristic of this dataset is the obvious difference between the Large follicles and the others. This is due to the biological properties of the data mainly including the appearance of LH receptors between the Medium and Large follicles (Figure 5). Medium-sized and Small follicles are still in the growth process whereas the Large follicles are completely differentiated to produce steroid hormones. Moreover, during the measurements that assign each follicle its class, the diameters of the Small and the Medium-sized follicles are very similar (1-2mm and 3 mm) whereas the Large ones cannot be mistaken (5-6mm). Another factor to consider is the vast majority of regulated cDNAs (clones) overexpressed in the Large follicles and hence the minority of regulated cDNAs (refered to as *genes* instead of clones) that are overexpressed in the Small ones.
We are clearly here in the practical case where classes are unbalanced, and where the number of original samples is extremely small, as some of the microarray experiments were duplicated.

4.2 Results and biological interpretation

The analysis of this data set with Random Forests and F-test was performed in Bonnet et al. (2007) and gave biologically relevant results. We focus here on the application of OFW+CART/SVM and their weighted variants.

Application of OFW

When the number of original samples is extremely small, the e.632+ bootstrap error rate must be considered with caution and should not be the only argument to favour a gene selection from a feature selection method rather than an other. Fig. 6 displays the weighted e.632+ error rate for all approaches. Both OFW+SVM and wOFW+SVM seem to give the best performance.
However, our experience show that the most biologically relevant results do not always give the best statistical performance (Lê Cao et al. (2007)). This is why biological interpretation is a crucial step when analysing microarray data.

Interpretation of the results

In these four gene lists we identified genes GSTA1 and Cyp19A3 known to be overexpressed during follicular development (Keira et al. (1994), Slomczynska et al. (2003)) and nexin, ACTA2, ATF7, UBC, that were not selected by F-test and Random Forest in the previous analysis.
Figure 7 displays the boxplots of the 9 top genes selected either with OFW+CART or OFW+SVM for each class (L, M or S). They show that while a minority of selected genes are overexpressed in the S class with OFW+CART (left), a majority of them are overexpressed in the S class in the OFW+SVM selection (right). This tendency can be generalised for a larger list of genes. It seems here that the construction of the one-vs-one SVM tends to favour mostly genes discriminating the minoritary class S rather than the majoritary class L, as L seems too easy to classify.
When applying wOFW+CART and wOFW+SVM, this tendency is still observed, with more genes overexpressed in S for the wOFW+CART selection (not shown).
The biological analysis shows that most of the overexpressed genes in the S class code for ribosomic proteins that may be associated with a decrease of proliferation during follicular growth from Small to Medium follicles. The wOFW+SVM selection seems hence to give a better discrimination between S and M classes. However, we also identify in this selection a great number of unknown genes that will need further investigation. The wOFW+CART selection seemed not appropriate here since two negative controls were selected and the OFW+SVM selection missed the known discriminative gene CYP11A3.
This section shows that depending on the experimental design, as well as the accurate biological questions, the statistician might not answer the study's aim if the conclusions are only drawn from the statistical results.

5 General remarks

5.1 Computation time.

The experiments were performed in R with a 1.6 GHz 960 Mo RAM AMD Turion 64 X2 PC for OFW+SVM (implementation in R) and OFW+CART (implementation in C in a R package). The learning time of OFW mostly depends on the initial number of variables in the feature space and the step of the stochastic scheme, as well as the size p of ω and the number of trees aggregated for OFW+CART. For Brain (Lymphoma) that contains 1963 (4026) genes, the learning took about 1 (1.5) hour for OFW+SVM for 200 000 iterations. Note that this algorithm would be much faster if it was implemented in C. It took 1 (3.5) hour for OFW+CART for 5000 iterations.

5.2 General remarks

This paper shows that microarray data sets have various levels of difficulty and are quite unpredictable if there is not a solid biological knowledge background of the data set. The analysis on public data set shows that there is no data set that seems to behave like the other. Without biological expertise, it is extremely difficult to assess the biological relevancy of the results.
Simulating a set of data would not help giving more insight in the applied methodologies, as simulating a data set like microarray is an extremely complex work.
The performance assessment of the methods could be computed, but had sometimes serious limits, due to the evaluation method and the applied algorithms or the small number of samples. This study shows that the evaluation part has to be taken with caution by the user in search of the "best" method. Furthermore, although there seemed to be no improvement of the performance of the method when applying wOFW+SVM, the resulting gene selection seemed to contain more information on the minoritary class. Our evaluation performance method might hence not be adequate in this context, especially for OFW+CART where a 'double bootstrap sampling' is performed during evaluation. We also believe that the performance of wOFW+CART can be improved by including weights in the construction of the tree.
Both multicategory classifiers CART and one-vs-one SVM applied with OFW seem to perform better than the other tested methods. Regarding the performances, choosing between these two methods seems difficult. If the user is interested in biological relevancy of the gene selection, then OFW+CART might be the best as the construction of CART really fits this requirement (finding genes with differential expression in different classes at each node of

