SIMPLE FULLY AUTOMATED GROUP CLASSIFICATION ON BRAIN FMRI

Jean Honorio[1,2], *Dimitris Samaras*[1], *Dardo Tomasi*[2,3], *Rita Goldstein*[2]

[1] Computer Science Dept., Stony Brook University [2] Medical Dept., Brookhaven National Laboratory
[3] National Institute on Alcohol Abuse and Alcoholism

ABSTRACT

We propose a simple, well grounded classification technique which is suited for group classification on brain fMRI datasets that have high dimensionality, small number of subjects, high noise level, high subject variability, imperfect registration and capture subtle cognitive effects. We propose *threshold-split region* as a new feature selection method and majority vote as the classification technique. Our method does not require a predefined set of regions of interest. We use average across sessions, only one feature per experimental condition, feature independence assumption, and simple classifiers. The seeming counter-intuitive approach of using a simple design is supported by signal processing and statistical theory. Experimental results in two block design datasets that capture brain function under distinct monetary rewards for cocaine addicted and control subjects, show that our method exhibits increased generalization accuracy compared to commonly used feature selection and classification techniques.

Index Terms— Pattern classification, magnetic resonance imaging

1. INTRODUCTION

Despite the tremendous progress in brain studies, still little is known about brain function for specific activities. One way to understand this process is through statistical analysis of brain imaging datasets. Neural activity can be captured by functional magnetic resonance imaging (fMRI) by taking advantage of the hemodynamics response.

In this paper, we propose a method for group classification on brain fMRI data. We apply our method to find functional differences between cocaine addicted versus healthy non-drug-using control subjects. In this two-class classification problem, we need to address two main questions: (i) what are the appropriate features and (ii) which classifier or ensemble of classifiers are good for discrimination.

In order to devise appropriate techniques for classification from brain fMRI, the main issues that must be taken into account when dealing with this modality are: (i) the datasets are very high dimensional, with tens of thousands of voxels per subject (ii) the number of available subjects (sample size) is small due to the cost and time needed to capture information; in practice, most datasets have only a few tens of subjects (iii) high noise level due to the high magnetic field in the acquisition (iv) high subject variability, such that intra-class differences could potentially hide inter-class differences and (v) imperfect spatial registration between subjects, due to non-affine deformations in the fMRI data with respect to the true anatomy which are caused by the high magnetic field and fast acquisition (i.e. the functional versus the anatomical MRI of the same subject show important deformations that are not due to scale, translation, rotation).

Several group classification techniques have been proposed for Schizophrenia [1, 2], Alzheimer and mild traumatic brain injury [2] and depression [3]. Prediction of cognitive states has been applied for lie detection [4], for prediction of size and shape of an observed object (e.g. chair) [5] or for detecting if a person is examining a sentence or a picture [6, 7].

Some methods require a predefined set of regions of interest (ROIs) [1, 7]. The main drawback for such methods is that in the absence of prior knowledge they are of little help. Even if the researcher has remarkable knowledge regarding the underlying brain process, such prior could significatively skew the results. In practice, people might perform “double dipping” [8] in the dataset in order to find the set of ROIs for classification, and therefore destroy the significance of the cross-validation results. Given those disadvantages, fully exploratory methods are preferred. Our method does not require a predefined set of ROIs.

Section 2 introduces the design principles behind our method. Section 3 presents our technique. Experimental results are shown and discussed in Section 4. Main contributions and results are summarized in Section 5.

2. DESIGN PRINCIPLES

We present the principles from signal processing and statistical theory that guide the design of a simple, well grounded classification technique.

Our datasets were collected at 4Tesla which allows capturing subtle cognitive effects, e.g. monetary reward processing. Such high magnetic field introduces several signal artifacts. We use the average of contrast maps across different sessions in order to reduce noise and avoid using single session contrasts or difference of contrasts.

In order to alleviate the effects of imperfect spatial registration of brains between subjects, we use the average activity in a brain region instead of independent voxel activity.

