MUTUAL INFORMATION-BASED FEATURE SELECTION ENHANCES FMRI BRAIN ACTIVITY CLASSIFICATION

Vincent Michel[1,2], *Cécilia Damon*[1], *Bertrand Thirion*[1]

[1] INRIA Saclay Parietal [2] Université Paris-Sud 11

ABSTRACT

In this paper, we adress the question of decoding cognitive information from functional Magnetic Resonance (MR) images using classification techniques. The main bottleneck for accurate prediction is the selection of informative features (voxels). We develop a multivariate approach based on a mutual information criterion, estimated by nearest neighbors. This method can handle a large number of dimensions and is able to detect the non-linear correlations between the features and the label. We show that, by using MI-based feature selection, we can achieve better perfomance together with sparse feature selection, and thus a better understanding of information coding within the brain than the reference method which is a mass univariate selection (ANOVA).

***Index Terms*—** Features selection, Mutual information, Brain reading, Classification, fMRI

1. INTRODUCTION

Over the past five years, there has been considerable interest in classifying brain activity from functional MR (fMRI) images to compare the response to different conditions and find which brain regions discriminate between two states ("brain-reading", see [1]). This technique consists in finding a combination of voxel-based (or ROI-based) responses that best predicts some target information (e.g. the stimulation condition). This problem, as a classification problem in high-dimension spaces (fMRI images are about $N = 1000 - 2000$ voxels when considering ROIs, but are most typically of 10^4 to 10^5 voxels when considering the whole brain), is plagued with the curse of dimensionality and thus requires the use of features selection. Some standard techniques for features selection have been used in fMRI (see [2] for review) : Anova, possibly in conjunction with spatial averaging (parcellation), univariate mutual information, but also multivariate methods, e.g. Singular Value Decomposition (SVD), Manova.

Mutual Information (MI) is known to characterize the dependence between random variables beyond the second order moment (correlation) and can be used for multivariate selection, by choosing the features which jointly maximize the prediction given a set of previously selected features. Given the few number of samples (here, the number of fMRI images used for learning), MI cannot be reliably estimated by the joint density of the features and the target ; in the case of high dimensional problem, some better estimators are based on the k nearest neighbors (knn) (see [3]), which can handle a large number of dimensions, with reasonable variance.

Rossi et al. ([4]) have developed a features selection technique based on Mutual Information (MI) for regression in spaces of very high dimension. In this work, we adapted this method to classification problem in which the target variable Y takes discrete values. Our approach can be seen as an improvement of the method developed in [5]. This new algorithm (MIFS) has been tested on simulated and real data and the selected features have been used in conjunction with SVM and RVM classifiers. The performances in generalization are shown to outperform ANOVA feature selection.

2. METHODS

2.1. Estimation of Mutual Information

Let X be a set of random variables that may be used to explain Y. The entropy of X is defined as $H(X) = -\int P(x) \log P(x) dx$, and the MI between X and Y is defined as the Kullback-Leibler divergence between the distributions $P(X,Y)$ and $P(X) \times P(Y)$:

$$MI(X,Y) = \int_{x,y} P(x,y) \times \log \frac{P(x,y)}{P(x)P(y)} dxdy$$

Let ϵ be twice the distance from a point z in Z to is k^{th} nearest neighbor in the space $Z = (X,Y)$ (with the maximum norm). Let d be the dimension of X and c_d the volume of the d-dimensional unit ball. Kraskov et al. ([3]) propose the following estimators of $H(X)$ and $MI(X,Y)$ where Y is real:

$$H(X) = -\psi(k) + \psi(N) + \log(c_d) + \frac{d}{N} \sum_{i=1}^{N} \log(\epsilon(i))$$

$$MI(X,Y) = \psi(k) + \psi(N) - \frac{1}{N} \sum_{i=1}^{N} (\psi(n_x+1) + \psi(n_y+1))$$

where ψ is the digamma function : $\psi(x) = \Gamma(x)^{-1} d\Gamma(x)/dx$, and nx (respectively ny) is the number of points with a distance to z in the space X (respectively in the space Y) strictly inferior to $\epsilon/2$.

Let us adapt it to the case where Y is finite : MI is equivalently defined as $MI(X,Y) = H(X) - H(X|Y)$:

$$MI(X,Y) = H(X) - \sum_{l=1}^{l_{\max}} H(X|Y=l)p(Y=l)$$

Then, let $A = \log(c_{dx}) + \frac{dx}{Nl}\sum_{i=1}^{N}\log(\epsilon(i))$, $n_x^l(i)$ the number of points between $x(i)$ and its knn which have the same label l, and N_l the number of points having the label l :

$$H(X|Y=l) = -\frac{1}{N_l}\sum_{i=1}^{N_l}\psi(n_x^l(i)+1) + \psi(N_l) + A \text{ , thus}$$

$$MI(X,Y) = B + \frac{1}{N}\sum_{i=1}^{N}\left(\psi(n_x^{l(i)}(i)+1) - \psi(N_{l(i)})\right)$$

with $B = \psi(N) - \psi(k)$.

