Available online at www.sciencedirect.com
ScienceDirect
Magnetic Resonance Imaging 26 (2008) 1007–1014
MAGNETIC RESONANCE IMAGING

Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys

Shih-pi Ku*, Arthur Gretton, Jakob Macke, Nikos K. Logothetis
Max Planck Institute for Biological Cybernetics, 72076 Tuebingen, Germany
Received 21 January 2008; accepted 22 February 2008

Abstract

Pattern recognition methods have shown that functional magnetic resonance imaging (fMRI) data can reveal significant information about brain activity. For example, in the debate of how object categories are represented in the brain, multivariate analysis has been used to provide evidence of a distributed encoding scheme [*Science* 293:5539 (2001) 2425–2430]. Many follow-up studies have employed different methods to analyze human fMRI data with varying degrees of success [*Nature reviews* 7:7 (2006) 523–534]. In this study, we compare four popular pattern recognition methods: correlation analysis, support-vector machines (SVM), linear discriminant analysis (LDA) and Gaussian naïve Bayes (GNB), using data collected at high field (7 Tesla) with higher resolution than usual fMRI studies. We investigate prediction performance on single trials and for averages across varying numbers of stimulus presentations. The performance of the various algorithms depends on the nature of the brain activity being categorized: for several tasks, many of the methods work well, whereas for others, no method performs above chance level. An important factor in overall classification performance is careful preprocessing of the data, including dimensionality reduction, voxel selection and outlier elimination.

Keywords: Multivariate analysis; Brain imaging; Functional MRI; Monkey; SVM; Naïve Bayes; Linear discriminant analysis; Correlation analysis; Feature selection; High field; High resolution

1. Introduction

Traditional functional magnetic resonance imaging (fMRI) studies focus on identifying which brain regions are most active while a subject is performing a specific cognitive task or perceiving a specific type of stimulus. Other studies describe the effect of different stimuli on activity or correlate responses across brain regions. These approaches treat each voxel as an independent entity [1] and perform averaging across multiple voxels or even subjects. Conversely, a class of qualitatively different methods, known collectively as multivariate analysis techniques, can be used to recognize distributed activity patterns across multiple voxels [2]. The latter statistical pattern recognition algorithms are powerful because they project the activity of multiple voxels to achieve a discriminative separation of the activity patterns. It has been demonstrated that with an optimized algorithm, one can even classify activity patterns instantaneously [3–5].

In the present study, we address the representation of object categories in the brain. Using traditional analysis, one could define category specific activity in the brain, as found in the fusiform face area (FFA), which responds more to faces than other object categories [6–8]. These results support a domain-specific encoding scheme. In contrast, by correlating the blood oxygenation level dependent (BOLD) responses across multiple voxels, Haxby et al. [9] could identify faces by the activity patterns in whole visual cortex, even with FFA effectively excluded. The latter results support a distributed information coding scheme. The similarity method used in Haxby's study demonstrates how category-related, distributed brain activity patterns can be measured. Other researchers have since applied various multivariate methods in analyzing distributed response patterns in human fMRI data sets: approaches include linear discriminant analysis (LDA) [2,10], support vector machines (SVM) [5] and Gaussian naïve Bayes (GNB)

* Corresponding author.
E-mail address: shihpi@tuebingen.mpg.de (S. Ku).

0730-725X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.mri.2008.02.016

[3,11]. In monkey fMRI, category-related activity has only been investigated using a correlation method with data acquired at low field (3 T) [12].

In the present study we acquired the data at high field (7 T) in a monkey. The high-field magnet provides a higher signal-to-noise ratio [13–15], which enables us to perform high resolution imaging. We applied several pattern recognition techniques and data preprocessing approaches to compare their performance in classifying brain activity patterns. We used four classification procedures — the SVM, GNB, correlation analysis and linear discriminant analysis. SVM and GNB have been demonstrated to be efficient in classifying human fMRI data set in a trial-based regime [3,5] but have not been compared directly.

The present study is intended to answer the following questions: first, which multivariate analysis approach is suitable for high-field, high-resolution fMRI data? Second, how do different pre-processing procedures affect classification performance? Third, how does single trial performance compare with performance averaged over varying numbers of stimulus presentations?

2. Material and methods

2.1. Subjects

Experiments were performed on one healthy monkey (*Macaca mulatta*), weighing 14 kg. All experiments were approved by local authorities (Regierungspräsidium) and were in full compliance with the guidelines of the European Community (EUVD 86/609/EEC) for the care and use of laboratory animals. The scanner, primate setup and hardware for the awake monkey experiments were described in detail in Pfeuffer et al. [16] and Keliris et al. [17] Briefly, the monkey was trained using standard operant conditioning techniques in a mock environment that mimics the scanner. He was trained to tolerate the scanner environment and gradient noise and to remain motionless during the trials. Movements were tracked using custom-made motion sensors. The animal learned to fixate a central fixation spot, whereas his eye movements were tracked using an infrared eye tracker.