We use a very small number of features in order to avoid the curse of dimensionality problem, since the number of samples is small. To illustrate the possible pitfalls of having more features than samples, consider a dataset with S subjects and $V \gg S$ voxels. We can train a linear support vector machine with several training sets of $S-1$ subjects (in a leave-one-out fashion) and S voxels picked at random. Every training set will obtain 0% training error, which shows over-fitting, and therefore the generalization error will be extremely high.

We observed that if feature selection is unstable under cross-validation, i.e. different regions are picked for different training sets, then generalization accuracy drops. We use stable features under cross-validation in order to ensure good performance.

Several authors have noticed that for high dimensional datasets with small number of samples, assuming full independence often performs better than learning dependencies among features, even for domains where the full independence assumption may not be valid [9, 10]. We assume independence of features, since the number of samples is insufficient for reliably learning dependencies among features.

Due to the very small number of samples, we use very simple classifiers in order to reduce the number of parameters to be learnt. We also avoid the assumption of underlying probabilistic distributions or the use second-order statistics as parameters (e.g. variance), since those also require having larger number of samples to obtain reliable results.

3. METHODS

The above principles lead us to propose the following feature selection and classification techniques.

Our feature selection method, *threshold-split region*, picks the biggest region (on the training set) that activates for one class and deactivates for the other class. This feature selection method is very simple, but it leads to regions that allow good classification and are very stable under cross-validation. We also experimentally observed that picking two regions is unstable under cross-validation and leads to poor generalization accuracies.

We propose using decision stump classifiers since we are only interested in activations and deactivations, i.e. voxel values being higher or lower than a specified threshold. Let x be the value of a feature, e.g. the activation for one voxel. A decision stump, formally defined as in eq.(1), classifies its input x by comparing it with a threshold θ and a polarity $p \in \{-1, +1\}$. We learn the parameters p and θ by minimizing the classification error in the training set.

$$h_{p,\theta}(x) = \begin{cases} \text{“Cocaine”}, \text{ if } px < p\theta \\ \text{“Control”}, \text{ otherwise} \end{cases} \tag{1}$$

For each experimental condition (e.g. 45¢, 1¢, 0¢), we rank each voxel independently according to their training error. We keep only voxels in the top 99.5% percentile and compute spatial clusters by using 18-connectivity. We take the mean activation from the voxels in the biggest cluster as the single feature that is used for each condition.

Our classifier is a decision stump on the single feature that was previously selected on the training set. When several conditions are used, we perform majority vote on the decision stump classifiers. When there is a tie, the classifier outputs the “Control” class.

4. EXPERIMENTS

4.1. Datasets

We apply our method on two fMRI datasets that capture the difference in brain function under distinct monetary rewards for cocaine addicted as well as healthy control subjects.

First Money Task. The overall neuropsychological experiment follows a block design that includes six sessions, each consisting of three monetary reward conditions (45¢, 1¢, 0¢). The dataset contains 16 cocaine addicted individuals and 12 control subjects [11]. Only sessions complying to the following requirements were used: motion <2mm translation, <2° rotation. The previous requirements reduce the effect of motion confounders in the contrast maps. At least four out of the six sessions were used per subject.

Second Money Task. The overall neuropsychological experiment follows a block design that includes six sessions, each of them under different conditions, i.e. one of three monetary reward conditions (50¢, 25¢, 0¢) and one of two cues (drug words, neutral words). In this paper, we focus on monetary conditions only. The dataset contains 16 cocaine addicted individuals and 17 control subjects [12]. Only subjects with all six sessions complying to the following requirements were used: motion <2mm translation, <2° rotation, and at least 50% performance of the subject in an unrelated task (see [12] for details). The previous requirements reduce noise by having the maximum number of sessions to average for all subjects. They also minimize the effect of confounders in the contrast maps, i.e. motion or poor performance.

Contrast maps were computed by using the statistical parametric mapping package SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). We applied grand mean scaling, since scale between different subjects and experimental conditions can be significantly different.

4.2. Generalization Accuracy

When performing cross-validation in order to approximate the generalization accuracy, the common practice is to use leave-one-out, as it is evident in all referenced papers. Therefore, we chose that cross-validation method for our experiments.

