Prediction of Successful Memory Encoding from fMRI Data

S.K. Balci[1], M.R. Sabuncu[1], J. Yoo[2], S.S. Ghosh[3], S. Whitfield-Gabrieli[2], J.D.E. Gabrieli[2] and P. Golland[1]

[1] CSAIL, MIT, Cambridge, MA, USA
[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[3] RLE, MIT, Cambridge, MA, USA

Abstract. In this work, we explore the use of classification algorithms in predicting mental states from functional neuroimaging data. We train a linear support vector machine classifier to characterize spatial fMRI activation patterns. We employ a general linear model based feature extraction method and use the t-test for feature selection. We evaluate our method on a memory encoding task, using participants' subjective prediction about learning as a benchmark for our classifier. We show that the classifier achieves better than random predictions and the average accuracy is close to subject's own prediction performance. In addition, we validate our tool on a simple motor task where we demonstrate an average prediction accuracy of over 90%. Our experiments demonstrate that the classifier performance depends significantly on the complexity of the experimental design and the mental process of interest.

1 Introduction

An important component of human learning is to evaluate whether information has been successfully committed to memory. Humans with superior judgments of learning are shown to perform better in learning tasks [1]. Recent functional neuroimaging studies have identified brain regions correlated with actual and predicted memory encoding using univariate analysis techniques [2]. In this work, we adopt the discriminative approach to predicting successful encoding. We view this work as a first step toward the development of tools that will enhance human learning. One of the possible applications is human-machine interfaces which employ a feedback mechanism to ensure successful acquisition of skills in critical applications.

Univariate techniques, such as the general linear model (GLM), are traditionally used to identify neural correlates in fMRI data [3]. In contrast, multivariate discriminative methods train a classifier to predict the cognitive state of a subject from the spatial brain activation pattern at that moment [4–6]. Most studies use linear classifiers [7–17], while others employ nonlinear classifiers [16–19].

Functional MRI classification is challenging due to the high dimensionality of the data, noisy measurements, motion artifacts and the small number of available training examples. Feature selection and dimensionality reduction techniques

promise to alleviate some of these problems. One approach is to restrict the analysis to anatomical regions of interest [7, 18]. Another is to compute univariate statistics to rank the features according to their discriminative power between the conditions of interest [7, 8, 11, 12]. Multivariate feature selection methods can evaluate the information content of subsets of features. However, such methods have to work in a large search space of all possible combinations of features. This problem is addressed by constraining the search space to local neighborhoods [20] or by adding one feature at a time to the feature set [18].

Pattern classification methods have been successfully applied to fMRI experiments on visual [7, 9, 10, 15, 16], motor [14], cognitive [11, 13] tasks, and experiments where subject's cognitive state cannot be inferred from simple inspection of the stimulus, such as memory retrieval [8].

The performance of the classifier depends on the complexity of the experimental paradigm [21]. O'Toole *et al.* [15] show that the classifier's ability to discriminate between different object categories decreases as the visual similarity of the objects increases. In our experiments, we observe that the classifier performance depends greatly on the complexity of the cognitive task of interest. While we achieve high accuracy in a simple motor task, classification accuracy is lower in a high level memory encoding task.

In this work, we explore the use of classification methods in the context of an event related functional neuroimaging experiment where participants viewed images of scenes and predicted whether they would remember each scene in a post-scan recognition-memory test. We trained support vector machines on functional data to predict participants' performance in the recognition test and compared the classifier's performance with participants' subjective predictions. We show that the classifier achieves better than random predictions and the average accuracy is close to that of the subject's own prediction.

2 Methods

Here we describe all the computational steps of the analysis, including feature extraction, feature selection and classification. We choose to use a GLM-based feature extraction method, which increases the classification accuracy by extracting the signal related to experimental conditions. We employ a feature selection method based on univariate statistics to decrease the dimensionality of the data. We then train a linear support vector machine and evaluate its accuracy on functional neuroimaging data using a set of cross-validation procedures.

