Random Forests Based Feature Selection for Decoding fMRI Data

Robin Genuer[1,2], Vincent Michel[1,2,3,5], Evelyn Eger[4,5], and Bertrand Thirion[3,5]

[1] Université Paris-Sud 11, Mathématiques, Orsay, France
[2] Select team, INRIA Saclay-Île-de-France, France
[3] Parietal team, INRIA Saclay-Île-de-France, France
[4] INSERM U562, Gif/Yvette, France
[5] CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

Abstract. In this paper we present a new approach for the prediction of a behavioral variable from Functional Magnetic Resonance Imaging (fMRI) data. The difficulty in this problem comes from the huge number of image voxels that may provide relevant information with respect to the limited number of available images. A very common solution consists in using feature selection techniques, i.e. to evaluate the significance of each individual brain region with respect to the target information, and then to use the best ranked features as input to a classifier, such as linear Support Vector Machines (SVM; we take this as the *reference method*). However, this kind of scheme ignores the correlations between features, so that it is potentially suboptimal, and it does not generally provide an interpretable pattern of predictive voxels. Based on Random Forests, our approach provides an accurate auto-calibrated framework for selecting a set of very few jointly informative regions. Comparisons with the reference method on real data show that our approach yields a little bit higher classification performance, but the real gain comes from the sparsity of our variable selection.

Keywords: Feature selection, Variable Importance, Random forests, Classification, fMRI

1 Introduction

A new way of analyzing neuroimaging data consists in assessing how well behavioral information or cognitive states can be predicted from brain activation images such as those obtained with functional magnetic resonance imaging (fMRI) (Cox and Savoy (2003)). This approach opens the way to understanding the mental representation of various perceptual and cognitive parameters. Indeed, certain neuronal populations are thought to activate specifically when a certain perceptual or cognitive parameter reaches a given value. The accuracy of the prediction of the target behavioral variable, as well as the spatial layout of predictive regions can provide valuable information about functional brain organization; in short, it helps to *decode* the brain system (Dayan and Abbott (2001)).

The main difficulty in this procedure is the huge dimensionality of the data, with far more features than samples. In this article, the samples will refer to the activation parameter maps resulting from a General Linear Model (GLM), the features being the voxel-based activation values. The large number of features leads to overfitting and thus to a dramatic decrease in prediction accuracy. Feature selection is thus mandatory, and is often performed by a mass-univariate selection based on F-test statistics. However, this classical approach is not well suited for neuroimaging as it does not cope with the multivariate structure of the data.
In order to improve the predictive framework, we introduce a new multivariate method of feature selection based on Random Forests (RF henceforth). RF is an increasingly used statistical method introduced in Breiman (2001). It gives outstanding results in prediction for lots of diverse applications. In addition, it computes a variable importance that can be used to select variables. Our RF-based algorithm uses the variable importance index in a feature selection framework. This variable selection procedure comes from Genuer et al. (2010), where one can find more information about RF variable importance.
After introducing the Random Forests and the RF-based algorithm, we show that our self-calibrated method performs an accurate feature selection, yielding a little bit better classification score than the reference technique, while keeping much less jointly informative variables. And this very sparse aspect of our variable selection method can help understanding functional brain organization. Let us finally emphasize that all along this paper, we distinguish two objectives: interpretation, which aims at selecting all the variables the most related to the response variable (even if they are correlated to each other); and prediction, which focuses on building a model involving the smallest subset of variables sufficient to make accurate predictions.

2 Methods

Let $(Y_1, \ldots, Y_n)$ represent the behavioural data to be fitted ($\forall i$, $Y_i \in \{1, \ldots, c\}$, where c is the number of classes) related to a set of n parameter maps obtained with a GLM, where each image corresponds to one stimulus presentation; $(X_1, \ldots, X_n)$ are the m-dimensional activation maps ($X \in \mathbb{R}^m$) and m is the number of features (voxels or parcels). In fMRI data, we have $n \ll m$, so that feature selection is mandatory.