Table 1 shows the generalization accuracy of our method for different sets of experimental conditions on the first

Conditions(s)	45¢	1¢	0¢	45¢,1¢	1¢,0¢	***45¢,1¢,0¢***
Accuracy (chance=57.1%)	82.1%	82.1%	82.1%	85.7%	85.7%	***89.3%***

Table 1. Leave-one-out classification accuracy of our method on the first money task. Note that better accuracy is obtained by mixing different conditions in the classifier.

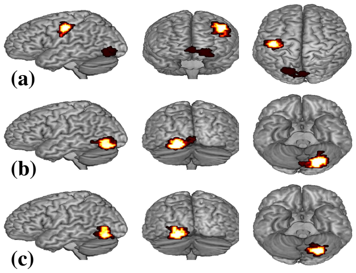

Fig. 1. Most selected regions by our method under leave-one-out for the first money task for different conditions: (a) 45¢: center (42,-15,44), 100 voxels, Brodmann areas 3,4,6, frequency 92.9% (b) 1¢: center (22,-76,-13), 147 voxels, Brodmann areas 18,19 and (c) 0¢: center (22,-72,-11), 114 voxels, Brodmann areas 18,19. Images were generated on the MRIcroN package (`http://www.mricro.com/`).

money task. Note that better accuracy is obtained by mixing different conditions in the classifier. Figure 1 shows the brain regions associated with each condition. Brodmann areas 3,4,6 (sensorimotor cortex) are selected for the 45¢ condition. We hypothesize that those areas are affected due to the fact that cocaine is a stimulant. Broadmann areas 18,19 (visual association cortex) are selected for the 1¢ and 0¢ conditions. We hypothesize that since cocaine addicted subjects are withdrawn from cocaine for the experiments, they possibly undergo vivid visual experiences.

Table 2 shows the generalization accuracy of our method for different sets of experimental conditions on the second money task. Note that better accuracy is obtained by mixing different conditions in the classifier. Figure 2 shows the brain regions associated with each condition. Brodmann areas 24,32 (anterior cingulate cortex) are selected for the 50¢ condition. The cerebellar tonsil is selected for the 25¢ and 0¢ conditions. It is very interesting to observe that, in both datasets, prefrontal cortical regions (Brodmann areas 3,4,6,24,32) are associated with the high monetary conditions, while the posterior regions (Brodmann areas 18,19 and cerebellum) are implicated in the lower monetary conditions. We hypothesize that only high monetary reward elicits such a prefrontal cortex response, possibly due to more effort or anticipation.

4.3. Comparison to Other Techniques

We compare our proposed technique to several feature selection and classification methods, common in the literature. The feature selection methods in our evaluation include: threshold-split region (our proposed method), principal component analysis (PCA) [2, 3], independent component analysis (ICA) [1], average value on a coarse image resolution by using non-overlapping $16\times16\times16$mm^3 cubes of voxels [4], most discriminative voxels [6] by ranking them independently with Gaussian classifiers, most active voxels [7] by ranking them independently with a two sample T-statistic for the difference of means, unequal sample sizes and unequal variances; searchlight accuracy [13] by using a Gaussian Naïve Bayes classifier on the $3\times3\times3$ voxel neighborhood as feature set.

Conditions(s)	50¢	25¢	0¢	50¢,25¢	***50¢,25¢,0¢***
Accuracy (chance=51.5%)	78.8%	66.7%	72.7%	81.8%	***90.9%***

Table 2. Leave-one-out classification accuracy of our method on the second money task. Note that better accuracy is obtained by mixing different conditions in the classifier.

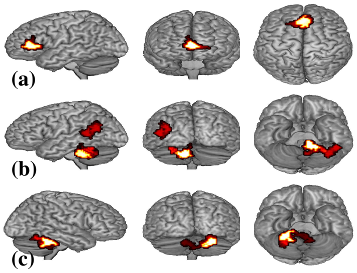

Fig. 2. Most selected regions by our method under leave-one-out for the second money task for different conditions: (a) 50¢: center (0,35,5), 116 voxels, Brodmann areas 24,32 (b) 25¢: center (13,-42,-34), 93 voxels, cerebellar tonsil, frequency 75.8% and (c) 0¢: center (-23,-43,-30), 148 voxels, cerebellar tonsil, culmen. Images were generated on the MRIcroN package (`http://www.mricro.com/`).

