Exploiting tree-based variable importances to selectively identify relevant variables

Vân Anh Huynh-Thu VaHuynh@ulg.ac.be

Louis Wehenkel L.Wehenkel@ulg.ac.be

Pierre Geurts P.Geurts@ulg.ac.be

Department of Electrical Engineering and Computer Science, Systems and Modeling
GIGA-Research, Bioinformatics and Modeling
University of Liège, B-4000 Liège, Belgium

Editor: Saeys et al.

Abstract

This paper proposes a novel statistical procedure based on permutation tests for extracting a subset of truly relevant variables from multivariate importance rankings derived from tree-based supervised learning methods. It shows also that the direct extension of the classical approach based on permutation tests for estimating false discovery rates of univariate variable scoring procedures does not extend very well to the case of multivariate tree-based importance measures.

1. Introduction

In the context of supervised learning, feature selection may be decomposed into two complementary subproblems. The first subproblem aims at identifying among a given set of candidate input variables a maximal subset of so-called *relevant* variables, i.e. variables which convey information about the target output variable, *in isolation or in conjunction with other relevant variables*. The second subproblem aims at identifying among a given set of variables, maximal subsets of so-called *redundant* variables, i.e. maximal subsets of variables which *conditionally to the other variables in that given set* do not convey complementary information about the target output variable. Unfortunately, when the sole information available about the problem under consideration is limited to a training sample of input-ouput pairs, it is not possible to exactly identify maximal subsets of relevant or redundant inputs. Thus, any feature selection algorithm will be at risk of either missing some sought features (false negatives) or of erroneously selecting some truly non desired ones (false positives).

In this paper we consider the problem of identifying *relevant* features from a (typically very large) set of candidate features. Among the different possible selectivity/sensitivity compromizes, we aim at high selectivity, i.e. at identifying subsets of relevant variables while maintaining the rate of false positives as small as possible. This type of compromize is typically sought in the context of 'biomarker discovery' in bioinformatics, where input variables correspond for example to RNA or protein expressions, or genetic polymorphisms, and where one seeks to identify a maximum number of them which truly provide information about some biological condition (disease status, treatment response, etc.) for further analysis and biological insight, while aiming at a very low false positive rate, because of high costs of subsequent experiments (see e.g. Saeys et al. (2007)).

Univariate statistical hypothesis tests provide only a partial answer to this question, because they can only identify variables that provide a significant amount of information about the output variable in isolation from the other inputs. When one seeks for interacting effects, one could resort to importance measures provided by supervised learning methods, such as tree-based methods. However, while these importance measures indeed allow to rank input variables by decreasing order of relevance, they still lack of a statistically sound procedure for selecting from their ranking a maximal set of variables while keeping the rate of false positives below a specified level. In particular, subset selection based on estimating generalization error rates by cross-validation generally does not imply a low false positive rate.

One possible way to derive from tree-based variable importance measures a procedure for selectively identifying relevant variables would be to mimic the procedures that are used in the context of multiple hypothesis testing of univariate statistics (see Ge et al. (2003) for a review): the approach there is to rank the features according to a relevance score derived from a (univariate) hypothesis test, and then for each score threshold to estimate the rate of false positives (called the false discovery rate, or FDR) among the variables that have a score greater than the threshold. In this context, the FDR is usually estimated in a non parametric way by assessing the average scores derived when randomly permuting the output variable values of the dataset.

In this paper, we first assess this FDR estimation approach when the relevance scores are derived by tree-based (multivariate) importance measures. We show empirically that this simple procedure typically overestimates in an unpredictable way the real FDR and thus can lead to unreliable selections of relevant subsets. We explain this by the fact that, contrary to the univariate case, the tree-based importance scores of different variables are not independent of each other. We then propose a novel alternative procedure for assessing the presence of irrelevant variables among a subset of variables top-ranked by tree-based variable importance measures. For a given importance threshold, this procedure first assumes that all and only those variables which have received an importance higher than the threshold are truly relevant, and then estimates the probability that any of the other variables would receive an importance higher than the given threshold. Experiments suggest that the latter quantity (estimated by an appropriately adapted permutation scheme) allows indeed to rather well identify an importance threshold below which the risk of having at least one false positive rapidly increases. The procedure may thus serve to identify in a more robust way than the FDR based approach a maximal subset of truly relevant variables among those proposed by the importance scoring method.