2.2. Features selection

The features selection process is adapted from [4] and [6] (see Fig. 1). Firstly, let S and R be the sets of selected features and the group of features that might be chosen : we start with $S = \emptyset$ and $R = \{x_i\}, i = 1..N$ and the algorithm will stop when R is empty. This algorithm uses an hybrid stepwise selection. The forward strategy adds at each step the most informative feature given the previously selected ones. The backward strategy removes from R all the features which are not informative at this step : we indeed assume that those features will not be informative in the next steps.

In order to select a feature, we compute at each step, for each dimension x in R, the value $MI_1 = MI(S\bigcup\{x\}, Y)$, which yields the amount of information about Y present in S and x. Let x^π be a permutation of the values of x across samples, and let $MI_2 = MI(S\bigcup\{x^\pi\}, Y)$. The distribution of MI_2 is computed by drawing randomly P permutations. We obtain the following approximate p-value : $p = \frac{1}{P}\sum_{k=1}^{P}(MI_1 < MI_{2,k})$. If this p-value is below a predefined threshold α, one can consider that this dimension is informative; otherwise we can remove it from R. In order to avoid redundancy of information, we also remove all the previously pre-selected features x for which $MI(S\bigcup\{x\}, Y) < MI(S,Y)$ Finally, we select the dimension with the highest value of MI_1 and keep the other pre-selected ones in R.
In this algorithm, three parameters are used :
The threshold for the p-value α: typical values are between 0.05 and 0.001. It is the most important factor in the algorithm and can be interpreted as a quality control that we require for the dimensions to be selected : a low value of α will discard usefull information, while a high value of α will allow the inclusion of weakly informative features, thus yielding overfitting.
The number of neighbors in the MI estimator k: typical values are between 10 and 30 (see Kraskov et al. [3]).
The number of permutations P: chosen given the value of α, by $P \gg \frac{1}{\alpha}$.

We also combine the feature selection by MI with a preprocessing, parcellation, which allows important unsupervised reduction of the dimension of X. Parcellation uses hierarchical agglomerative clustering to create groups of voxels which have similar time courses : from 1500 voxels, we will create about 50 parcels. The signal is averaged in each parcel.

Finally, cross-validation is used to obtain a performance in generalization, and to compare the different features selection techniques. We use two types of classifiers: SVM (Support Vector Machines) with a linear Kernel (see [7]), and Relevance Vector Machines) with a linear Kernel (see [8]).

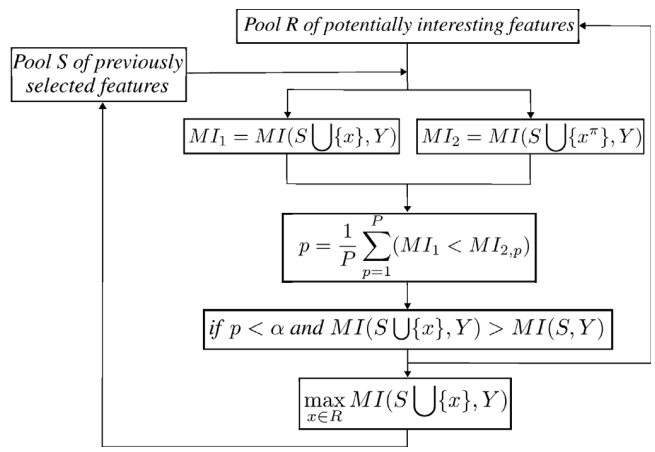

Fig. 1. Flowchart of the MISF algorithm.

3. EXPERIMENTS

We have tested different algorithms of selection on real and simulated data: **Knn** (MI algorithm); **Anova-ltd** (Anova with a number of selected features equal to the number of features selected by the MI algorithm); **Anova** (Anova with a number of selected features equal to 1/5 of the set of features, or fixed by a threshold on the p-value).

3.1. Simulated Data

We have tested the selection on simulated data inspired by Friedmann (see [9]) :

$$Y = 10\sin X_1.X_2 + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon$$

where ϵ is a gaussian noise with an unitary variance, X is 100-dimensional , and we discretize Y into 4 labels : $y \in \{1,2,3,4\}$. The dataset is split into two subsets: the training set (3/4 of the data) on which we perform the selection process and train the classifier, and the test set, on which we test the performance in generalization of the classifier. The

chance level is 25% for 4 labels, and the tests are performed 10 times in order to have a good assessment of the performances. We take $\alpha = 0.05$ (then $P = 400$) and $k = 20$.