The classification methods in our evaluation include: majority vote (MV) on decision stump classifiers (our proposed method), Gaussian naïve Bayes (GNB) [6], k-nearest neighbors (kNN) [7] with number of neighbors k selected by nested cross-validation from $\{1,2,5,10,20\}$; Fisher linear discriminant (FLD) [1, 2], logistic regression (LR), linear support vector machines (LSVM) [3, 5, 6, 7], Gaussian support vector machines (GSVM) [4] with kernel size γ selected by nested cross-validation from $\{1,10,100,1000,10000\}$, and Adaboost (AB) on decision stump classifiers with number of iterations selected by nested cross-validation from $\{5,10,20,50,100\}$.

We report the generalization accuracy for the optimal set of experimental conditions and the ranking for each method in all possible sets of conditions for the first and second money tasks in Table 3 (e.g. for the first money task: {45¢},{1¢},{0¢},{45¢,1¢},{45¢,0¢},{1¢,0¢},{45¢,1¢,0¢}). The ranking uses a normalized scale from 1 to 10, where 10 means that the method outperforms all others for every possible set of conditions, and 1 means that the method loses against all others. We observe that: (i) the only combination of feature and classifier that obtains very good classication accuracy in both datasets, is our proposed method (around 90%); furthermore, it almost always outperforms all other

Accuracy on First Money Task (chance=57.1%)								
Feature	**Classifier**							
	MV	**GNB**	**kNN**	**FLD**	**LR**	**LSVM**	**GSVM**	**AB**
Threshold	***89.3%***	82.1%	82.1%	78.6%	82.1%	71.4%	75.0%	***85.7%***
split	***10.0***	8.9	9.1	8.3	9.1	5.1	8.0	***9.6***
PCA	64.3%	82.1%	78.6%	57.1%	64.3%	71.4%	64.3%	82.1%
(3 comps)	3.0	8.4	6.1	2.3	3.0	6.6	4.0	6.4
ICA	71.4%	82.1%	78.6%	57.1%	60.7%	71.4%	64.3%	71.4%
(3 comps)	6.9	7.9	6.1	2.3	3.0	5.9	4.0	5.1
Cubes	57.1%	71.4%	67.9%	60.7%	57.1%	64.3%	53.6%	***85.7%***
	4.1	6.1	3.6	2.3	2.0	4.0	2.3	***8.7***
Discrim.	71.4%	78.6%	71.4%	85.7%	82.1%	82.1%	***85.7%***	71.4%
(100 voxels)	6.9	8.0	6.3	4.4	4.0	9.3	***8.6***	3.9
Active	75.0%	75.0%	82.1%	60.7%	64.3%	***85.7%***	82.1%	78.6%
(100 voxels)	8.0	6.7	7.6	3.9	3.6	***9.0***	8.3	4.7
Searchlight	71.4%	75.0%	71.4%	67.9%	60.7%	75.0%	78.6%	67.9%
(100 voxels)	5.1	7.4	5.7	3.6	2.3	6.6	9.3	5.1

Accuracy on Second Money Task (chance=51.5%)								
Feature	**Classifier**							
	MV	**GNB**	**kNN**	**FLD**	**LR**	**LSVM**	**GSVM**	**AB**
Threshold	***90.9%***	***81.8%***	***81.8%***	***81.8%***	72.7%	66.7%	75.8%	***78.8%***
split	***9.9***	***9.9***	***9.3***	***9.7***	9.0	6.4	7.1	***9.6***
PCA	48.5%	57.6%	60.6%	48.5%	57.6%	54.5%	63.6%	63.6%
(3 comps)	2.3	4.0	4.7	3.4	3.4	5.0	3.1	4.7
ICA	63.6%	48.5%	60.6%	48.5%	63.6%	78.8%	63.6%	66.7%
(3 comps)	6.7	2.7	4.7	3.4	3.6	5.4	3.1	6.4
Cubes	36.4%	66.7%	63.6%	63.6%	60.6%	63.6%	54.5%	48.5%
	1.3	7.0	4.9	6.3	6.6	6.1	2.7	2.7
Discrim.	75.8%	69.7%	69.7%	72.7%	63.6%	63.6%	66.7%	72.7%
(100 voxels)	8.0	7.1	8.0	6.3	6.9	7.0	6.1	7.9
Active	72.7%	72.7%	69.7%	72.7%	69.7%	69.7%	69.7%	69.7%
(100 voxels)	6.7	5.3	4.1	5.4	6.7	6.6	6.0	6.3
Searchlight	33.3%	63.6%	63.6%	66.7%	75.8%	72.7%	69.7%	75.8%
(100 voxels)	1.1	5.1	5.0	6.1	7.7	6.7	7.0	7.1