The rest of the paper is organized as follows. Section 2 describes the particular tree-based supervised learning method, the corresponding variable importance measure, and the synthetic datasets that we will use in the paper for our empirical tests. Section 3 studies the classical permutation based FDR estimation scheme on these synthetic datasets, highlighting its main properties when applied to tree-based variable importances. Section 4 describes the proposed alternative approach, its permutation based estimation procedure and the empirical results obtained with the same synthetic datasets. Section 5 shows some results on a real biological dataset. Finally, section 6 concludes and gives a few directions for further research.

2. Tree-based ensemble learners and importance measure

In this paper, we focus on classification problems. We assume that we have at our disposal a learning sample of N input-output pairs drawn from some unknown probability distribution. The m input variables are denoted $f_i, i = 1, \ldots, m$.

Tree-based methods. The basic idea of classification trees (Breiman et al., 1984) is to recursively split the learning sample with binary tests based each on one input variable trying to reduce as much as possible the uncertainty about the output classification in the resulting subsets of examples. Single classification trees are usually very much improved by ensemble methods, which aggregate the predictions of several trees. In our experiments, we will consider two tree-based ensemble methods based on randomization, namely Random Forests (Breiman, 2001) and Extra-Trees (Geurts et al., 2006). For these two methods, we use the default parameter setting (i.e., the number K of variables randomly selected at each node fixed at the square root of m and no pruning) and grow ensembles of $T = 100$ trees, unless specified otherwise.

Variable importance measure. Several variable importance measures have been proposed in the literature for tree-based methods. In this paper, we consider a measure based on information theory (Wehenkel, 1998), which at each test node n computes the total reduction of the class entropy due to the split, defined by:

$$I(n) = \#S.H_C(S) - \#S_t.H_C(S_t) - \#S_f.H_C(S_f), \tag{1}$$

where S denotes the set of samples that reach node n, S_t (resp. S_f) denotes its subset for which the test is true (resp. false), $H_C(\cdot)$ is the Shannon entropy of the class frequencies in a subset, and # denotes the cardinality of a set of samples. For a single tree, the overall importance v_i of the ith variable is then computed by summing the I values of all tree nodes where this variable is used to split. For an ensemble of trees, the importances are averaged over all individual trees.

A property of this measure is that the sum of the importances of all variables for a tree is equal to the total mutual information brought by the tree about the classification variable, which in the case of unpruned trees is usually very close to the initial total entropy of the classification variable (Wehenkel, 1998). The sum of the importances, for a single tree as well as for an ensemble of trees, is thus usually almost constant for a given problem. For presentation purpose, the importances are often normalized for the different variables so that they sum up to 100%. Each importance is thus the percentage of the total information brought by the ensemble of trees that is due to this variable.

Artificial problems. To validate our methods in a context where relevant variables are perfectly known, we generated two artificial problems by adapting the Matlab code originally used to produce the *Madelon* dataset for the NIPS 2003 feature selection challenge[1]. Both problems are binary classification problems with continuous input variables.

[1] http://www.clopinet.com/isabelle/Projects/NIPS2003/

- **Dataset-3-20:** This dataset is composed of 200 objects and 20 variables. The first three are really relevant, while the others are pure Gaussian noise. The problem is such that the third variable is only relevant in combination with the first two.
- **Dataset-50-1000:** The second (larger) dataset is composed of 2000 objects and 1000 variables. 50 variables are relevant, among which 6 have been directly used to define the output and 44 are linear combinations of these 6 variables (these latter are thus redundant given the first 6 ones). The remaining 950 irrelevant variables are pure Gaussian noise.

3. False discovery rate

We assume now that from the learning sample, we have computed with a tree-based method the importance v_i of each input variable f_i ($\forall i = 1, \dots, m$) and, without loss of generality, we further assume that the features are ordered according to their importances, i.e.

$$v_1 \geq v_2 \geq \dots \geq v_m.$$

For a given importance threshold v, we consider that all variables whose importance is greater than v are relevant and our concern is to estimate the expected rate of truly irrelevant features among these variables, the so-called false discovery rate (FDR) (Storey and Tibshirani, 2003).

More formally, for a given importance threshold v, the FDR is defined as:

$$\mathrm{FDR}(v) = E\left[\frac{F(v)}{S(v)}\right], \tag{2}$$

where $S(v)$ is the number of variables considered relevant at threshold v and $F(v)$ is the number of those variables that are truly irrelevant. The expectation is taken over different random learning samples drawn from the (usually unknown) joint distribution of the variables (assuming that the algorithm is deterministic).