For scanning, the monkey was positioned in a custom-designed primate chair that has integrated equipment to immobilize the monkey's head, monitor jaw and body motion, track eye movements, present the stimulus and give the reward. Earplugs and foam covering the ears were used to decrease acoustical noise. Software to monitor motion, track eye movements, present the stimulus and give the reward was custom written, and ran on a real time quantum UNIX system (QNX Software System, Canada) interfaced with Microsoft Windows computers. Visual stimuli were presented binocularly using an SVGA fiber-optic system (AVOTEC, Silent Vision) with a resolution of 800×640 pixels and a frame rate of 60 Hz. The stimuli were five-degree (visual angle) gray-scaled images of faces, houses, fractals and fruit, presented to both eyes. The trials were initiated by the monkey, by not moving and fixating on a central fixation spot for 6 s, after which six different images randomly selected from the database of the same category were presented. Each image lasted for 1 s, and the total stimulation time in each trial was 6 s. The animals were required to fixate within two-degree window during the stimulation. To receive the reward, the monkey had to remain motionless for another eight seconds. Trials were aborted when the animal moved or broke fixation. The exact onset of the trials was triggered by the scanner, and the timing of all events was recorded. Only successful trials were analyzed.

2.2. MRI

All the imaging parameters were described in details elsewhere [18], and we provide only a short summary here. Images were acquired using a vertical 7-T scanner (Bruker BIOSPEC) with a 60-cm diameter bore (Bruker, Ettlingen, Germany), running ParaVision (version 4) software. The scanner was temporarily equipped with a 33-cm inner diameter (after acoustic shielding) actively shielded gradient coil (Siemens AC44) that could attain gradients up to 40 mT/m with rise times of 140 μs. The radiofrequency (RF) coil used was a 16-cm saddle coil that was optimized for the temporal lobes. In addition to a global shim, local shimming was performed using a 25-mm^3 quadrant in the center of the brain, using the Fastmap algorithm [19]. RF power was adjusted manually.

The field of view (FOV) was 12.8×12.8 cm, with a matrix size of 72×64, resulting in a final resolution of 1.5 mm in the read direction (L-R) and 2 mm in the phase direction (A-P). Seventeen slices with a thickness of 2 mm were acquired to cover the entire ventral pathway. Slices were acquired along the temporal lobe (−20 degrees from the Frankfurt zero plane) to reduce susceptibility artifacts.

Two-segment spin-echo echo planar imaging (SE-EPI) sequence was used, BW 159 kHz, resulting in an acquisition window of 14.58 ms. Because the sensitivity to susceptibility gradients depends on the length of the acquisition window, the readout window was minimized by using a segmented EPI and by reducing the FOV and image matrix. TE was 40 ms, and repetition time (TR), 1000 ms.

For anatomical reference images, a high-resolution T_1-weighted 3D Modified Driven Equilibrium Fourier Transform (3D-MDEFT) image of the entire brain was acquired under general anesthesia at 4.7 T (for procedure, see Ref. [20]). For each session, anatomical reference images were acquired using a fast low-angle shot (FLASH) [21] sequence.

2.3. Data analysis

2.3.1. Functional activity and voxel selection

EPI images were reconstructed using Bruker ParaVision software, including ramp sampling. Data were analyzed using custom-written routines in MatLab (The MathWorks, Natick, MA, USA, version 6.5), Statistic Parametric Mapping 2 (SPM2) (Wellcome Department of Cognitive Neurology, London, UK) and FSL (FMRIB Software Library, Analysis group, FMRIB, Oxford, UK). Details of the analysis

procedure are described in Keliris et al. [17]. Briefly, volumes that were contaminated by motion artifacts were removed, based on information from the motion sensors. Images were realigned to correct for motion and B_0-changes. Statistical analysis was done using SPM2 using general linear model (GLM) analysis with the default hemodynamic response function. Using SPM region of interest analysis toolbox (Marsbar 0.38.2), the voxels were selected according to two criteria: anatomy and functional activity. The inferotemporal (IT) cortex was manually segmented on MDEFT (high-resolution anatomy image). Activation was thresholded at six significance levels [F test, $P<10^{-3}$–10^{-7} family-wise error rate (FWE) corrected] in order to include different number of significant responding voxels for further pattern recognition analysis, yielding regions of size 139–550 voxels in the inferotemporal cortex. This procedure is used to measure the effect of dimensionality on classifier performance. The ratio between peak amplitude (8–12 s after stimulus onset) of each trial and its baseline (4–0 seconds before stimulus-onset) was used as a feature for classification. Each category contains about 80 trials.