Random Forests
The principle of random forests is to aggregate many binary decision trees built on several bootstrap samples drawn from the learning set. The bootstrap samples are obtained by uniformly drawing n samples among the learning set with repetition. The decision trees are fully developed binary trees and the split rule is the following:
First, the whole dataset (also called the root of the tree) is split into two sub-

sets of data (called two children nodes). To do that, one randomly chooses a given number m_{try} of variables, and computes all the splits only for the previously selected variables. A split is of the form $\{X^i \leq s\} \cup \{X^i > s\}$, which means that data with the i-th variable value less than the threshold s go to the left child node and the others to the right one. Finally the selected split is the one leading to the most homogeneous children nodes (i.e. subsets associated to the same class).
Then, one restraints to one child node, randomly chooses another set of m_{try} variables and calculates the best split. And so on, until each node is a terminal node, i.e. it comprises observations associated with the same class.
A new data item X, starting in the root of the tree, goes down the tree following the splits and falls in a terminal node. Then the tree predicts for X, the common class $\hat{Y}$ of the data in this terminal node. To finally get the RF classifier, one aggregates all the tree classifiers through a majority vote heuristic: for a new observation, each tree predicts a class and RF finally returns the most popular class.
Inside the variable selection procedure, we use an estimation of the prediction error directly computed by the RF algorithm. This is the Out Of Bag (OOB) error and is calculated as follows. Fix one data in the learning sample, and consider all the bootstrap samples which do not contain this data (i.e. for which the data is "out of bag"). Now perform a majority vote only among trees built on these bootstrap samples. After doing this for all data, compare to the true classes and get an estimation of the prediction error (which is a cross-validated error estimate).
Let us now detail the computation of the RF variable importance for the first variable X^1. For each tree, one has a bootstrap sample associated with an OOB sample. Predict the OOB data with the tree classifier. Now, randomly permute the values of the first variable of the OOB observations, predict these modified OOB data with the tree classifier. The variable importance (VI henceforth) of X^1 is defined as the mean increase of prediction errors after permutation. The more the error increases, the more important the variable is (note that it can be slightly negative, typically for irrelevant variables).

Variable selection procedure
Let us give (following Genuer et al. (2010)) some details about the variable selection procedure that we use here. We apply it on a simulated learning set of size $n = 100$ from the classification toys data model, introduced in Weston et al. (2003), with $m = 200$. It is an equiprobable two-class problem, $Y \in \{-1, 1\}$, with 6 true variables, the others being some noise.
The results are summarized in Figure 1. The true variables (1 to 6) are respectively represented by $(\triangleright, \triangle, \circ, \star, \triangleleft, \square)$. Based on to the learning set, we compute 50 forests with $n_{tree} = 2000$ and $m_{try} = 100$, which are values of the main parameters considered as well adapted for VI estimation (for more details, see Genuer et al. (2010)).

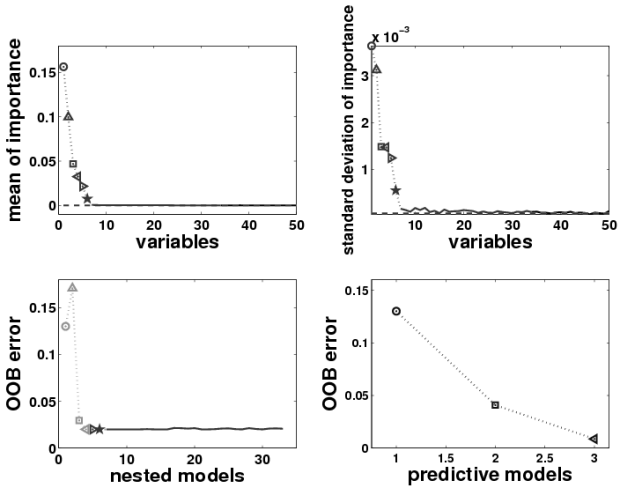

Fig. 1. Variable selection procedure for a toy dataset. The top left graph shows the variable ranking. The curve of the top right graph is used to determine the threshold (represented by the horizontal dashed line) needed in Elimination step. OOB errors of the nested models are plotted in the bottom left graph to illustrate the Interpretation step. The bottom right graph stands for Prediction step.