3.2. Results on Simulated Data

The results on simulated data show a slight improvement for the performance in generalization, by using the MIFS algorithm (see Fig. 2).

	Knn	Anova-ltd	Anova
	SVM		
Mean	54%	53%	46%
Std	11%	10%	13%
	RVM		
Mean	53%	52%	44%
Std	11%	12%	14%
Size Selection	2.5	2.5	20

Fig. 2. Comparison (mean and standard deviation) of the different methods of reduction of dimensions for simulated data, for $\alpha = 0.05$, $P = 400$ and $k = 20$. The MI feature selection performs sligthy better than the other techniques.

3.3. Real Data

We have applied the algorithm MI on a real fMRI data set. This dataset comprises images acquired while the subjects were viewing some chairs of different sizes and shapes (see [10] for more details on the data). We split our data set into two parts, a training set (3/4 of the data) and a test set(1/4 of the data), and we applied a features selection and set up a classifier, to check if we could retrieve which size of object is seen, whatever the shape. This problem is an intra-subject problem: we work only on the data of one subject, and we make an average of the results across 12 subjects.

3.4. Results on Real Data

Effect of α: The parameter α has a strong influence on the outcome of the selection. We have studied the size of the selection for different values of α (see Fig. 3), on the set of images of subject 6, where we have pre-selected the 300 voxels with the highest F-score in Anova, in order to reduce the computation time. We can see that the final number of features depends on α. When α increases, the selection is less strict, and the number of selected features is higher. The performance in generalization of the selection of features is constant, which seems to imply that the first set of 4 voxels contains all the information needed to classify the images. However, the voxels added by increasing α do not seem to decrease the performance in generalization for the SVM. It is interesting to notice that for very low threshold (i.e. for a low α), our method is more efficient than the reference method, but is more time consuming. In the following parts of the study, we will keep a medium threshold ($\alpha = 0.05$) to allow an easier computation of the results.

α	10^{-3}	5.10^{-3}	10^{-2}	5.10^{-2}	0.1
P	10^4	10^4	10^3	5.10^2	10^2
Size MI	4	6	6	7	8
Size Anova-Pval	75	122	148	256	300
	RVM				
Knn (%)	78	72	72	61	61
Anova-Pval (%)	72	78	72	83	67
Anova-60 (%)	67				
	SVM				
Knn (%)	**89**	**94**	**94**	**89**	**89**
Anova-Pval (%)	78	83	83	83	78
Anova-60 (%)	83				

Fig. 3. Size of the selection and performances in generalization for different values of α and for the reference method (ANOVA-60, with 60 voxels selected, and Anova-Pval with the threshold for the p-value equal to alpha), on the data of subject 12. The MI selection gives better results for low values of α, and the SVM is globally more efficient than the RVM.

Comparison of different methods: We have studied the different methods with voxels and parcels (Fig. 4). The results were averaged over 5 trials and 12 subjects, and we use $k = 20$, $\alpha = 0.05$ and $P = 400$. In Fig. 4, we can see that when using RVM, the MI algorithm gives equivalent results than the reference method (for the same number of voxels than selected by the MI algorithm, ANOVA is less efficient). In the bottom part of Fig. 4, we can notice that, when using SVM classification, the reference method performs better than the MI algorithm, and SVM performs better than RVM for classification. However, those results are to be related to the number of selected voxels : for both SVM and RVM, with an equal number of selected voxels, given by the MI algorithm, our method is more efficient than the ANOVA. The parcellation is less accurate than the use of voxels, but we selected very few parcels (about 2 or 3). However, there is a very high variability in performances of the methods between subjects.

4. DISCUSSION

4.1. Classification and Mutual Information

In this paper, we have compared different methods for classification and to elaborate a method of features selection by using MI criterion. It seems that MI is an interesting way to find areas of activity in the brain by selecting very few, but strongly informative, voxels (from 1500 to 8 voxels). In our experiments, RVM classifiers are less efficient than SVM.

