Computational Statistics and Data Analysis 53 (2009) 3601–3615

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Multiclass classification and gene selection with a stochastic algorithm

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

[a] Institut de Mathématiques de Toulouse, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[b] Station d'Amélioration Génétique des Animaux UR631, INRA, F-31326 Castanet-Tolosan, France
[c] Laboratoire de Génétique Cellulaire UMR 444, INRA, F-31326 Castanet-Tolosan, France

ARTICLE INFO

Article history:
Received 28 October 2007
Received in revised form 23 February 2009
Accepted 24 February 2009
Available online 11 March 2009

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. Starting from the application of the stochastic meta-algorithm "Optimal Feature Weighting" (OFW) for selecting features in various classification problems, focus is made on the multiclass problem that wrapper methods rarely handle. From a computational point of view, one of the main difficulties comes from the unbalanced classes situation that is commonly encountered in microarray data. From a theoretical point of view, very few methods have been developed so far to minimize the classification error made on the minority classes. The OFW approach is developed to handle multiclass problems using CART and *one-vs-one* SVM classifiers. Comparisons are made with other multiclass selection algorithms such as Random Forests and the filter method F-test on five public microarray data sets with various complexities. Statistical relevancy of the gene selections is assessed by computing the performances and the stability of these different approaches and the results obtained show that the two proposed approaches are competitive and relevant to selecting genes classifying the minority classes.

Application to a pig folliculogenesis study follows and a detailed interpretation of the genes that were selected shows that the OFW approach answers the biological question.

1. 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 classification methods. Dimensionality must therefore 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, with the filter methods or implicitly, with the 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 they may fail to find the "useful" set of variables, as they usually select variables with redundant information. On the contrary, the aim of the wrapper methods is to measure the usefulness of a subset of features in the whole set of variables. However, when dealing

* Corresponding author at: Institut de Mathématiques de Toulouse, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France. Tel.: +61 733462623; fax: +61 733462101.
E-mail address: k.lecao@uq.edu.au (K.-A. Lê Cao).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2009.02.028

with a large number of variables as is the case here, it is computationally impossible to do an exhaustive search among all subsets of features. Furthermore, these methods are prone 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 and avoid local minima. The "Optimal Feature Weighting" algorithm (OFW) proposed by Gadat and Younes (2007) allows for the selection of an optimal discriminative subset of variables. This meta-algorithm can be applied with any classifier. For example Support Vector Machines (SVM, Vapnik, 1999) and Classification And Regression Trees (CART, Breiman et al., 1984) were applied to this stochastic approach 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 and a filter method on public microarray data sets. The relevancy of the results was assessed in a statistical manner by measuring the performance of each gene selection and with a thorough biological interpretation. The selections obtained with OFW were statistically competitive and biologically relevant, even for complex data sets.

From this point, we investigate OFW 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 very small number of cases in at least one of the classes. Furthermore, this "rare" minority class is 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 at minimizing 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 (Li et al., 2004). Several methods have been proposed in recent years. Chen et al. (2004) proposed balanced or weighted random forests, McCarthy et al. (2005) compared sampling methods and cost sensitive learning, but with no clear winner in the results. More recently Eitrich et al. (2007) and Qiao and Liu (in press) also addressed the unbalanced multiclass issue with cost sensitive machine learning techniques and SVM.

In the specific context of multiclass microarray data, Li et al. (2004) applied various classifiers with various feature selection methods and concluded that the accuracy was 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 the selected genes, Tibshirani et al. (2002) proposed the Shrunken Centroid approach and Yeung and Burmgarner (2003) applied uncorrelated or error-weighted Shrunken Centroid. More recently Chakraborty (2008) proposed a Bayesian Nearest Neighbor model.

In this study we compare two ways of handling multiclass data, either with an internal weighting procedure in OFW to take into account the minority classes or without. 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 factors to evaluate the relevancy of the results will be given for one case study. The reader can also refer to Lê Cao et al. (2007) that highlights the importance of biological interpretation in the analysis.

We apply the multicategory classifier CART and the *one-vs-one* SVM approach with OFW on five public microarray data sets. Numerical comparisons are done with Random Forests, that are known to perform efficiently on such data sets, and one filter method (F-test). We compute the e.632+ bootstrap error from Efron and Tibshirani (1997) for each feature selection method, assess the stability of the results with the Jaccard index and compare the different gene lists. The weighted and non-weighted approaches are then compared in OFW + CART and OFW + SVM with the same tools. Finally, application and biological analysis are performed on a pig folliculogenesis data set.

The first section introduces the theoretical adaptation of the OFW model to the multiclass framework. In the 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 public data sets and on a practical data set follow. The paper ends with further elements of discussion.

2. The OFW model

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

2.1. Measure of the classification efficiency

Let $\mathcal{G}$ be a large set of genes numbered from 1 to N that describe a signal $\mathcal{I}$ that belongs to one of the classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$, $k = 1, \ldots, K$. 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 this probability $\mathbb{P}$ that fits the efficiency of each gene for the classification of $\mathcal{I}$ in $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$. In this probability, important weights are given to genes with a high discriminative power and lower weights to those that have the poorest influence on the classification task. We denote by p any small integer compared to N. A gene subset of size p is drawn from $\mathcal{G}$ with respect to $\mathbb{P}$. 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 for any p-tuple $\omega \in \mathcal{G}^p$, 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. Genes that are selected with respect to $\mathbb{P}$ in (1) are drawn with replacement, although it would look more reasonable to use subsets of genes without replacement. This mainly comes from the mathematical derivations to optimize $\mathcal{E}$ that are described below.

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

2.2. Stochastic optimization method

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 a 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

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

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

$$\mathbb{P}_{t+\mathrm{d}t} = \mathbb{P}_t - \nabla\mathbb{P}_t \mathrm{d}t. \tag{3}$$

The main clue is that the Euclidean 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*

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

This last expression is numerically intractable since it requires the computation of ϵ for every possible p-uple from $\mathcal{G}$. To deal with such gradient, a computable Robbins–Monro algorithm can be used, which exhibits similar asymptotic behavior as (3), see for instance Gadat and Younes (2007) and 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 $\mathbb{P}_n(g) = 0$, 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 $I(t) = \sup\{k \mid \tau_k \leq t\}$, 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:*

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

This last result ensures 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.