2.1 Feature Extraction

Let $y(v)$ be the fMRI signal of N time points measured at a spatial location v, X be the matrix of regressors, $\boldsymbol{\beta}(v)$ be the coefficients for regressors in the columns of X, and M be the total number of stimulus onsets. The general linear model [3] explains $y(v)$ in terms of a linear combination of regression variables $\beta(v)$:

$$y(v) = X\boldsymbol{\beta}(v) + e(v), \tag{1}$$

where $e(v)$ is modeled as i.i.d. white Gaussian noise. Each of the first M columns of X is obtained by convolving the hemodynamic response function with a reference vector which indicates the onset of a particular stimulus. The remaining columns of X include nuisance regressors that include motion correction and detrending parameters. The maximum likelihood estimate $\hat{\beta}(v) = (X^T X)^{-1} X^T y(v)$ also corresponds to the least-squares solution. We obtain a GLM-beta map by combining m'th elements of $\hat{\beta}(v)$ over all spatial locations v into a vector $\hat{\beta}_m$ which represents the spatial distribution of activations for the m'th stimulus. $\hat{\beta}_m$ contains V elements, one for each voxel in the original fMRI scan.

2.2 Feature Selection

Let $L = \{l_1, ..., l_M\}$ be a vector denoting the class label of each stimulus, $l_i \in \{+1, -1\}$. The t-statistic $t(v)$ for voxel v,

$$t(v) = \frac{\mu_{+1}(v) - \mu_{-1}(v)}{\sqrt{\frac{\sigma^2_{+1}(v)}{n_{+1}} + \frac{\sigma^2_{-1}(v)}{n_{-1}}}}, \tag{2}$$

is a function of $n_l(v)$, $\mu_l(v)$ and $\sigma^2_l(v)$, $l = -1, +1$. $n_l(v)$ is the number of stimuli with label l. $\mu_l(v)$ and $\sigma^2_l(v)$ are, respectively, the mean and the variance of the components of $\hat{\beta}(v)$ corresponding to stimuli with label l. A threshold is applied to the t-statistic to obtain a subset of coefficients that we denote $\bar{\beta}$.

2.3 Weighted SVM

Since we work with unbalanced data sets, we choose to use the weighted SVM variant, which imposes different penalties for misclassification of samples in different groups [22, 23]. Given the penalty for positive class C_+, and the penalty for the negative class C_-, the weighted SVM with a linear decision boundary solves the following constrained optimization problem:

$$\begin{aligned} \langle w^*, b^*, \xi^* \rangle = \underset{w,b,\xi}{\operatorname{argmin}} \left\{ \frac{1}{2} w^T w + C_+ \sum_{l_m=1} \xi_m + C_- \sum_{l_m=-1} \xi_m \right\} \\ \text{s.t. } \; l_m(w^T \bar{\beta}_m + b) \geq 1 - \xi_m \text{ and } \xi_m \geq 0 \text{ for } \; m = 1, \ldots, M. \end{aligned} \tag{3}$$

The resulting classifier predicts the hidden label of a new GLM-beta map $\bar{\beta}$ based on the sign of $w^{*T} \bar{\beta} + b^*$.

2.4 Experimental Evaluation

To evaluate the performance of this training scheme over a range of penalties C_+ and C_-, we construct the ROC curves. In all experiments in this paper,

each subject participated in several runs of the experiment. We employ a cross-validation procedure by holding out one of the functional runs, training the classifier on the remaining runs and testing it on the hold-out run. In the feature selection step, we evaluate a range of threshold values and choose the threshold value corresponding to maximum cross-validation accuracy within the training set. We obtain the ROC curves by training the SVM classifier using varying weights for the class penalties C_+ and C_- in equation (3), and averaging the testing accuracy across runs. The values of C_+ and C_- are equally spaced on a log scale where the ratio of penalties vary between 10^{-5} and 10^5. In addition, we identify the point on the ROC curve that corresponds to the smallest probability of error. We report the classification accuracy of that point which we call min-error classification accuracy.

In the motor task experiments, we demonstrate the benefit of feature selection by comparing our method to an SVM classifier trained on all features. For memory encoding experiments, we have two labels for each stimulus available to us: the actual memory encoding and the subject's prediction of the performance. We employ three different training strategies which aim to explore the challenging nature of this experiment. The first strategy corresponds to the standard training setup. We perform feature selection on the training set only, train the classifier on all samples in the training set and evaluate the accuracy on the test set. The second strategy restricts the training set to samples where the subject's prediction is correct. One of the main challenges in our experimental design is to obtain correct labels for the samples as we rely on subject's response for the actual memory encoding. With the second setup we aim to improve reliability of training samples by requiring the predicted and the actual labels to agree. For the third strategy, we perform feature selection using both the training and test sets while still training the classifier on samples in the training set. This setup is impractical as in real applications we do not have access to test data. However, it serves as an indicator of the best accuracy we could hope to achieve.