To select a subset of variables, one can then check the FDR for increasing values of the threshold v and choose the minimum value of v such that $\mathrm{FDR}(v) < \alpha$, where α is typically small and reflects the risk one is ready to accept in terms of false positives when selecting the variables.

3.1 Estimation by random permutations

To estimate this FDR, we adopt the same approximations as in Listgarten and Heckerman (2007). When the number of features is large, one can approximate the expectation of the ratio by the ratio of the expectations:

$$\mathrm{FDR}(v) = E\left[\frac{F(v)}{S(v)}\right] \approx \frac{E[F(v)]}{E[S(v)]}. \tag{3}$$

$E[S(v)]$ can be simply approximated by the observed $S(v)$, i.e. the number of variables with an importance greater than v in the original data. $E[F(v)]$ is approximated by the expectation $E[F(v)|H_0]$ over the null distribution H_0 stating that all variables are truly

Table 1: Permutation algorithm for the estimation of the FDR

Compute variable importances from the original data and assume, without loss of generality, that the variables are ordered according to their importance v_i, i.e. $v_1 \geq v_2 \geq \ldots \geq v_m$.

1. for $p = 1$ to P (typically $P = 1000$):
   - (a) Randomly permute the class labels.
   - (b) Compute variable importance values v_j^p, for $j = 1, \ldots, m$, from the permuted data.
   - (c) Compute $R_i^p = \#\{k : v_k^p \geq v_i\}$, for $i = 1, \ldots, m$.
2. Then at v_i, the FDR is estimated by

$$\widehat{\mathrm{FDR}}(v_i) = \frac{1}{P} \frac{\sum_{p=1}^{P} R_i^p}{i}, \quad \text{for } i = 1, \ldots, m.$$

irrelevant. In other words, $E[F(v)]$ is taken as the expected number of variables that get an importance greater than v when none of them are truly relevant. We simulate this null hypothesis by applying the same tree-based method that was used to produce the original importances on datasets obtained from the original data by randomly permuting the class labels. This permutation decorrelates the classification variable from the inputs, making them all irrelevant, but keeps the dependencies that exist between the features in the original data.

The algorithm of Table 1 describes the procedure that we use to estimate the FDRs for all observed importance value thresholds in a learning sample.

3.2 Experiments on artificial data

We apply this procedure on the two artificial problems described in Section 2. Since we perfectly know the relevant variables for these two problems, we are also able to compute the *observed* FDR for a given subset of selected variables, i.e. the proportion of irrelevant variables among those selected. Figure 1 plots the FDR estimated by the procedure of Table 1 and the observed FDR as a function of the rank of the variables on the two artificial datasets with Extra-trees and Random Forests. There is no important difference between the two methods. In all cases, we observe that the estimated FDR overestimates the observed FDR. On the small dataset, both methods are able to find the three relevant variables (as illustrated by the fact that the observed FDR is equal to 0 until rank 3). However, the curve of the estimated FDR that already starts increasing for the third variable wrongly suggests that this variable is a false positive. On the larger dataset, using a typical threshold of 0.05 on the estimated FDR leads to the selection of 26 variables with Extra-trees and 25 variables with Random Forests (all relevant in both cases), while the

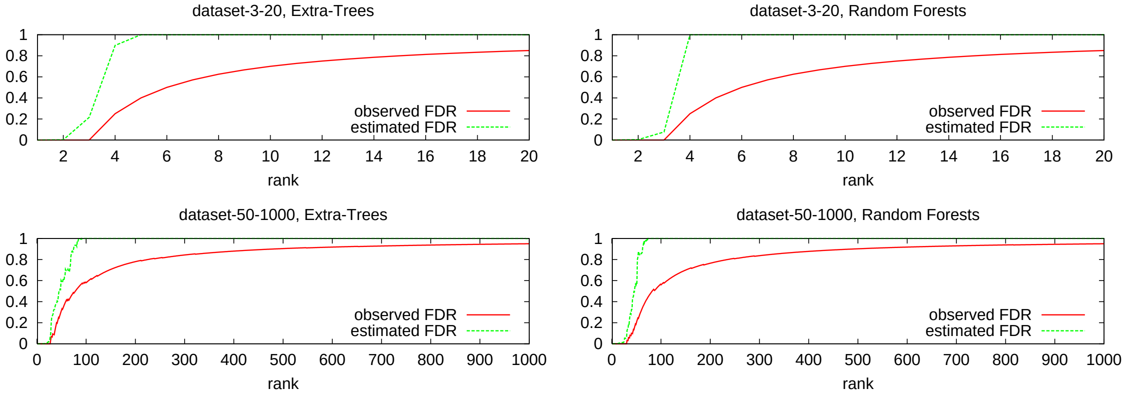

Figure 1: Estimated and observed FDR for increasing rank, top on Dataset-3-20, bottom on Dataset-50-1000, left with Extra-trees, right with Random Forests. ($P = 1000$, $T = 100$, $K = \sqrt{m}$)