2.3. Detailed algorithm

We detail 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 $\mathbb{P}_n^{\otimes p}$,
  - . 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 $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n d_n]$,
  - . $n = n + 1$.

The last lines introduce a projection Π_S which corresponds to the natural affine projection into the simplex S of discrete probability measures. More precisely, we have

$$\Pi_S(q) = \arg\min_{p \in S} \|q - p\|^2.$$

Note that since $\mathbb{P}_n - \alpha_n d_n$ may have some negative coordinates, this projection is slightly different from a simple normalization step. Several details are provided in Gadat and Younes (2007).

3. Application of OFW and performance evaluation

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

3.1. CART and SVM multiclass applied to OFW

3.1.1. CART

OFW is applied with the 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.

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 (1996), the trees were unpruned, but there is no overfitting thanks to the aggregation technique. Note that recently, Zhang et al. (2008) proposed a boosting-based double bagging procedure that seemed to perform better than boosting or bagging alone.

In OFW + CART, for each iteration n, B trees were constructed on B bootstrap samples and on different variable subsets ω_n^b drawn with respect to $\mathbb{P}_n$, $b = 1, \dots, B$. We then defined the efficiency criterion ϵ as the mean classification error rate on the out-of-bag samples. The detailed bagging version of OFW + CART is described in 3.3.

3.1.2. 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-rest* approach, the approach proposed by Lee and Lee (2003), by Joachims (1999) or the multiclass version from Weston and Watkins (1999). Unlike CART, SVM is very stable and ϵ was therefore computed on only one bootstrap sample ($B = 1$).

3.2. Different computations of the gradient

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:

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

where b is the bootstrap sample on which each CART tree is constructed and $\alpha_i = A/(B + i)$ is the step sequence referred to in Section 2.2.

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

3.3. 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 3.2,
  - . update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n D_n]$,
  - . $n = n + 1$.

3.4. Weighting procedure

An efficient way to take into account the unbalanced characteristic of the data is to weight the internal error rate $\epsilon(\omega)$ according to the number of samples of each class in the learning set. This would penalize a classification error made on the minority class and therefore put more weight on the variables that help in classifying this class instead of the majority class.

Let n be the total number of cases and m_k, $k = 1 \dots K$ the number of cases in class k. We define the (normalized) weight of *each* case in class k by $w_k = \frac{1}{m_k \times K}$. Then, for each out-of-bag test case (*i.e.* the sample not drawn in the bootstrap sample), we denote by mis_k the number of misclassified cases from class k. The weighted internal error rate is defined as:

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

instead of $\epsilon(\omega) = \frac{\sum_k^K mis_k}{n}$ in the non-weighting case. This weighting procedure also stands for the evaluation step, see following Section 3.5.

3.5. Performance measurement

3.5.1. Comparison of the prediction performance

The 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 size data sets. Each algorithm is learned on a bootstrap sample to avoid any overfitting during the gene selection evaluation (see Ambroise and McLachlan (2002)). However, note that this performance evaluation does not dictate the optimal number of genes to select. The e.632+ only allows for the comparison of the performances of the different selection methods.

3.5.2. Stability

One can define the feature stability as the level of agreement between the set of genes selected in each bootstrap sample with the set of genes selected using the full training set. The Jaccard index that is then computed lies between 0 (low level of agreement) and 1 (high level of agreement) and will be used to compare the stability of all four 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 as 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 $\Delta_{\max}$ selected genes.

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

3.6. Ranking methods

Multicategory ranking methods are still rare in the context of classification, especially in the microarray data context. We compare three wrapper methods: OFW + CART, OFW + SVM, Random Forests (RF, Breiman, 2001) and the filter method F-test that is still widely used for selecting genes in the context of microarrays. Note that during the evaluation performance, the F-test gene selection was assessed with a *one-vs-one* linear SVM.

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.

4. Statistical assessment on public data sets

A short description of the five public data sets is first given before we apply OFW + CART, OFW + SVM, RF and F-test with no weighting procedure. We compare the results obtained in terms of performance, stability and differences in the gene selections. We then focus on OFW and compare the weighted *vs.* non-weighted procedure with the same criteria cited above.

4.1. Multiclass data sets

Five public multiclass data sets were analyzed in this study.

(1) Lymphoma (Alizadeh et al., 2000) compares 3 classes of cells (42, 9 and 11 cases 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 cases) and the AML class (Acute Myeloid Leukemia, 25 cases). The classes AML-B and AML-T are known to be biologically very similar.
(3) The Small Round Blue-Cell Tumor Data of childhood (SRBCT, Khan et al., 2001) includes 4 different types of tumors with 23, 20, 12 and 8 microarrays per class and 2308 genes.
(4) The Brain data set compares 5 embryonal tumors (Pomeroy et al., 2002) with 5597 gene expression. Classes 1, 2 and 3 count 10 microarrays each, the remaining classes 4 and 8.
(5) The Multiple Tumor data set initially compared 14 tumors (Ramaswamy et al., 2001) and 7129 gene expressions. We used the normalized data set from Yeung and Burmgarner (2003) with 11 types of tumors. To fit into a usual microarray framework (*i.e.* a small number of samples), we randomly selected 90 samples (out of 192) that have tumor types coming from breast (8), central nervous system (4), colon (7), leukemia (26), lung (4), lymphoma (15), melanoma (3), mesotheolima (7), pancreas (6), renal (5) and uterus (5).

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). The Multiple Tumor data set was also pre-filtered with an F-test, leaving 2000 genes, to reduce the computation time of the algorithms. These data sets are succinctly described in Table 1.

All these data sets were chosen for their unbalanced characteristics as the minority class for each data set represents a small percentage of the total number of cases. All data sets were assumed to be correctly normalized.