the tree). However if the interest mostly lies in the classification task and finding predictive genes, then OFW+SVM might be the best. By construction, it searches the best hyperplane between two of the classes. In contrary to CART, SVM optimizes a cost criterion based on the classification performance.

6 Conclusion

Starting from Lê Cao et al. (2007) that provided interesting results for binary problems, we extended the application of OFW+CART and OFW+SVM one-vs-one for multiclass microarray problems. These data sets are known to be difficult because of their high dimensionality with a small sample size and at least one of the classes that is under represented. For most classifiers, this often results in a good overall classification accuracy even though the minority classes are misclassified.
We first compared OFW+CART and OFW+SVM with two other methods, Random Forests and the still widely used F-test in gene selection. All methods were performed with no weighting procedure. Our results showed that our two methods gave good results in terms of error rate estimation and that the filter method F-test might not be appropriate for multiclass datasets. The stability of the results tended to be better in OFW+SVM than CART.
We then compared the weighted version of wOFW+CART or SVM. There seemed to be no difference in the performance evaluation between the weighted and the non-weighted version of OFW+SVM, that performed the best. The performances of the two versions of OFW+CART differed largely, due to the extensive use of bootstrap samples during the learning step. The relevancy of the selected genes with wOFW should however be improved as they aim to discriminate the minoritary classes. An intensive study should now be performed on the public data sets.
Application and biological interpretation on a real world data set show that the wOFW+SVM selection might give relevant results that are complementary with a previous analysis.

Availability

OFW is being implemented in an R package and can be available upon request to the corresponding author.

References

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., Hudson J. Jr, Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C,. Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Byrd J.C., Botstein D., Brown P.O., Staudt L.M., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000 403, 503-511.

Ambroise, C. and McLachlan, G., 2002. Selection Bias in Gene Extraction in Tumour Classification on Basis of Microarray Gene Expression Data. Proc. Natl. Acad. Sci. USA 99(10), 6562-6566.

Bonnet, A., Lê Cao, K-A., SanCristobal, M., Benne, F., Tosser-Klopp, G., Robert-Granié, C., Law-So, G., Besse, P., De Billy, E., Quesnel, H., Hatey, F., Identification of gene networks involved in antral follicular development. (Laboratoire de Génétique Cellulaire, INRA, 2007).

Breiman, L., Friedman, J.H., Olshen, R. A., Stone, C.J., 1984. Classification and Regression Trees. Chapman and Hall.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123-140.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-22.

Chen, C. Liaw, A., Breiman, L., Using Random Forest to Learn Imbalanced Data. (Dept. of Statistics, University of Berkeley, 2004).

Chen, D., Hua, D., Reifman, J. and Cheng, X., 2003. Gene Selection for Multi-Class Prediction of Microaray Data. (Proceedings of the Computational Systems Bioinformatics: CSB'03).

Duan, K-B. and Keerthi S., 2005. Which Is the Best Multiclass SVM Method? An Empirical Study. Lecture Notes in Computer Science, Springer Berlin-Heidelberg 3541/2005, 278-285.

Efron, B. and Tibshirani R.J., 1997. Improvements on cross-validation: the e.632+ bootstrap method. Journal of American Statistical Association 92, 548-560.

Gadat, S. and Younes, L., 2007. A Stochastic Algorithm for Feature Selection in Pattern Recognition. Journal of Machine Learning Research 8(Mar), 509-547.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286 531-537.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389-422.

Joachims, T., 1999. Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. In: B. Schölkopf and C. Burges and A. Smola (Ed.), MIT Press, 1999.

Keira M., Nishihira J., Ishibashi T., Tanaka T. and Fujimoto S., 1994. Iden-

tification of a molecular species in porcine ovarian luteal glutathione S-transferase and its hormonal regulation by pituitary gonadotropins. Arch Biochem Biophys. Jan 308(1):126-32.