same threshold on the observed FDR would lead to 28 variables with Extra-trees and 33 variables with Random Forests, with 5% of irrelevant among them.

3.3 Discussion

The overestimation of the FDR by the procedure of Table 1 can be explained, at least partially, by the fact that this procedure does not take into account the dependence between the importance values for different variables[2]. Indeed, as mentioned in Section 2, the sum of importances is roughly a constant for a given problem[3]. In consequence, if a variable brings a lot of information about the classification, there is much less left to be explained by the remaining variables, whether they are relevant or not. Thus, if a relevant variable receives a high importance, it potentially hides a less important but still relevant variable that may consequently receive an importance v_i which is small or even similar to that of an irrelevant variable in the permuted data. In this case, our estimation of $E[F(v_i)]$ from random permutations will be positively biased and thus our estimate of the FDR will be too conservative. This phenomenon is clearly apparent on the small artificial dataset, where the relevant variable f_3 gets an importance in the original ranking that is lower than the average importance obtained by the most important variable in the permuted data (see Figure 2).

Listgarten and Heckerman (2007) observed a similar effect when trying to estimate the false discovery rate among the edges predicted by a Bayesian network learning algorithm.

[2] Note that we are not talking here about the statistical dependence of the features that induces a dependence of their importances, however they are computed, but rather about the dependence between the importances that results from their joint computation by a multivariate approach.

[3] Note that this applies whether or not the importances are normalized.

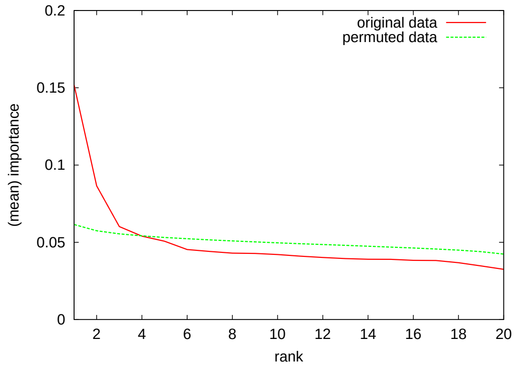

Figure 2: Variable importance as a function of the rank with Extra-trees on Dataset-3-20, from the original and the permuted data. In the latter case, importances are averaged over the P permutations. ($P = 1000$, $T = 100$, $K = \sqrt{m}$)

4. An alternative measure

In order to overcome this limitation of the FDR, we propose an alternative measure to be associated with each importance threshold v_i that takes into account the importances of the variables that are ranked above f_i. For a given importance threshold v_i, the procedure consists in computing the following conditional probability, which we call the conditional error rate (CER):

$$\mathrm{CER}(v_i) = P(\max_{k=i,\ldots,m} V_k^H \geq v_i | H_R^{1\to i-1}, H_I^{i\to m}), \tag{4}$$

where $H_R^{1\to i-1}$ denotes the hypothesis that all variables above f_i in the original ranking are relevant, $H_I^{i\to m}$ is the hypothesis that f_i and all the variables below f_i are irrelevant, and V_k^H ($k = 1, \ldots, m$) is the random variable denoting the importance of f_k under these two hypotheses. $\mathrm{CER}(v_i)$ is thus the probability that at least one irrelevant variable among m-i+1 gets an importance greater or equal to v_i, when these importances are computed under the assumption that variables $f_1, \ldots, f_{i-1}$ are all relevant.

This value can be interpreted as a measure of evidence against the hypothesis that all variables above the threshold v_i are relevant: a $\mathrm{CER}(v_i)$ close to one means that it is very likely to observe an irrelevant variable with an importance above v_i while a $\mathrm{CER}(v_i)$ close to zero means that it is very unlikely that an irrelevant variable could reach the threshold v_i. The limit between relevant and irrelevant variables in the ranking can then be determined by looking for the minimal threshold v_i such that $\mathrm{CER}(v_i) < \alpha$, for some small value of α.