4.2. Comparison of the ranking methods with no weighting procedure

4.2.1. Performance comparison

Fig. 1 displays the e.632+ error rates obtained on all data sets with respect to the number of the genes selected 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), Brain (d) and Multiple Tumor (e), the error rates vary between 25% to 50% for a selection of 10 genes.

OFW is generally among the best performers, and the error rates of OFW + CART and OFW + SVM are often very close, except for Multiple Tumor, where OFW + SVM gives a poor performance. We suspect that the aggregation of this binary SVM (*one-vs-one*) may not be adapted in this extreme multiclass setting.

RF achieves good results on Leukemia, SRBCT and Multiple Tumor, whereas on Lymphoma and Brain, the performance of the RF selection is the worst. RF might therefore not succeed in selecting genes with sufficient relevant information, especially in Lymphoma, where all classes are easy to classify.

On the contrary, the F-test achieves good results on Lymphoma and Brain. This filter method orders genes that are differentially expressed (*i.e.* significant) for at least one of the classes. If some genes are differentially expressed for more than one class (or for all classes), they will all be informative enough and the performance will be good, which is likely

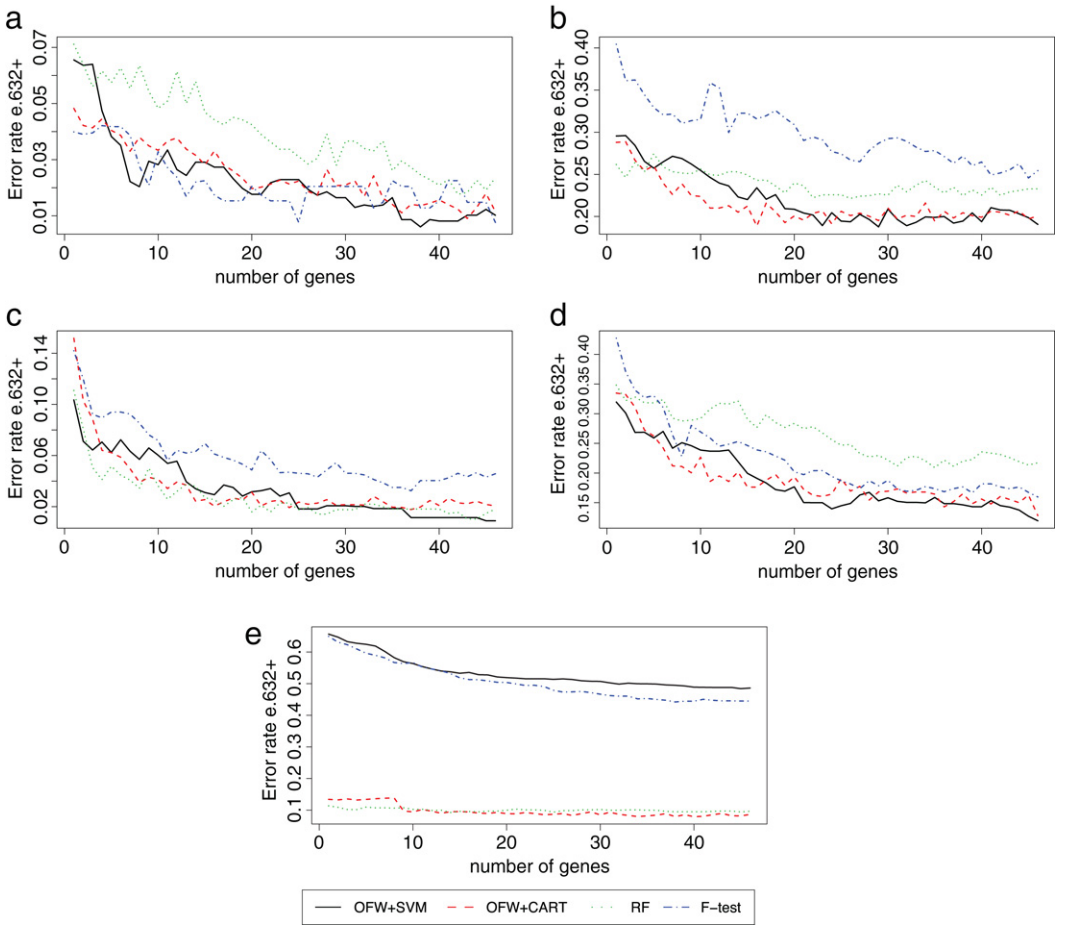

Fig. 1. Error e.632+ bootstrap of several algorithms with respect to the number of genes on Lymphoma (a), Leukemia (b), SRBCT (c), Brain (d) and Multiple Tumor (e).

Table 1
Summary of the five data sets.

	Lymphoma	Leukemia	SRBCT	Brain	Multiple tumor
# Genes	4026	3000[a]	2308	1963[a]	2000[a]
# Classes	3	3	4	5	11
# Obs.	62	72	63	42	90
# Obs. per class	42/9/11	38/9/25	23/20/12/8	10/10/10/4/8	8/4/7/26/ 4/15/3/7/6/5/5

[a] Pre-filtered with a very large F-test *p*-value.

the case for Lymphoma and Brain. With Leukemia however, the F-test performs the worst. This data set is more difficult to classify as the 2 classes ALL-B and ALL-T are similar and ALL-B is the majority class while ALL-T is the minority class. The F-test thus first ordered significant genes that discriminated the easiest class (ALL-B), 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. In general, OFW + SVM and OFW + CART were the best performers, especially OFW + CART in a high multiclass setting.

4.2.2. Remark on the performance assessment 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 large (as more noise enters the selection). This was not the case for any method using the SVM classifier and RF, which are known to base their classification task on the good variables among numerous and possibly noisy variables. The results that we obtained are in agreement with this fact. We did not observe this tendency with OFW + CART, as during the evaluation step, each aggregated tree is constructed on a small variable subset from the selection (see Lê Cao and Chabrier (2008) for the details of the algorithm).