Let us detail the four main steps of the procedure:
Variable ranking: First the variables are sorted according to the VI (averaged from the 50 runs) in descending order. Note that true variables are significantly more important than the noisy ones.
Elimination step: Keeping this order in mind, the corresponding standard deviations of VI are plotted. A threshold for importance is computed using this graph, and only the variables with an importance exceeding this level are kept. More precisely, the threshold is set as the minimum prediction value given by a CART model fitting this curve (for details, see Breiman et al. (1984)).
Interpretation step: Then, OOB error rates of the nested random forests models are computed; starting from the one with only the most important variables, and ending with the one involving all important variables kept previously. The set of variables leading to the smallest OOB error is selected.
Prediction step: Finally a sequential variable introduction with testing is performed: a variable is added only if the error gain exceeds a data-driven threshold (see Genuer et al. (2010)). The rationale is that the error decrease must be significantly greater than the average variation obtained by adding noisy variables.

3 Experiments and Results

Real Data

We used a real dataset related to an experiment on the representation of objects Eger et al. (2008). During the experiment, twelve healthy volunteers

viewed objects of three different sizes and four different shapes, with 6 repetitions of each stimulus (refering to 6 sessions), resulting in a total of $n = 72$ images by subject. Functional images were acquired on a 3-T MR system with eight-channel head coil (Siemens Trio, Erlangen, Germany) as T2*-weighted echo-planar image (EPI) volumes. Twenty transverse slices were obtained with a repetition time of 2 s (echo time, 30 ms; flip angle, 70°; $2 \times 2 \times 2$-mm voxels; 0.5-mm gap). Realignment, normalization to MNI space and GLM fit were performed with the SPM5 software. For our analysis we used the resulting session-wise parameter estimate images. The four different shapes of objects are pooled across the three sizes, and we are interested in discrimination between shapes. We used parcellation as a preprocessing, which allows important unsupervised reduction of dimensions. Parcellation uses Ward's algorithm (hierarchical agglomerative clustering) to create groups of voxels which have similar activity across trials. Thus, the signal is averaged in each parcel. The number of parcels created is fixed to 1000 for the whole brain.

Feature selection results for one subject

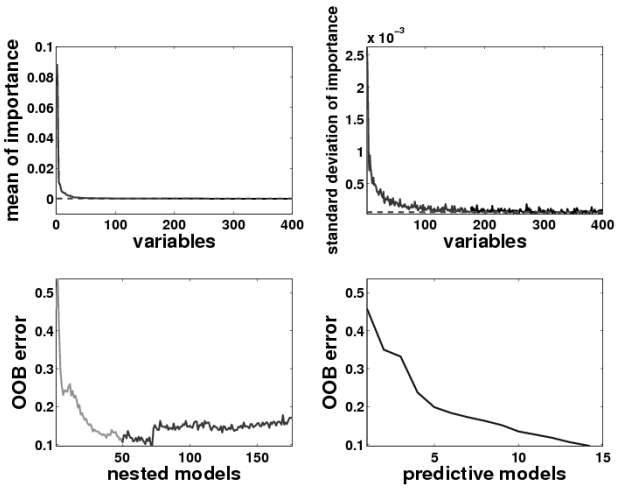

Fig. 2. Variable selection procedure for one subject. The graphs follow the exact same description as in Figure 1.

We apply the procedure described in Section 2 for the subject 2 of the study. The results are plotted in Figure 2. The horizontal dotted line of the top graphs indicates the threshold, computed using standard deviations of VI (see the top right graph) and used in the top left graph to eliminate variables of small importance. Starting with all the 1000 variables, this elimination step retains 176 variables. The minimum OOB error rate in the bottom left graph is obtained by the RF model involving 50 variables, which constitute the interpretation set. Finally, the prediction procedure, illustrated by the bottom right graph, selects 15 variables.