3 fMRI Experiments and Data

We acquired fMRI scans using a 3T Siemens scanner. We obtained functional images using T2-weighted imaging (repetition time=2s, echo time=30s, $64 \times 64 \times 32$ voxels, 3mm in-plane resolution, 4mm slice thickness). We collected 1,500 MR-images in five functional runs, each run 10 minutes long. We used Statistical Parametric Mapping (SPM5) [3] to perform motion correction using 6-parameter rigid body registration of images to the mean intensity image and smoothing with a Gaussian filter(FWHM=8mm) to decrease the effects of motion artifacts and scanner noise.

In the memory encoding task, we scanned 10 participants with normal visual acuity. We used five hundred pictures of indoor and outdoor scenes and randomly divided them into ten lists of 50 pictures. We presented five lists during the scan and scanned the subjects in five functional runs as they studied 50 pictures in each run. We presented each picture for three seconds with a nine second rest interval and instructed participants to memorize the scenes for a later memory

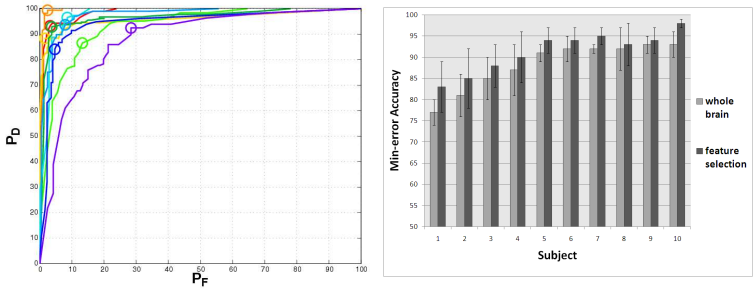


Fig. 1. Left: ROC curves for the motor task for 10 subjects for classification with feature selection. Circles show the operating points corresponding to min-error classification accuracy. Right: Min-error classification accuracy for classification without feature selection (light-gray) and with feature selection (dark-gray).

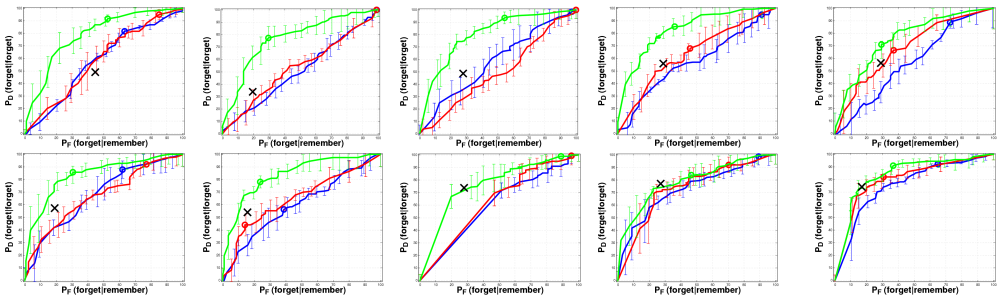
test. For each picture, participants predicted whether they would remember or forget it, by pressing a response button. Following the scan we gave participants a recognition test where we presented them all 500 pictures, including the 250 images they had not seen before. The participants judged whether they had seen the picture during the scan. In our classification experiments, we used participants' responses in the recognition test to derive the binary labels and their predictions during the scan as a benchmark for our classifier.

In the motor task, we scanned another 10 subjects, using the same setup and acquisition parameters as in the memory encoding task with the only difference that the subject's prediction was acquired using two buttons. We instructed subjects to press the left button using their left hand if they thought they would remember the presented picture and press the right button using their right hand otherwise. We use this dataset to train the classifier to predict the hand used to press the button.