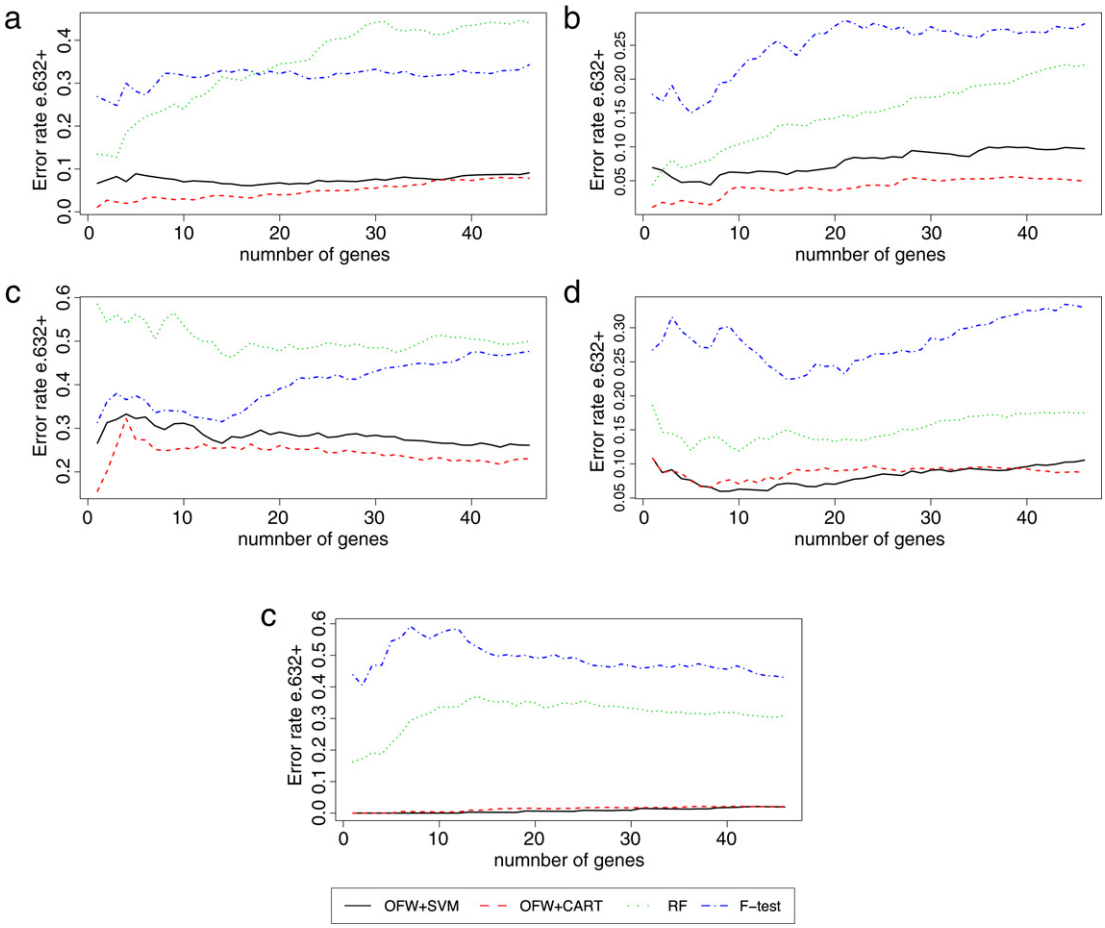

Fig. 2. Jaccard index of OFW + SVM, OFW + CART, RF and F-test with respect to the number of genes on Lymphoma (a), Leukemia (b), SRBCT (c), Brain (d) and Multiple Tumor (e).

The evaluation error rate should thus be solely used to compare the ranking methods between each other, and not to give an accurate classification error rate of a given variable selection.

4.2.3. Stability

The computation of the Jaccard index with respect to the number of selected genes are displayed in Fig. 2. 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 undoubtedly the most stable method on complex data sets (Leukemia (b), Brain (d), Multiple Tumor (e)), although its performance is very poor (as previously discussed). RF is in general very stable compared to OFW + SVM and OFW + CART.

The good stability results of the filter method can be explained as the F-test selects redundant information mostly only on the majority class, whereas the other methods select genes with relevant information on all classes. As the gene selection might strongly depend on the cases drawn in the bootstrap sample, especially if one class size is small, these methods will consequently be less stable.

OFW + SVM and OFW + CART are stochastic methods and are therefore less stable. When the number of classes becomes large (Brain, SRBCT and Multiple Tumor), the stability results seem largely affected. Thus, a compromise needs to be taken between information (on all classes) and stability.

4.2.4. Insight into the different gene selections

Tables 2 and 3 provide more insight into the different genes that were selected with all methods on each data set (not shown for Multiple Tumor). For example in Table 2 for the Lymphoma data set (upper triangle), OFW+SVM and OFW+CART commonly selected 12 genes among the 50 selected.

The most striking point is the very few number of shared genes between all methods. This highlights the differences between 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,

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	#

where the number of classes is larger than 3 (SRBCT and Brain), the 3 methods RF, OFW + CART and OFW + SVM generally shared 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 very different: CART searches the best variable in the feature space and the best split to divide each node in the tree, while SVM looks for the optimal hyperplane between two classes. Note that the same tendency was observed if we reduced or increased the size of the selection (*e.g.* 10 or 100).

The difficulty of the Multiple Tumor data set was strongly highlighted (not shown) as no method shared more than 4 common genes. Given the poor performances of the F-test and OFW + SVM (Section 4.2.1), this small overlapping result is to be expected.

4.3. 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.

4.3.1. 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 recall 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 favor the majority class to the detriment of the minority class and would still give a (wrongly) low error rate.

Fig. 3 displays the weighted e.632+ error rate of OFW and wOFW with the application of either CART or SVM for the five 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 therefore extremely well adequate for unbalanced classes. wOFW + CART seems to improve the error rate compared to OFW + CART on the easy data set Lymphoma (a). For SRBCT (c), all methods perform similarly, whereas for Multiple Tumor (e), wOFW + SVM is still affected by the high number of classes.