Table 3. Leave-one-out classification accuracy for different feature selection and classification methods on the first (top) and second money task (bottom) for the optimal set of conditions (first line) and ranking for each method in all possible sets of conditions (second line, from 1 to 10, higher is better). Methods with an accuracy in the top 90% quantile and a ranking $\geq$ 8.5 are highlighted.

methods for all possible sets of conditions (ranking almost 10) (ii) some combinations of features and classifiers obtain good results on the first money task but not as good results on the second money task, e.g. linear SVM on most active voxels (iii) some combinations of features and classifiers appear to achieve good accuracy for the optimal set of conditions but otherwise perform rather poorly on other sets of conditions (low ranking), e.g. FLD on most discriminative voxels for the first money task.

5. CONCLUSIONS

We have shown that the use of principles from signal processing and statistical theory allowed for the design of very simple, fully automated, and successful methods for feature selection and classification. This led to a model with very low complexity (six parameters, i.e. polarity and threshold of one brain region per each of three experimental conditions) and good generalization accuracy.

Our method has shown approximately 90% of generalization accuracy in two completely different datasets: different subjects, different tasks and different acquisition protocols, i.e. interscan interval (TR=3500ms for the first money task, TR=1600ms for the second money task). Furthermore, it almost always outperforms all other methods for all possible sets of experimental conditions. Since only one region per condition is used, our model is also easy to interpret from a neuropsychological point on view.

There are several ways of extending this research. It would be very interesting to apply our method to other group classification problems (e.g. canabis addicted, Schizophrenia or Alzheimer versus control patients), or to the prediction of cognitive states. Another very interesting line of research is to measure the generalization accuracy of our method for different magnetic field and noise levels, larger number of samples, or for event-related tasks.

Acknowledgments. This work was supported in part by NIDA Grants 1 R01 DA020949, R 21 DA02062, 1 R01 DA023579, NIAAA Grant 2 R01 AA09481 and NIBIB Grant 1 R01 EB007530

6. REFERENCES

[1] O. Demirci, V. Clark, and V. Calhoun, "A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia," *NeuroImage*, 2008.

[2] J. Ford, H. Farid, F. Makedon, L. Flashman, T. Mc Allister, V. Megalooikonomou, and A. Saykin, "Patient classification of fMRI activation maps," *MICCAI*, 2003.

[3] C. Fu, J. Mourão-Miranda, S. Costafreda, A. Khanna, A. Marquand, S. Williams, and M. Brammer, "Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression," *Biological Psychiatry*, 2008.

[4] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. Langlebenb, "Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection," *NeuroImage*, 2005.

[5] V. Michel, C. Damon, and B. Thirion, "Mutual information-based feature selection enhances fMRI brain activity classification," *ISBI*, 2008.

[6] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning*, 2004.

[7] X. Wang, R. Hutchinson, and T. Mitchell, "Training fMRI classifiers to discriminate cognitive states across multiple subjects," *NIPS*, 2003.

[8] N. Kriegeskorte, W. Simmons, P. Bellgowan, and C. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nature Neuroscience*, 2009.

[9] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, 1997.

[10] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, 1997.

[11] R. Goldstein, D. Tomasi, N. Alia-Klein, L. Zhang, F. Telang, and N. Volkow, "The effect of practice on a sustained attention task in cocaine abusers," *NeuroImage*, 2007.

[12] R. Goldstein, N. Alia-Klein, D. Tomasi, J. Honorio, T. Maloney, P. Woicik, R. Wang, F. Telang, and N. Volkow, "Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction," *PNAS*, 2009.

[13] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, 2009.