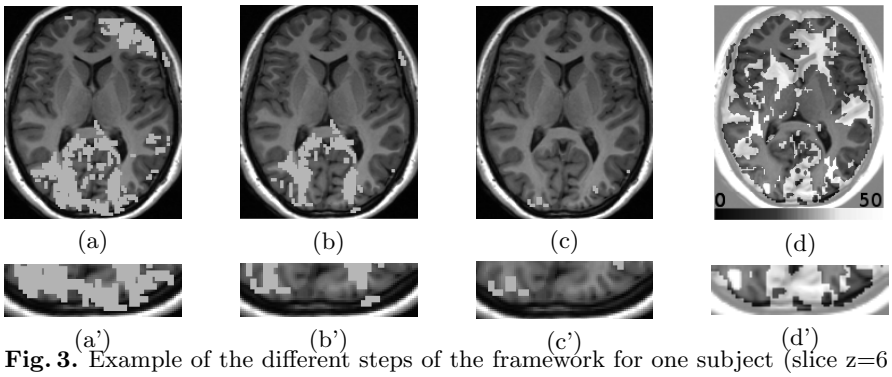

Fig. 3. Example of the different steps of the framework for one subject (slice z=6 mm). (a) Selected parcels after Elimination Step. (b) Selected parcels after Interpretation Step. (c) Selected parcels after Prediction Step. (d) Shows the parcels selected by the reference method, and their F-test values. (a'), (b'), (c') and (d') are magnifications of the occipital part.

Figure 3 shows the selected parcels for the different steps of the algorithm in one axial slice for subject 2: sub-figures (a), (b) and (c) represent the variables selected in the Elimination step, Interpretation step and Prediction step, and (d) represent the variables selected by the reference method. Sub-figures (a'), (b'), (c') and (d') are magnifications of the occipital part. During the interpretation step, our algorithm keeps only three regions of the occipital cortex, reducing the features to a much smaller sets while keeping an accurate prediction (see Figure 4). In addition, the prediction step (c) allows to avoid redundancy in the features. The selected regions are different between the two hemispheres, while the interpretation step retained more symmetric regions. Finally, comparison whith sub-figure (d) highlights the most beneficial aspect of our method: we select very localised informative regions, while the reference method keeps lots of regions distributed in all brain.

Prediction results for the whole data

We perform a leave-one-session-out cross-validation: we successively train the classifier with all the sessions except one, and report the performance of the trained classifier on the left out session. Importance-based feature selection was applied independently on the twelve datasets. The results are shown in Figure 4. The first row represents the classification score of RF for each subject (from left to right: all parcels, after Elimination step, after Interpretation step and after Prediction step). The average number of selected parcels across subjects is noted above each histogram, with the average classification score across all subjects.

The first graph of the second row shows the results of a cross-validated linear SVM: the optimal number of parcels to be kept (from 50 to 1000 parcels with a step of 50) for the linear SVM is selected using the F-statistic, by leave-one-out validation on the training set. The average number of selected

parcels across subjects is equal to 350. The three last histograms of the second row show the results of a linear SVM: the parcels are selected by using a F-statistics, and the number of features used is equal to the number of parcels found by the three different steps of the RF-based algorithm. We can see that our algorithm gives better accuracy for the three steps of selection than the reference method (cross-validated linear SVM). And the three last histograms of the second row illustrate the fact that a linear SVM (coupled with F-test) do not manage to keep good accuracy with as few features as selected by our method.