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 minority, as long as the number of classes remains reasonable (≤ 5 here). On the contrary, for OFW + CART, the weighting procedure might be needed as by construction, CART tends to favor the majority classes.

4.3.2. Stability

The comparisons of the Jaccard index for both versions of the algorithm are displayed in Fig. 4. wOFW + SVM seems to improve the stability of the results of the 3-class data sets Lymphoma (a) and Leukemia (b). When the number of classes is larger, the non-weighted versions are the most stable. These Jaccard indices are very low as the proportion of the minority cases is often diminished during the bootstrap sampling and the selected variables that discriminate the minority classes must strongly depend on each bootstrap sample. This explains the poor results obtained in Multiple Tumor (e).

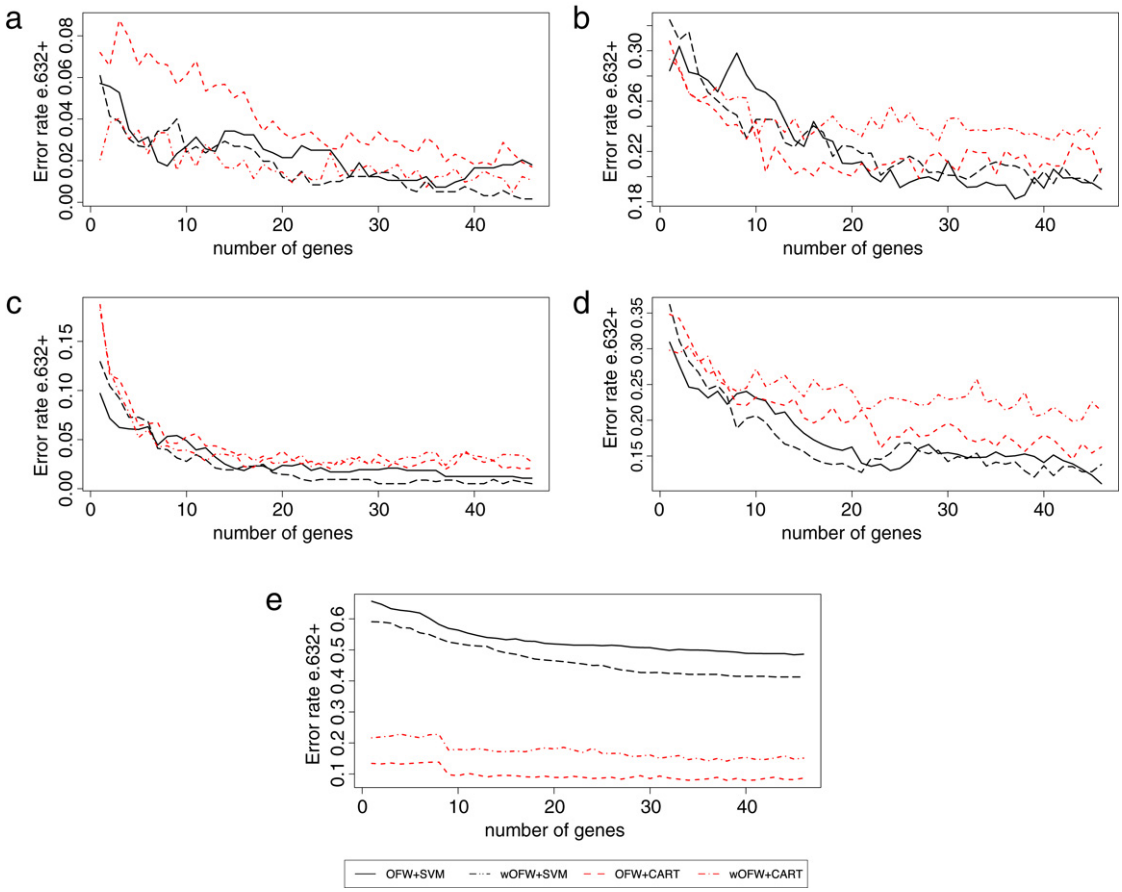

Fig. 3. Weighted e.632+ bootstrap error of OFW + CART and OFW + SVM with both procedures weighted and non-weighted with respect to the number of genes on Lymphoma (a), Leukemia (b), SRBCT (c), Brain (d) and Multiple Tumor (e).

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

We compared the gene 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 versions of OFW. For example with Lymphoma, OFW + SVM and wOFW + SVM shared 13 genes out of the 50 selected ones. This is surprising as there did not seem to be a strong difference in the performance of both methods (Fig. 3(a)). However, with SRBCT where all performances of the four tested versions were similar (Fig. 3(c)), the number of shared genes was quite high and similar compared to the other data sets (from 24 to 31 in Table 4).

The less numerous the genes that are shared between OFW and wOFW, the better the improvement of the selection in terms of relevancy (as wOFW aims to favor minority 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. However, the high complexity of the Multiple Tumor data set shows the limitation of the algorithm OFW. It also highlights a strong difference between all proposed versions of OFW.

5. Application and biological interpretation

When developing feature selection algorithms for microarray data, it is useful to check if the actual gene selection is biologically relevant for the study. The biological interpretation of the results is therefore valuable to show the applicability of such algorithms.