Since the CER tries to detect at least one false positive above the threshold, we expect it to evolve much more abruptly than the FDR and thus that it will indicate more clearly the risk of selecting some irrelevant variables in the ranking.

Table 2: Permutation algorithm for the estimation of the CER

```
Compute variable importances from the original data and assume, without loss of
generality, that the variables are ordered according to their importance $v_i$, i.e.
$v_1 \geq v_2 \geq \ldots \geq v_m$.

  1. for $i = 1$ to $m$ :
     (a) for $p = 1$ to $P$ (typically $P = 1000$):
         • Keep the class labels and the values of the $i-1$ first variables of
           the original ranking fixed and randomly permute the values of the
           remaining variables using the same permutation vector for all
           variables.
         • Compute variable importance values $v_j^p$, for $j = 1, \ldots, m$, from the
           permuted data.
     (b) Then at $v_i$, the conditional probability (4) is estimated by:
           $\widehat{\mathrm{CER}}(v_i) = \frac{\#\{p : \max_{j=i,\cdots,m} v_j^p \geq v_i\}}{P}$,   for $i = 1, \ldots, m$.
```

4.1 Estimation by random permutations

We propose to estimate the probabilities (4) by random permutations as well. $H_R^{1 \to i-1}$ is approximated by keeping the class labels and the first i-1 variables unchanged (which amounts at considering that variables 1 to i-1 are truly relevant), while hypothesis $H_I^{i \to m}$ is simulated by randomly permuting the values of the variables i to m, that have an importance equal or smaller than v_i in the original ranking. To adhere as much as possible to the original joint distribution of the variables, they are furthermore permuted jointly, i.e. using the same permutation vector. The resulting procedure is described in Table 2.

Because the importance of the variables i to m in the random permutations are computed with the values of the variables 1 to i-1 being unchanged, these importances should not suffer from the same bias as in the estimation of the FDR. We thus expect that the algorithm of Table 2 will produce unbiased estimates of the CER and thus be more adapted to highlight the true frontier between relevant and irrelevant features than the procedure of Table 1 based on the FDR.

4.2 Experiments on artificial data

Figure 3 compares the CER as estimated by the procedure of Table 2 with the FDR estimated by the procedure of Table 1 on both datasets and with the two ensemble methods.

On the small dataset, the CER correctly starts increasing at the fourth variable. It thus gives more chance than the FDR to the third variable to be selected.

On the larger dataset, the transition region between low and high CER is quite well centered at the point where irrelevant variables start appearing in the ranking (indicated

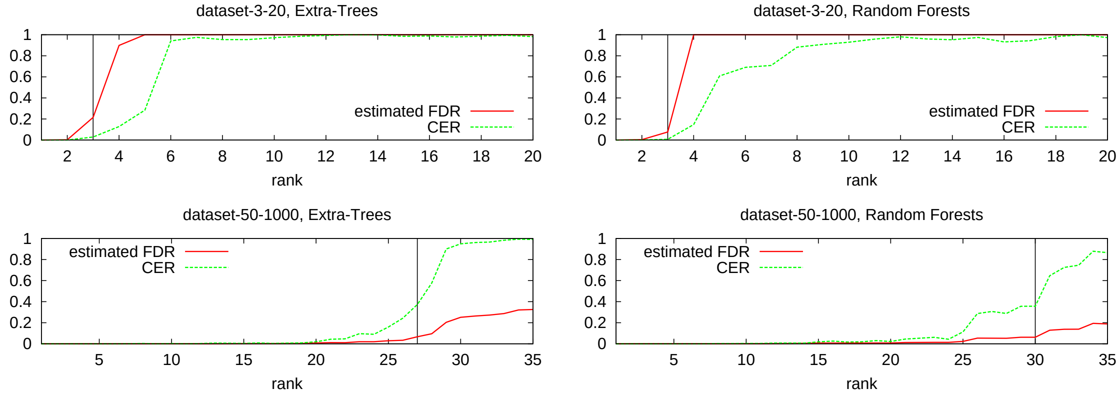

Figure 3: CER and FDR for increasing rank, top on Dataset-3-20, bottom on Dataset-50-1000, left with Extra-trees, right with Random Forests. The vertical line indicates the rank beyond which the observed FDR is greater than zero and the position where the observed CER switches from 0 to 1. ($P = 1000$, $T = 100$, $K = \sqrt{m}$)

by the vertical line). Setting a small threshold $\alpha = 0.05$ on the CER and looking for the last variable f_i in the ranking such that $\text{CER}(v_i) < \alpha$ leads to the selection of 22 variables with Extra-trees and 21 variables with Random Forests. In both cases, all the selected variables are truly relevant but the selection remains however quite conservative (with Extra-trees, the first 27 variables are relevant and with Random Forests, the first 30). This is because the transition region between low and high CER is quite large (especially for Random Forests).