Khan J., Wei J.S., Ringnr M., Saal L.H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu C.R., Peterson C., Meltzer P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7 Number 6, June.

Kushner, H. and Clark, DS., 1978. Stochastic Approximation Method for Constrained and Unconstrained Systems. Springer-Verlag.

Lê Cao, K-A., Gonçalves, O., Besse, P. and Gadat, S., 2007. Selection of biologically relevant genes with a wrapper stochastic algorithm, Statistical Applications in Genetics and Molecular Biology, Vol 6: Iss. 1, Article 29.

Lee, Y. and Lee, C-K., 2002.Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19, 1132-1139.

Li, T., Zhang, C. and Ogihara, M., 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20, 2429-2437.

McCarthy, K., Zabar, B. and Weiss, G., 2005. Does cost-sensitive learning beat sampling for classifying rare classes? (Conference on Knowledge Discovery in Data, Proceedings of the 1st international workshop on Utility-based data mining, Chicago, Illinois), 69-77.

Pomeroy S.L., Tamayo P., Gaasenbeek M., Sturla L.M., Angelo M., McLaughlin M.E., Kim J.Y., Goumnerova L.C., Black P.M., Lau C., Allen J.C., Zagzag D., Olson J.M., Curran T., Wetmore C., Biegel J.A., Poggio T., Mukherjee S., Rifkin R., Califano A., Stolovitzky G., Louis D.N., Mesirov J.P., Lander E.S., Golub T.R., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002 415, 436-442.

Slomczynska M., Szoltys M., Duda M., Sikora K. and Tabarowski Z., 2003. Androgens and FSH affect androgen receptor and aromatase distribution in the porcine ovary. Folia Biol (Krakow) 51(1-2):63-8.

Yeung, K.Y., Burmgarner, R.E., 2003. Multi-class classification of microarray data with repeated measurements: application to cancer. Genome Biology 2003 4:R83

Vapnik, V., 2000. The nature of statistical learning theory. Springer-Verlag.

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M., 2003. Use of the Zero-Norm with Linear Models and Kernels Methods. Journal of Machine Learning Research 3, 1439-1461.

Table 1
Summary of the four data sets.

	Lymphoma	Leukemia	SRBCT	Brain
# genes	4026	3000 (pf[1])	2308	1963 (pf[1])
# classes	3	3	4	5
# obs.	62	72	63	42
# obs. per class	42/9/11	38/9/25	23/20/12/8	10/10/10/4/8
% obs. per class	68/14.5/17.5	53/12.5/34.5	36.5/32/19/12.5	24/24/24/9/19
% obs. per class if balanced	33.33	33.33	25	20

[1]pre-filtered with a very large F-test p-value (0.1 and 0.2)