5.1. The pig folliculogenesis data set

This experiment was designed to compare different sizes of healthy follicle granulosa cells during the last stages of antral phase. RNA from Large (L), Medium-sized (M) and Small (S) follicles was extracted from three different sows per size category. The RNA isolated from these cells was used to hybridize 42 microarrays including duplicates, resulting in 20 Large,

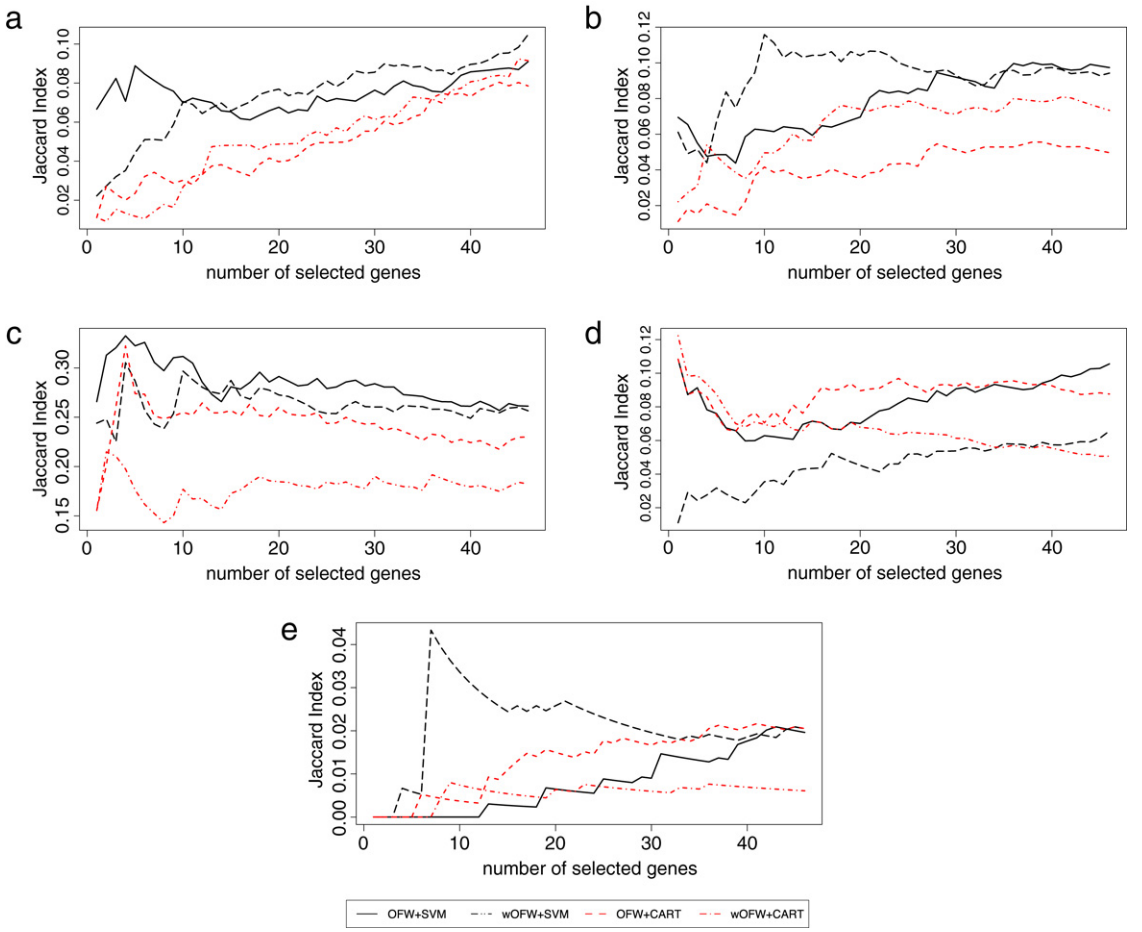

Fig. 4. Comparison of the Jaccard index with the weighted and non-weighted versions of OFW + SVM and OFW + CART on Lymphoma (a), Leukemia (b), SRBCT (c), Brain (d) and Multiple Tumor (e).

Table 4
Number of genes commonly selected 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	Multiple tumor
OFW + SVM ∩ OFW + CART	12	7	29	8	0
wOFW + SVM ∩ wOFW + CART	16	5	24	4	0
OFW + SVM ∩ wOFW + SVM	13	13	31	18	5
OFW + CART ∩ wOFW + CART	27	11	25	13	2

14 Medium-sized and 8 Small follicle cases (GEO accession number: GSE5798). After a normalization and a pre-process step, the expression of 1564 clones are left on each microarray for the analysis.

The main characteristic of this data set is the obvious difference between the Large follicles and the others. This is due to the biological properties of the data, where LH receptors appear between the Medium and Large follicles (Fig. 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 (1–2 mm and 3 mm) whereas the Large follicles cannot be mistaken (5–6 mm). Another factor to be considered is the vast majority of regulated clones that are over-expressed in the Large follicles and hence the minority of regulated clones (from now on referred to as *genes*) that are over-expressed in the Small follicles.

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.

5.2. Results and biological interpretation

The analysis of this data set with Random Forests and F-test was performed in Bonnet et al. (2008) and gave biologically relevant results. We focus here on the application of OFW + CART/SVM and their weighted variants.

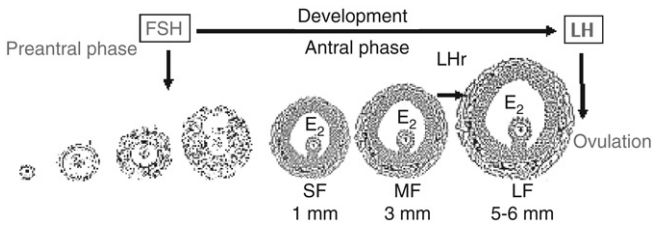

Fig. 5. The three follicle classes: Small, Medium-sized and Large.

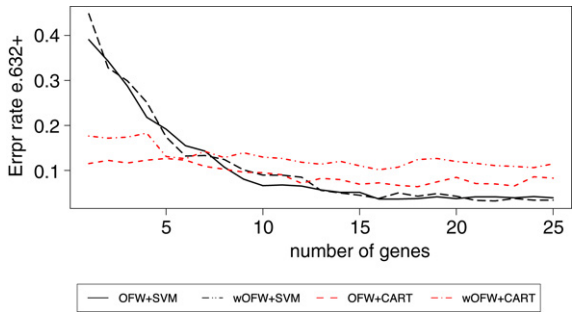

Fig. 6. Weighted e.632+ bootstrap error of OFW + CART and OFW + SVM with both procedures weighted and non-weighted with respect to the number of genes on the follicle data set.

5.2.1. 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 favor a feature selection method. 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 shows 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 analyzing microarray data.