4.3 Link with FWER based univariate procedures

The CER has a nice interpretation when the importances are actually derived from univariate statistics and a variable is, by definition, irrelevant when it satisfies the null hypothesis H_0 of the corresponding statistical test (e.g. v_i is the t-statistic associated to variable i). Indeed, in this case, importances v_i are computed independently of each other and probability (4) can thus be rewritten as follows:

$$\begin{aligned} P(\max_{k=i,\ldots,m} V_k^H \geq v_i | H_R^{1\rightarrow i-1}, H_I^{i\rightarrow m}) &= P(\max_{k=i,\ldots,m} V_k^H \geq v_i | H_I^{i\rightarrow m}) \\ &= P(\max_{k=i,\ldots,m} V_k^H \geq v_i | H_I^{1\rightarrow m}), \end{aligned} \quad (5)$$

where $H_I^{1\rightarrow m}$ is the hypothesis that all variables satisfy the null hypothesis H_0. Expression (5) corresponds precisely to the definition of Westfall and Young's *stepdown maxT adjusted p-values* (see Westfall and Young (1993); Ge et al. (2003)). The direct application of the procedure of Table 2 in this case thus produces estimates of these adjusted p-values by random permutations. Under some conditions about the distribution of the statistic, Westfall and Young showed that selecting all variables such that their adjusted p-values is lower than some threshold α guarantees that the family-wise error rate, or FWER (i.e. the probability to include at least one false positive among the selected variables) is lower than α. In our

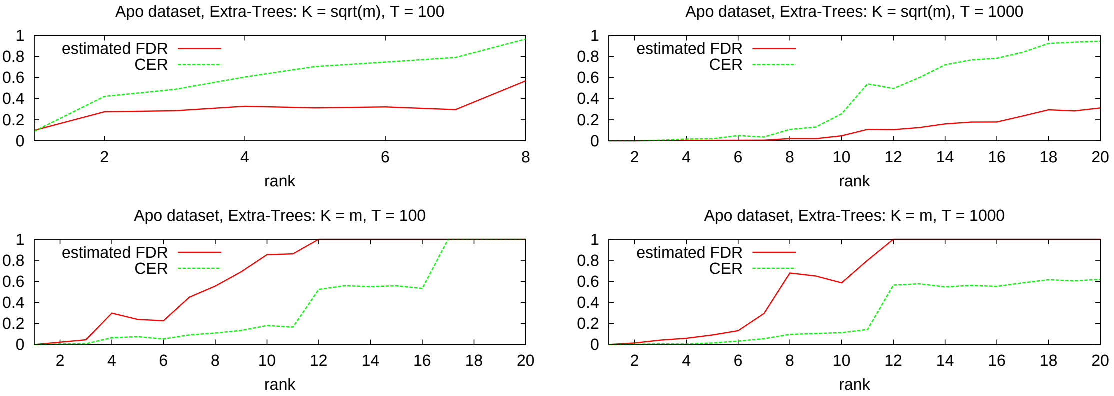

Figure 4: CER and FDR for increasing variable importance rank (Extra-Trees). Top left with ($K = \sqrt{m}, T = 100$), top right with ($K = \sqrt{m}, T = 1000$). Bottom left with ($K = m, T = 100$), bottom right with ($K = m, T = 1000$). ($P = 1000$, in all cases)

context, however, given the strong dependency between the importances that are computed by tree-based methods, it is not clear whether this guarantee still applies.

5. Experiments on a real dataset

To highlight the behaviour of both measures on a real problem, we run experiments on a biological dataset. The goal of the study (Callow et al., 2000) is the identification of the genes with altered expression in the livers of knock-out mice compared to control mice. The dataset[4] contains 5548 gene expression measurements for 16 mice divided into two classes: 8 wild-type mice and 8 mice whose Apo AI gene was knocked out. This dataset was also used in Ge et al. (2003) to compare several statistical procedures for controlling multiple testing issues for univariate statistical tests.