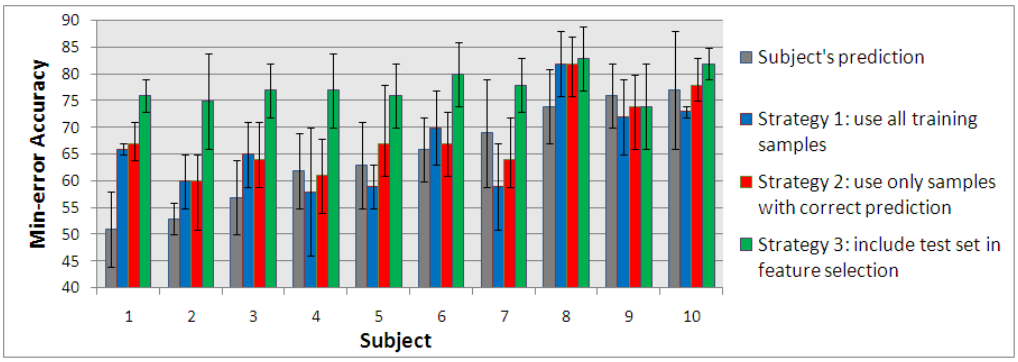
4 Results

We first evaluate the method on the simple motor task and then present the results for the memory encoding experiment. Figure 1 shows the ROC curves and the min-error classification accuracies for the motor task. We observe that in this simple motor task the classifier achieves highly accurate results, the min-error classification accuracy is over 90% for the majority of the subjects. Furthermore, the bar graph shows that feature selection improves classification accuracy compared to using all voxels for classification.

Figure 2(a) shows the results for the memory encoding task for all three strategies for training a classifier described in Sec 2.4. For the first strategy (blue), we note that the ROC curves of the classifier are better than random but are lower than subject's predictions. The ROC curves of the second strategy are shown in red. We note that the curves improve and are closer to subject's own predictions. A statistical comparison between the first and the second strategies reveals a significant difference (single-sided, paired T-test, $P < 0.05$). This observation confirms that the samples whose labels are correctly predicted by the



(a) ROC curves



(b) min-error graph

Fig. 2. (a) ROC curves for memory encoding experiment for 10 subjects. Crosses represent subject's prediction accuracy. Blue curves correspond to strategy 1, using the training set for feature selection. Red curves correspond to training the classifier only on correctly predicted samples (strategy 2). Green curves correspond to strategy 3, including test set in feature selection. Circles show the operating points corresponding to min-error classification accuracy. (b) Min-error classification accuracy.

subject indeed provide more reliable samples for training the classifier. Green curves correspond to the third strategy of performing feature selection on both the training and test sets. As expected, the ROC curves are much higher, even surpassing subject's own predictions. However, we note that even in this impractical setting where we use the test set for feature selection, the ROC curves are far from perfect, indicating the high level of noise present in the observations and the labels.

Figure 2(b) shows the min-error classification accuracy for the memory encoding task. The min-error accuracy of the classifier is very close to, and sometimes better than the subject's own predictions. We note that the highly uneven frequencies of the two labels significantly affect the min-error classification accuracy. In our dataset, the class sizes are unbalanced by a factor of about three-to-one as subjects remember pictures more often than they forget them. As a result, the operating points that correspond to min-error accuracy for the classifier occur at higher false alarm rates than those of subject's predictions. The classifier is more biased toward predicting the "remember" class, which in-

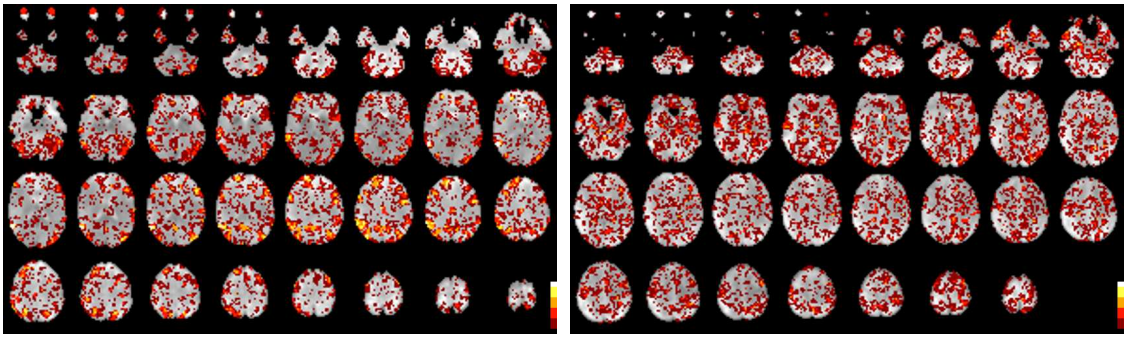


Fig. 3. Feature overlap maps for the best(left) and the worst(right) performing subjects for the memory encoding task. For all five functional runs feature selection is performed on each run. The color indicates the number of runs in which a voxel was selected. Dark red color shows the voxels selected only in one run and white color displays voxels selected in all runs.

creases the min-error accuracy by weighting the high false alarm rate with the relatively low probability of the "forget" class.