5.2.2. Interpretation of the results

In these four gene lists we identified the genes GSTA1 and Cyp19A3 which are known to be over-expressed 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.

Fig. 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 over-expressed in the S class with OFW + CART (left), a majority of them are over-expressed in the S class in the OFW + SVM selection (right). This tendency can be generalized for a larger list of genes. It seems here that the construction of the *one-vs-one* SVM tends to mostly favor genes discriminating the minority class S rather than the majority class L, as L seems too easy to be classified.

When applying wOFW+CART and wOFW+SVM, this tendency is still observed, with more genes that are over-expressed in S for the wOFW + CART selection (not shown).

The biological analysis shows that most of the over-expressed genes in the S class code for ribosomic proteins may be associated with a decrease of proliferation during follicular growth from Small to Medium follicles. The wOFW + SVM selection seems therefore to give a better discrimination between the S and M classes. We also identified in this selection a great number of unknown genes that will need further investigation. The wOFW + CART selection seemed not appropriate in this study since two negative controls were selected and the OFW + SVM selection missed some known discriminative genes such as CYP11A3 (Bonnet et al., 2008).

This section shows that depending on the experimental design, as well as the precise biological question, the statistician might not answer the aim of the study if the conclusions are solely drawn from the statistical results.

6. General remarks

6.1. 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 of ω and the number of trees aggregated for OFW + CART. For Brain (Lymphoma) that contains 1963 (4026) genes, the learning step took about 1 (1.5) h for OFW + SVM for 200 000 iterations. It took 1 (3.5) h for OFW + CART for 5000 iterations.

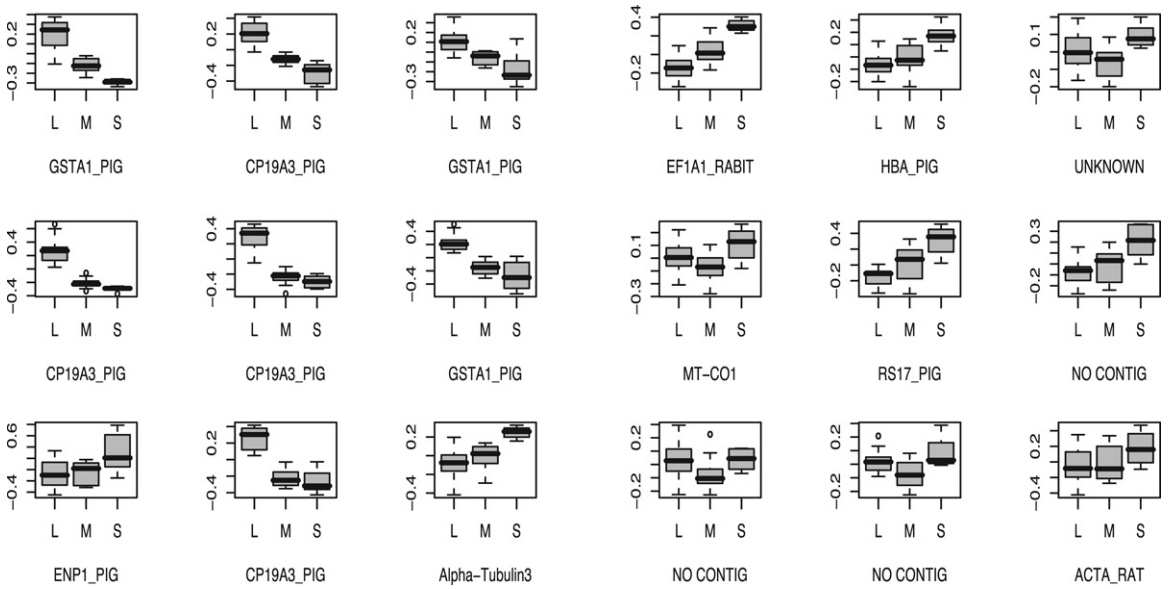

Fig. 7. Boxplots of the 9 top genes selected with OFW + CART (left) or with OFW + SVM (right) on the follicle growth data set. Boxplots are displayed for each class (L, M and S).

6.2. Complexity of OFW

The complexity of the meta-algorithm OFW depends on two points. The first one is the nature of the algorithm used with SVM. The second point is the convergence speed of the stochastic scheme towards a minimum of the energy $\mathcal{E}$.

The complexity of each algorithm used with OFW (CART, SVM, Multiclass SVM etc.) may be very variable and depends on the choice of the user. For instance, with this meta-algorithm, each iteration computes a SVM with N_s samples described by p variables and the complexity of each step is at most $p \times N_s^2$, since $p > N_s$ in this study (see the detailed computation of this complexity in Burges (1998)).

Regarding the second point, the convergence to an optimal state x^* using a standard (non-averaged) Robbins–Monro stochastic approximation scheme $(X_n)_{n\in\mathbb{N}}$ is described by the following assessment:

$$\sqrt{\frac{n}{\log n}}(X_n - x^\star) \to \mathcal{N}(0, \Lambda^*). \tag{5}$$

This last theoretical derivation can be found in Duflo (1997). In this last statement, Λ^* is the trace of the Hessian matrix of $\mathcal{E}$ computed on the optimal state x^*. If n iterations are run in the initial version of OFW (Gadat and Younes, 2007), the convergence speed is bounded by $O\left(\frac{\log n}{n}\mathrm{Tr}\,(\Lambda^*)\right)$. The interest in the OFW meta-algorithm is significant since an exhaustive search of p-uple among N features would require C_N^p iterations.

The aim of the averaging step introduced in Section 3.2 is to improve the rate of convergence of the stochastic scheme reducing the variance of the estimate D_n. The theoretical derivations concerning the rate of convergence are at the moment an open issue but it is likely to reduce the $\mathrm{Tr}\,(\Lambda^*)$ term introduced in (5).