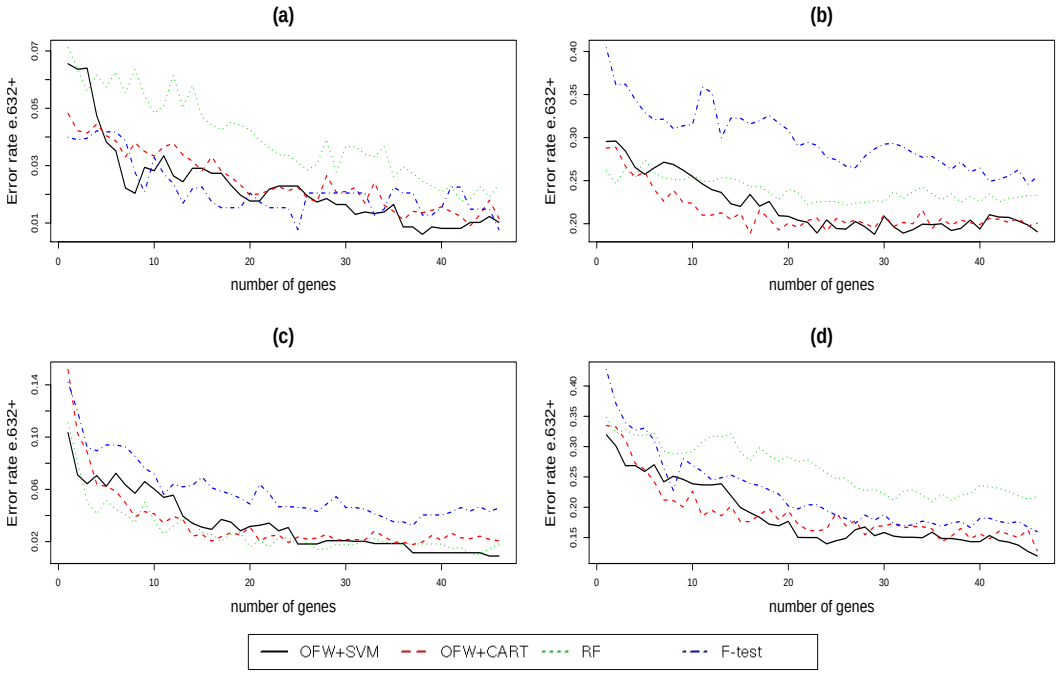

Fig. 1. Error e.632+bootstrap of several algorithms with respect to the number of genes on Lymphoma (a), Leukemia (**b**), SRBCT (**c**) and Brain (**d**).

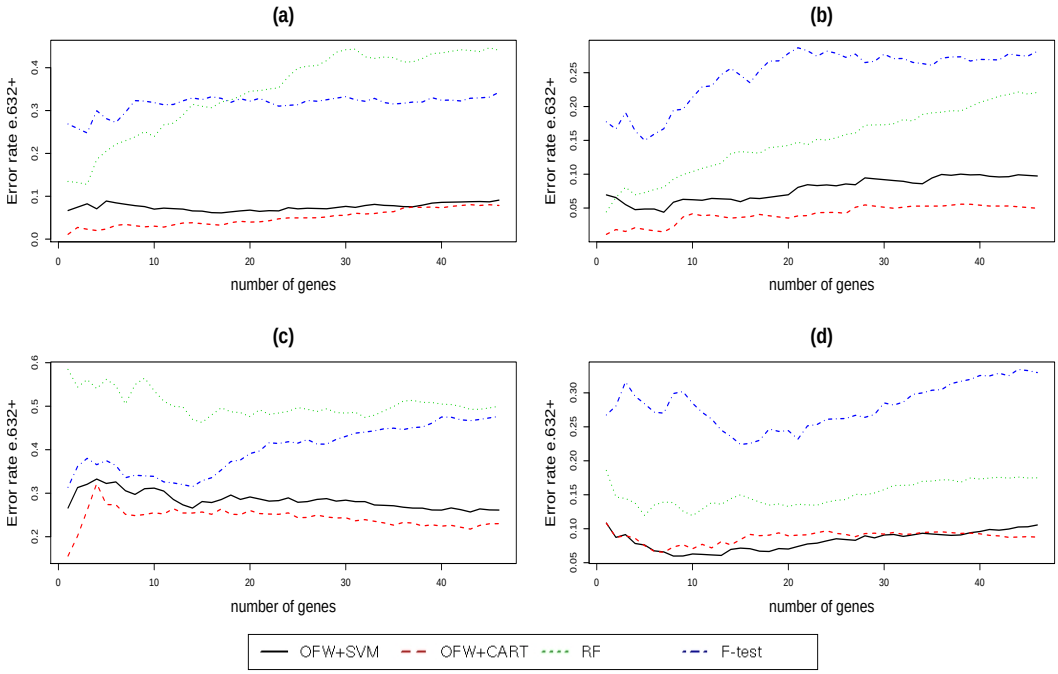

Fig. 2. Jaccard index of OFW+SVM, OFW+CART, RF and F-test with respect to the number of genes on Lymphoma (**a**), Leukemia (**b**), SRBCT (**c**) and Brain (**d**).

Table 2
Number of genes shared by several feature selection algorithms on Leukemia or Lymphoma for a selection of 50 genes.

Leukemia \ Lymphoma	OFW+SVM	OFW+CART	RF	F-test
OFW+SVM	#	12	11	12
OFW+CART	7	#	22	24
RF	17	18	#	30
F-test	3	6	11	#

Table 3
Number of genes shared by several feature selection algorithms on Brain or SRBCT for a selection of 50 genes.