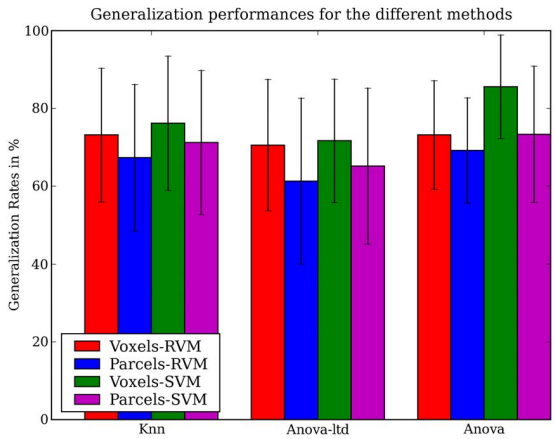

Fig. 4. Comparison of the different methods of selection for all the subject and 5 resamplings per subject. The chance level is 33%. The different approachs give similar results, but the reference method, associated with SVM slightly outperforms the others.

4.2. Neuroscientific aspects

The MI algorithm seems to be an interesting alternative to the classical mass-univariate method currently used in brain-reading. The small number of selected voxels allows an easy interpretation of the results : the selected areas are reduced to very few voxels, and seem to be the areas that are the most strongly related to the cognitive tasks. We can see (fig. 5) that MIFS selects only few voxels at the core of activated areas (by removing the redundant voxels). The most predictive regions for object size seems to be the occipital part of the LOC. By trying to predict the shape of the chairs viewed by the subjects, we obtain a classification rate of 75% (85% for sizes), but the informations to be extracted are more complex. These results emphazise one of the aims of the MIFS algorithm which is to find relatively precise regions, in order to better understand the discriminative patterns of functional activity.

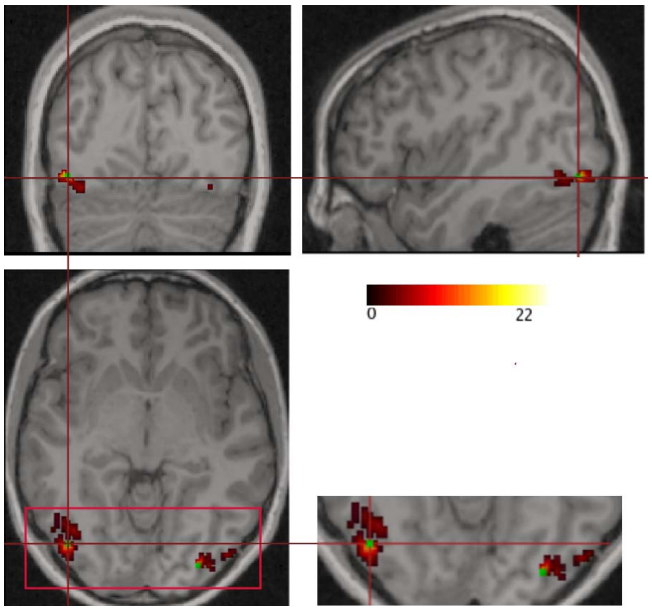

Fig. 5. Representation of the voxels selected by the MI algorithm (green) and the Anova method (values of the F-score between black and yellow), in the case of discrimination between the 3 different sizes, for the subject 6, with $\alpha = 0.05$ and $k = 20$. The main activity found by MI is at the localization (42,-75,-5) mm, in the lateral occipital cortex. The localization of the activity is far more defined with the MI selection than the reference method, and allows a better interpretation of the results.

5. REFERENCES

[1] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fmri) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex," vol. 19, no. 2, pp. 261–270, June 2003.

[2] John-Dylan Haynes and Geraint Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, pp. 523–534, 2006.

[3] Alexander Kraskov, Harald Stoegbauer, and Peter Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, 2004.

[4] F. Rossi, A. Lendasse, D. François, V. Wertz, and M. Verleysen, "Mutual information for the selection of relevant variables in spectrometric nonlinear modelling," *Chemometrics and intelligent laboratory systems*, vol. 80, pp. 215–226, 2006.

[5] A. Luo and P. Sajda, "Using single-trial eeg to estimate the timing of target onset during rapid serial visual presentation," in *Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS '06*, Aug. 2006, pp. 79–82.

[6] D. François, F. Rossi, V. Wertz, and M. Verleysen, "Resampling methods for parameter-free and robust feature selection with mutual information," *Neurocomput.*, vol. 70, no. 7-9, pp. 1276–1288, 2007.

[7] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] M. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems, San Mateo, CA*. 2000, Morgan Kaufmann.

[9] J. Friedman, "Multivariate adaptive regression splines (with discus- sion)," *Ann. Stat.*, vol. 9, 1991.

[10] Evelyn Eger, John Ashburner, John-Dylan Haynes, Raymond J. Dolan, , and Geraint Rees, "Functional magnetic resonance imaging activity patterns in human lateral occipital complex carry information about object exemplars within category.," *Journal of Cognitive Neuroscience*, vol. 20:2, 2007.