6.3. General remarks

This study shows that microarray data sets have various levels of difficulty and are quite unpredictable if there is no solid biological knowledge of the data set. The analysis of several public data sets shows that no data sets exhibit the same behavior. Without biological expertise, it is extremely difficult to assess the relevancy of the results. Simulating a set of data would not help give more insight into the applied methodologies. Realistically simulating a microarray data set is a complex work, and often, the technical effects on the data are not easily identifiable.

The performance assessment of the methods could be computed, but had sometimes serious limits, either 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 considered 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 instead of OFW + SVM, the resulting gene selection seemed to contain more biological information on the minority class. Thus, our evaluation performance method might not be adequate in this context, especially for OFW + CART where a "double bootstrap sampling" is performed during the evaluation step. We also believe that the performance of wOFW+CART could be improved by directly including weights in the construction of the trees.

Both multicategory classifiers CART and *one-vs-one* SVM that were applied to OFW seemed to perform better than the other tested methods, except when the number of classes was very high (here ≥ 5). In this case, aggregating binary *one-vs-one* SVMs seemed limited. Lee and Lee (2003) mentioned that the *one-vs-rest* SVM can also give poor results if several classes are similar, as is often the case with biological data. One should investigate instead the implementation of a multiclass SVM, as was proposed by Weston and Watkins (1999), to solve the multiclass optimization quadratic problem into the SVM directly.

Regarding the performances, choosing between these two proposed approached seems difficult. If the user is interested in biological relevancy of the gene selection, or if the number of classes is high, then OFW + CART might be adequate as the construction of CART really fits this requirement (*i.e.* 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 appropriate. By construction, the SVM searches the best hyperplane between two of the classes. Contrary to CART, SVM optimizes a cost criterion based on the classification performance.

7. 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 complex problems 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 for gene selection. All methods were applied with no weighting procedure. Our results showed that the two proposed approaches generally gave good results in terms of performance. The filter method F-test seemed not appropriate for multiclass data sets and 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 between the weighted and the non-weighted versions of OFW+SVM, which generally 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 at discriminating the minority classes.

In the case where the classes were numerous (≥ 5) and unbalanced, OFW+CART clearly outperformed OFW+SVM whose poor results were due to the types of binary SVMs that were aggregated for the multiclass purpose. The implementation of OFW with a multiclass SVM might improve these results.

The application and biological interpretation on a real world data set showed that the wOFW + SVM selection gave relevant results and answered the biological question.

Availability

OFW is implemented in an R package called `ofw` (see also Lê Cao and Chabrier (2008)).

References

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., et al., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. Proceedings of the National Academy of Sciences of the United States of America 99 (1), 6562–6566.

Bonnet, A., Lê Cao, K., SanCristobal, M., Benne, F., Tosser-Klopp, G., Robert-Granié, C., Law-So, G., Besse, P., De Billy, E., Quesnel, H., Hatey, F., Tosser-Klopp, G., 2008. Identification of gene networks involved in antral follicular development. Reproduction 136, 211–224.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167.

Chakraborty, S., 2008. Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach. Computational Statistics and Data Analysis 53, 1462–1474.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. Technical Report 666, Dpt. of Statistics, University of Berkeley.

Chen, D., Hua, D., Reifman, J., Cheng, X., 2003. Gene selection for multi-class prediction of microarray data. In: CSB '03: Proceedings of the IEEE Computer Society Conference on Bioinformatics. IEEE Computer Society, Washington, DC, USA, page 492.

Duflo, M., 1997. Random Iterative Models. Springer.

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

Eitrich, T., Kless, A., Druska, C., Meyer, W., Grotendorst, J., 2007. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. Journal of Chemical Information and Modeling 47 (1), 92–103.

Gadat, S., Younes, L., 2007. A stochastic algorithm for feature selection in pattern recognition. Journal of Machine Learning Research 8, 509–547.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286 (5439), 531.

Joachims, T., 1999. Making large-Scale SVM Learning Practical. MIT-Press.

Keira, M., Nihihira, J., Ishibasshi, T., Tanaka, T., Fujimoto, S., 1994. Identification of a molecular species in porcine ovarian luteal glutathione S-transferase and its hormonal regulation by pituitary gonadotropins. Archives of biochemistry and biophysics (Print) 308 (1), 126–132.

Khan, J., Wei, J.S., Ringnér, 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 (6), 673–679.

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

Lê Cao, K., Chabrier, P., 2008. Ofw: An R package to select continuous variables for multiclass classification with a stochastic wrapper method. Journal of Statistical Software 28 (9), 1–16.
Lê Cao, K.-A., Gonçalves, O., Besse, P., Gadat, S., 2007. Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology 6 (1), Article 1.
Lee, Y., Lee, C., 2003. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19 (9), 1132–1139.
Li, T., Zhang, C., Ogihara, M., 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20 (15), 2429–2437.
McCarthy, K., Zabar, B., Weiss, G., 2005. Does cost-sensitive learning beat sampling for classifying rare classes? In: Proceedings of the 1st International Workshop on Utility-Based Data Mining. pp. 69–77.
Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., et al., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415 (6870), 436–442.
Qiao, X., Liu, Y., 2008. Adaptive weighted learning for unbalanced multicategory classification. Biometrics 65 (1), 159–168.
Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., et al., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences 98 (26), 15149–15154.
Slomczynska, M., Szoltys, M., Duda, M., Sikora, K., Tabarowski, Z., 2003. Androgens and FSH affect androgen receptor and aromatase distribution in theporcine ovary. Folia Biol (Krakow) 51, 63–68.
Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences 99 (10), 6567.
Vapnik, V.N., 1999. The Nature of Statistical Learning Theory (Information Science and Statistics). Springer.
Weston, J., Watkins, C., 1999. Support vector machines for multi-class pattern recognition. In: Proceedings of the Seventh European Symposium On Artificial Neural Networks. 4 p. 6.
Yeung, K., Burmgarner, R., 2003. Multi-class classification of microarray data with repeated measurements: Application to cancer. Genome Biology 4 (83).
Zhang, C., Zhang, J., Zhang, G., 2008. Using Boosting to prune Double-Bagging ensembles. Computational Statistics and Data Analysis 53, 1218–1231.