Brain \ SRBCT	OFW+SVM	OFW+CART	RF	F-test
OFW+SVM	#	25	31	11
OFW+CART	8	#	29	15
RF	12	22	#	9
F-test	7	2	2	#

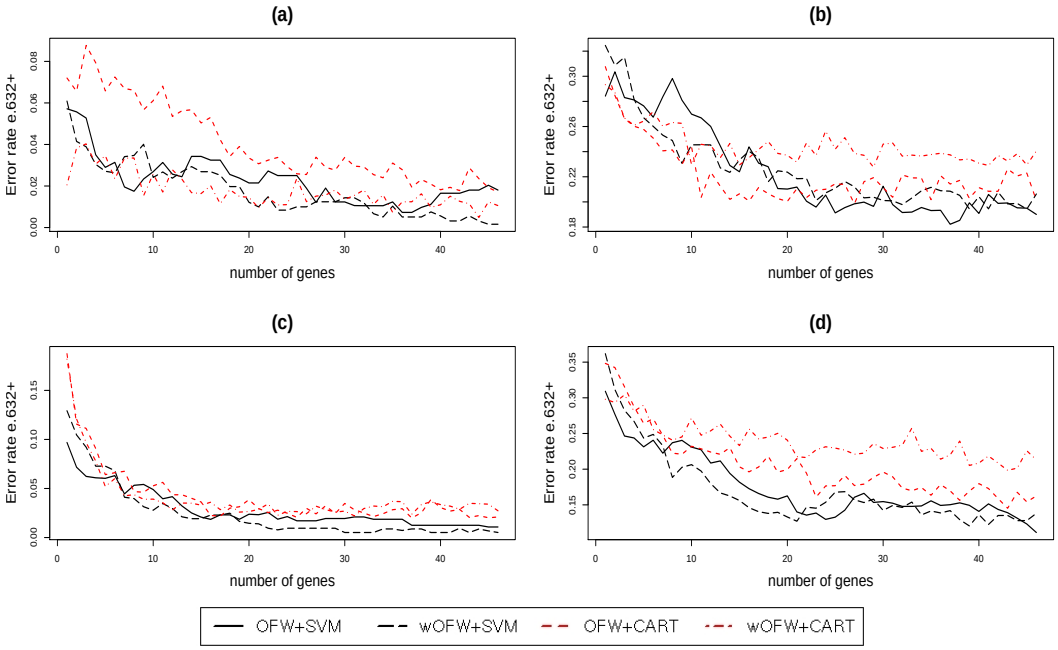

Fig. 3. Weighted e.632+ bootstrap error of OFW+CART and OFW+SVM with both procedures weighted and non weighted with respect to the number of genes on Lymphoma (**a**), Leukemia (**b**), SRBCT (**c**) and Brain (**d**).

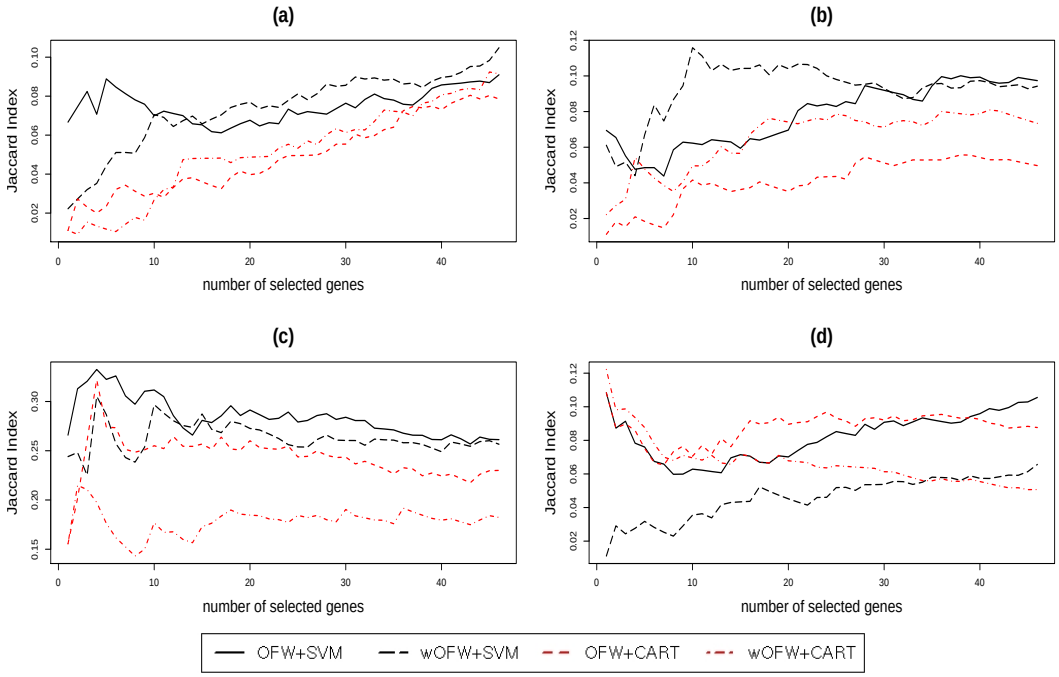

Fig. 4. Comparison of the Jaccard index with the weighted and non-weighted versions of OFW+SVM and OFW+CART on Lymphoma (**a**), Leukemia (**b**), SRBCT (**c**) and Brain (**d**).