Although the truly relevant variables are unknown, in Callow et al. (2000), eight variables were identified as differentially expressed using univariate statistical tests and their relevance was experimentally confirmed. We expect that multivariate approaches will at least highlight these eight variables and we will thus check their presence in the rankings below. Note however that this does not mean that only those eight variables are relevant. All additional variables found by our multivariate procedures certainly deserve to be checked experimentally.

On this dataset, we apply the Extra-trees algorithm with four different settings of its parameters, i.e. the number K of variables that are randomly selected at each node and the number T of trees in the ensemble: ($K = \sqrt{m}, T = 100$), ($K = \sqrt{m}, T = 1000$), ($K = m, T = 100$), and ($K = m, T = 1000$). The estimated FDR and CER as a function of the ranking are plotted in Figure 4 in all four cases.

[4] http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html

Several conclusions can be drawn from these plots. First, using $K = \sqrt{m}$ and $T = 100$ does not bring interesting results on this dataset. The method is unable to distinguish truly relevant variables from randomly permuted ones, which translates into a high value of both the FDR and the CER. Simply increasing the number of ensemble terms already gives much better results in terms of the number of variables that appear as relevant. A threshold of 0.05 on the CER selects a subset of 7 variables and a threshold of 0.05 on the FDR gives 10 variables that actually contains the 8 variables identified in Callow et al. (2000). Only 3 of them were present in the top 10 variables with $T = 100$, confirming that the ranking is indeed improved by increasing T. It is interesting to note that, on the other hand, increasing T from 100 to 1000 does not affect the error rate (which is equal to 25% in both cases, as estimated by leave-one-out), meaning that here accuracy would not be a relevant criterion to assess the quality of the ranking. The high improvement of the FDR and CER values when T is increased is here a consequence of the very high ratio between the number of variables and the number of examples that makes the random trees, and thus the corresponding rankings, highly unstable and thus requires to average a very large number of trees for stabilization.

When K is increased to its maximum value (i.e. randomization is reduced), the CER is also very much improved, even with $T = 100$. It shows an abrupt change between the 11th and the 12th variables, suggesting that about 11 variables are relevant. As a confirmation, the 8 variables identified in Callow et al. (2000) are again among these 11 ones. On the other hand, the estimated FDR seems to be highly overestimated in this case. We explain this by the fact that increasing K in Extra-trees makes the model less random and thus increases the importance of the top ranked variables relatively to the low ranked ones, thus emphasizing the phenomenon highlighted in Section 3.3 and responsible for the overestimation of the FDR.

6. Conclusions

In this paper, we have proposed and evaluated two statistical procedures to extract a subset of truly relevant variables from multivariate importance values obtained from tree-based supervised learning methods.

The first method is a direct adaptation of FDR estimation schemes based on permutations used in the context of univariate statistics. Unfortunately, we found that this procedure, because it does not take into account the dependencies of the importance values derived by the tree-based methods, often strongly overestimates the actual FDR and can thus potentially lead to overly conservative selections of relevant subsets.

We therefore have proposed a new statistic, called the *conditional error rate* (CER), that explicitly takes into account the dependencies between the importances and thus leads to more robust feature selection. We have also proposed a permutation procedure to empirically estimate the CER values associated to any given importance ranking scheme, and compared its performances with the FDR-based scheme on both artificial and real datasets. Although further experiments are needed to better validate the CER procedure, we believe that our results are already quite informative and very encouraging. In particular, they suggest that the CER-based procedure leads to a more robust scheme for the selection of relevant variables among large numbers of irrelevant ones. As a byproduct, our simulations also suggest that for reliable identification of relevant features with tree-based ensemble methods one should use very large ensembles, much larger than those needed for accurate prediction. This in turn highlights the fact that prediction accuracy may not be an appropriate measure for the identification of relevant features.

One drawback of the current CER estimation procedure is that it is very computationally demanding: $P \times m$ models are needed for the estimation of the CER for all possible importance thresholds, where m is the number of variables and P is the number of random permutations. The computing times can however be decreased, e.g. by stopping the procedure as soon as the estimated probability is greater than some threshold, as we are typically only interested in small values of the CER. In problems with a very high proportion of irrelevant variables, this truncation will lead to a very significant speed-up of the procedure.