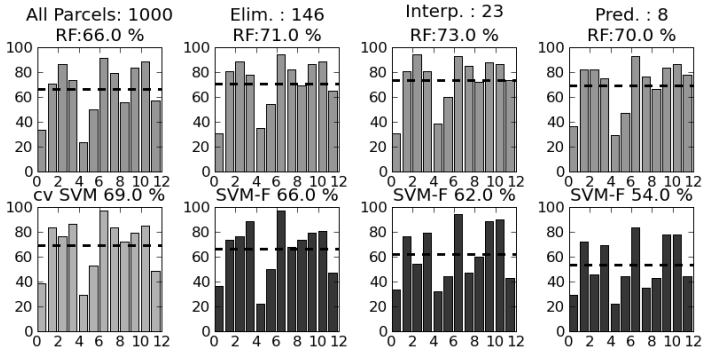

Fig. 4. Results on real data: rate of correct object identification, using the mean signal of 1000 parcels of the brain volume (chance level=25 %). The first row shows the prediction accuracy in each individual dataset, the mean classification score and the number of selected parcels for (from left to right) the whole brain, the Elimination step, the Interpretation step and the Prediction step. The first graph of the second row represents the results for the reference method. The three last histograms of the second row show the prediction accuracy of a linear SVM trained with the same number of parcels as above, but selected by F-statistics.

4 Discussion

This work presents the first application of a RF-based feature selection technique to brain state decoding. We show that it is competitive with state of the art method (univariate selection followed by linear SVM classifier). More importantly, the insensitivity of the correct classification rates along the different steps of feature reduction that is observed in Figure 4 for the RF model shows that our strategy manages to extract the statistical information of the data: it keeps much of the information while significantly reducing the dimension. This suggests that the multivariate RF variable importance index performs better than the classical univariate F-test score to detect the most predictive variables. Another noticeable aspect of the proposed procedure is that it is entirely data-driven: at each step of the procedure, thresholds are computed using only the data. So this procedure can adapt to lots of different applications, without the need of adding prior information (like e.g. a number of variables to be selected).

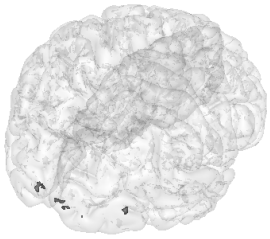

Fig. 5. Regions selected in at least 3 subjects among 12 by the last step of the RF-based selection. The MNI coordinates are :
[18, -102, 6]mm
[10, -100, 4]mm
[-12, -96, 0]mm
[50, -78, 6]mm.

From a neuroscientific point of view, we can notice that the spatial distribution of the selected parcels is quite informative: first, by avoiding redundancy, the algorithm is able to focus on few extremely precise regions of the brain without loss of accuracy. Moreover, starting from whole brain, the algorithm selects very few parcels in the occipital cortex, corresponding to visual areas. If we look at the regions selected for 3 subjects or more among the 12 subjects by the last step of the RF-based selection (see Figure 5), there are only few regions in the early visual cortex, and a slightly more anterior parcel. This is consistent with the fact that early visual cortex contains highly reliable signals discriminative of feature/shape differences between object exemplars, as long as no generalization across image changes is required (Cox and Savoy (2003) and Eger et al. (2008)).

Conclusion In this article, a multivariate and threshold-free feature selection algorithm based on Random Forests, yields an accurate selection for fMRI data analysis, and creates a highly informative set of very few features. Results on real data show the benefits of our approach for both interpretation and prediction, with higher accuracy and higher sparsity than the reference method.

References

BREIMAN, L. (2001): Random Forests. *Machine Learning 45 ,5-32.*

BREIMAN, L., FRIEDMAN, J., OLSHEN, R.A. adn STONE, C.J. (1984): *Classification and Regression Trees.* Chapman & Hall.

COX, D.D. and SAVOY, R.L. (2003): Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage 19 (2),261–270.*

DAYAN, P. and ABBOTT, L.F. (2001): *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* The MIT Press.

EGER, E., KELL, C. and KLEINSCHMIDT, A. (2008): Graded size sensitivity of object exemplar evoked activity patterns in human LOC subregions. *Journal of Neurophysiology 100 (4) , 2038-47.*

GENUER, R., POGGI, J.-M. and TULEAU, C. (2010): Variable selection using random forests. *Pattern Recognition Lett.* doi:10.1016/j.patrec.2010.03.014

WESTON, J., ELISSEEF, A., SCHOLKOPF, B., TIPPING, M., KAELBLING, P. (2003): Use of the Zero-Norm with Linear Models and Kernel Methods. *Journal of Machine Learning Research 3 , 1439-1461.*