5 Discussion

Our experiments demonstrate that the classification accuracy is significantly affected by the complexity of the neuroimaging experiment. While we achieve highly accurate results for the simple motor task, the classification accuracy drops for the memory encoding task. Compared to the motor task, memory encoding task involves more complex neural circuitry. In addition, it is challenging to design an experiment in which the actual encoding labels are obtained without subjective evaluation by the participants.

The feature maps in Figure 3 provide an insight into the performance of the classifier. To create these maps, we performed feature selection on each functional run for each subject and computed how often each voxel was included in the resulting feature maps, essentially quantifying the overlap among features selected for each run. Figure 3 shows these feature overlap maps for the memory encoding task for the subject with the best ROC curves and the subject with the worst ROC curves. We note that most included voxels for the worst subject only appear in one of the runs. Such unreliable features and noisy activation patterns lead to poor generalization performance of the classifier. On the other hand, the map for the best subject includes contiguous regions that are present in most of the runs. We observe a consistent spatial activation pattern across runs that explains the high accuracy of the classier.

The future work should clearly address the problem of obtaining better training labels, perhaps by eliminating the prediction part of the task, and investigating ways to bring better spatial consistency to the features selected for classification.

Acknowledgments: This work was in part supported by the NIH NIBIB NAMIC U54-EB005149, NAC P41-RR13218 and the NSF CAREER 0642971 grant.

References

1. King, J., *et. al.*: Judgements of knowing: the influence of retrieval practice. Am. J. Psychol. **93**(2) (1980) 329–343.
2. Kao, Y., Davix, E., Gabrieli, J.: Neural correlates of actual and predicted memory formation. Nature Neuroscience **8**(12) (2005) 1776 – 1783.
3. Friston, K., *et al.*: Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping **2**(4) (1995) 189–210.
4. O'Toole, A., *et al.*: Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. Journal of Cognitive Neuroscience **19**(11) (2007) 1735–1752.
5. Spiers, H., Maguire, E.: Decoding human brain activity during real-world experiences. Trends in Cognitive Sciences **11**(8) (2007) 356–365.
6. Norman, K., *et al.*: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences **10**(9) (2006) 424–430.
7. Haxby, J., *et al.*: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science **293**(5539) (2001) 2425–2430.
8. Polyn, S., *et al.*: Category-specific cortical activity precedes recall during memory encoding. Science **310**(5756) (2005) 1963–1966.
9. Haynes, J., Rees, G.: Decoding mental states from brain activity in humans. Nature Reviews Neuroscience **7**(7) (2006) 523–534.
10. Kamitani, Y., Tong, F.: Decoding the visual and subjective contents of the human brain. Nat. Neuroscience **8**(5) (2005) 679–685.
11. Mitchell, T., *et al.*: Learning to decode cognitive states from brain images. Machine Learning **57**(1-2) (2004) 145–175.
12. Miranda, J., *et al.*: Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. NeuroImage **28**(4) (2005) 980–995.
13. Hardoon, D., *et al.*: Unsupervised analysis of fMRI data using kernel canonical correlation. Trends in Cognitive Sciences **37**(4) (2007) 1250–1259.
14. Laconte, S., *et al.*: Support vector machines for temporal classification of block design fMRI data. NeuroImage **26**(2) (2005) 317–329.
15. O'Toole, A., *et al.*: Partially distributed representations of objects and faces in ventral temporal cortex. Journal of Cognitive Neuroscience **17** (2005) 580–590.
16. Cox, D., Savoy, R.: fMRI Brain Reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage **19**(2) (2003) 261–270.
17. Davatzikos, C., *et al.*: Classifying spatial patterns of brain activity with machine learning methods. NeuroImage **28**(3) (2005) 663–668.
18. Ramon, M., *et al.*: fMRI pattern classification using neuroanatomically constrained boosting. NeuroImage **31**(3) (2006) 1129–1141.
19. Friston, K., *et al.*: Bayesian decoding of brain images. NeuroImage **39**(1) (2008) 181–205.
20. Kriegeskorte, N., *et al.*: Information-based functional brain mapping. PNAS **103**(10) (2006) 3863–3868.
21. Strother, S., *et al.*: The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. NeuroImage **15**(4) (2002) 747–771.
22. Vapnik, V.: Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley and Sons (1998)
23. Osuna, E., *et. al.*: Support vector machines: training and applications. In: AI Memo 1602, Massachusetts Institute of Technology. (1997)