In terms of future works, it would be interesting to investigate better procedures to estimate the FDR. As noted in Section 3.3, one of the reasons of the overestimation of the FDR is the overestimation of $E[F(v)]$ that results from the simultaneous permutation of all features. We have tried to adopt the same sequential approach for estimating this term as for the CER, i.e. only permuting the m-i+1 last variables and estimate $E[F(v_i)]$ as the expected number of permuted variables whose importance exceeds threshold v_i. This yields better estimates of the FDR for small i but when i grows to m, the number of permuted variables, and thus the FDR, decreases to 0. Further investigations are thus necessary to really assess the conditions of validity of this procedure. Several improved permutation procedures have also been proposed to better estimate the FDR in the context of univariate tests (e.g. Xie et al. (2005)), and the extension of these latter approaches to our multivariate context is also an interesting future work direction.

Two related approaches to selectively identify relevant features from a ranking have been proposed in (Stoppiglia et al., 2003) and (Tuv et al., 2006). The common idea of both methods is to include random features in the learning sample and then to exploit their rank among the original features to determine a relevance threshold. Stoppiglia et al. (2003) suggest to introduce only one such random feature and they apply this idea in the context of linear models where the distribution of the rank of the random feature can be computed analytically. They also suggest in their conclusions that this distribution could be computed empirically for any other ranking method. Along a similar line, Tuv et al. (2006) introduce as many random features as there are input variables in the original problem and generate these features by permuting the original variables. Relevant variables are then defined as those variables that receive an importance significantly greater than their permuted counterpart. In Tuv et al. (2006), this idea is actually wrapped into a gradient boosting type algorithm that iteratively selects subsets of important variables but it could be also applied in a single step as an alternative to our approach.

While these two approaches are potentially faster than our CER procedure (as they require only one round of randomization/permutation) and take also into account multivariate effects, they nevertheless introduce an additional degree of freedom, which is the number of random features to introduce into the problem. Some preliminary simulations on our artificial datasets showed indeed that Stoppiglia et al. (2003)'s method (introducing a single random feature) does not seem to take into account multiple test problems. Indeed, the probability of the random feature to reach a high rank, and thus the number of rejected features, will actually decrease as the number of truly irrelevant features in the original problem increases. On the other hand, with as many random features as there are input variables, Tuv et al. (2006)'s method significantly changes the nature of the problem (by doubling the number of features), which could potentially affect the ability of the method to detect relevant variables. A detailed comparison of our method with these two approaches will be the subject of future works.

Another important direction of investigation is of course the application of our procedures in the context of other multivariate importance measures derived from machine learning algorithms. In the context of tree-based methods, our analysis should carry over straightforwardly to regression measures based on variance reduction (Breiman et al., 1984). It would be interesting also to consider other importance measures such as, for example, Breiman's permutation based importance measure (Breiman, 2003).

On the other hand, we have already carried out some preliminary experiments with variable importances derived from the weights of linear SVM models (Guyon et al., 2002). One problem however with these importances is that their interpretation is dependent on one hand on the scaling of the input variables and on the other hand on the specific margin that can be achieved given the features (since the margin is proportional to the norm of the weight vector). This scaling instability suggests that some further caution should be taken when comparing the original importances with importances derived from randomly permuted data as required in our procedure.

Acknowledgments

This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. VAHT is recipient of a F.R.I.A. fellowship and PG is Research Associate of the F.R.S.-FNRS. The authors thank the anonymous reviewers of this paper for their useful suggestions. The scientific responsibility rests with the authors.

References

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman. *Setting up, using, and understanding random forests V4.0.* University of California, Department of Statistics, 2003.

L. Breiman, J. H. Friedman, R. A. Olsen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International (California), 1984.

M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, 10(12):2022–2029, 2000.

Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. Technical Report 633, Department of Statistics, University of California, Berkeley, 2003.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 36 (1):3–42, 2006.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 9(1-3):389–422, 2002.

J. Listgarten and D. Heckerman. Determining the number of non-spurious arcs in a learned DAG model: Investigation of a bayesian and a frequentist approach. In *Proceedings of UAI*. UAI Press, 2007.

Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414, 2003.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, August 2003. ISSN 0027-8424. URL `http://dx.doi.org/10.1073\%2Fpnas.1530509100`.

E. Tuv, A. Borisov, and K. Torkkola. Feature selection using ensemble based ranking against artificial contrasts. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2006)*, 2006.

L. Wehenkel. *Automatic learning techniques in power systems.* Kluwer Academic, Boston, 1998.

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons, 1993.

Y. Xie, W. Pan, and A. Khodursky. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21 (23):4280–4288, 2005.