Table 4
Number of genes shared by the weighted and non-weighted versions of OFW+SVM or OFW+CART for each data set (selection of 50 genes).

	Lymphoma	Leukemia	SRBCT	Brain
OFW+SVM ∩ OFW+CART	12	7	29	8
wOFW+SVM ∩ wOFW+CART	16	5	24	4
OFW+SVM ∩ wOFW+SVM	13	13	31	18
OFW+CART ∩ wOFW+CART	27	11	25	13

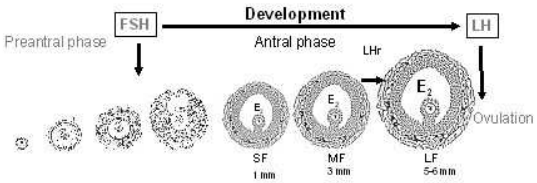

Fig. 5. The three follicle classes: Small, Medium-sized and Large.

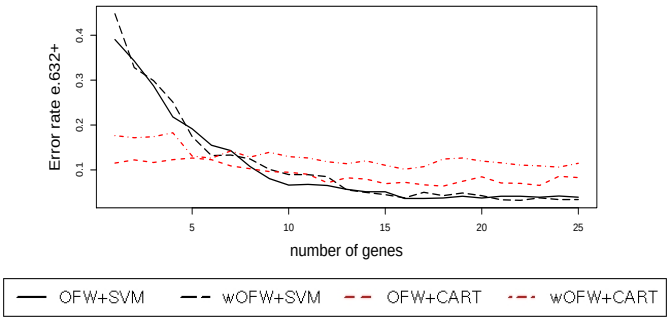

Fig. 6. Weighted e.632+ bootstrap error of OFW+CART and OFW+SVM with both procedures weighted and non weighted with respect to the number of genes on the follicle data set.

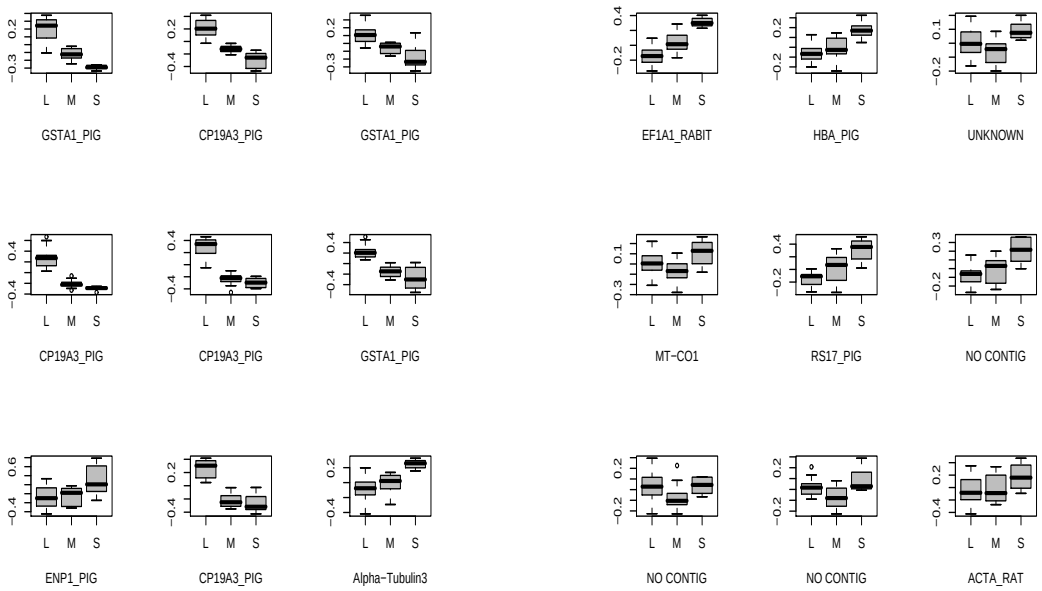

Fig. 7. Boxplots of the 9 top genes selection with OFW+CART (left) or with OFW+SVM (right) on the follicle growth data set. Boxplots are displayed for each class (L, M and S).