2.3.2. Pattern recognition

In this section, we describe our data selection and classification procedures. Varying numbers of voxels were extracted from the IT cortex, to measure the resulting gain in classification accuracy. A variety of preprocessing steps were also employed to reduce noise and enhance classifier performance. Four classifiers were used: a linear SVM, LDA, GNB and correlation analysis.

We determine the ability of each classifier to predict the stimulus from both single trials and the averaged activity of several presentations of the stimulus. As this averaging reduces the trial-to-trial variations in the signal, prediction from averaged activity is an easier task for linear algorithms than single trial prediction. On the other hand, averaging increases the dependence between test points in the data, which can result in a loss of statistical power. To investigate the effect of averaging on prediction performance, we varied the number of presentations that were averaged from one (single trial prediction) to 25. To estimate the generalization ability of the classifiers, we split each data set into two nonoverlapping subsets, trained each classifier on one subset (the training set) and tested on the second (the test set). The procedure was repeated 100 times for different random partitions of the data, and results were averaged across these 100 results. Statistical significance of the resulting performance scores was assessed by a permutation test, as described in [22].

2.3.2.1. Voxel selection. Before performing pattern recognition, we selected a subset of voxels for further analysis. The procedure of selecting relevant features (in this case, particular voxels) can greatly enhance classification performance [23,24], by reducing the dimension of the space of patterns to be labeled (avoiding the "curse of dimensionality") and removing "noise" features that can only degrade performance. This effect has been widely observed when classifying microarray data in bioinformatics, where the number of features greatly exceeds the number of observations [25].

Likewise, most classification of fMRI data depends on an effective feature selection procedure being applied beforehand [23]. In a typical fMRI study, time courses of more than 90,000 voxels are simultaneously acquired. Many of these are uninformative and could severely damage the performance of the algorithm. In order to perform pattern recognition more efficiently, one should use a priori knowledge to find a reasonable subset of voxels to feed the classifiers. For our voxel selection procedure, we simply chose those voxels with highest overall responsiveness, as measured by a GLM with respect to the stimulus convolved with the haemodynamic response.[1] We used different significance levels yielding six distinct voxel counts from 139 to 550.

2.3.2.2. Data preprocessing. Data were normalized by subtracting the mean value across conditions and presentations from each voxel and dividing by the overall standard deviation. Thus, each voxel had mean activity of zero and unit standard deviation. Means and standard deviations were estimated on the training set only and assumed to be of the same value on the test set. Removal of outliers consisted of setting all values that were beyond 3 standard deviations from the mean to a fixed value of 3 or −3, depending on the sign of the original value. Denoising of the data consisted of computing a singular value decomposition of the training set and only keeping as many principal components as were needed to preserve 90% of the total variance. The singular vector decomposition (SVD) was always done after outlier removal, since outliers would otherwise explain too great a proportion of the observed variance. The test set was denoised by projecting it onto components obtained by this SVD. We further discuss the effect of these procedures in our Results section.

2.3.2.3. Linear discriminant analysis. Linear discriminant analysis (LDA) is an algorithm which is used to separate two classes of data points by a linear hyperplane. In other words, it aims to find a linear combination of features (voxels) that discriminate between the two classes. In LDA, the weights of this linear combination are given by

$$\omega = \Sigma^{-1}(\mu_1 - \mu_2),$$

where μ_1 and μ_2 denote the respective means of the first and second classes and Σ is the aggregate covariance of data

[1] We also compared with an alternative voxel selection procedure, where we selected the 80 most active voxels per category, yielding 320 voxels in total to feed to the classifiers. In a comparison of the voxels selected according to overall responsiveness, we observe an overlap of only half of the voxels. This comes as no surprise, since it indicates that visually selective voxels are dominated by some of the categories. In other words, the information about individual categories is not equally distributed over the entire IT cortex. Classification results with this alternative selection technique were almost the same, however; thus, we do not report these results in detail.

points within each class. It is assumed that the covariance is the same for each of the two classes.

It can be shown that LDA yields the optimal decision boundary if the data in each class is well described by a normal distribution with equal covariances. It can be seen that the optimal weighting ω depends both on the means of each of the two classes and also on the covariance of the points within each class. In this respect, LDA is a more sophisticated classification technique than correlation analysis or GNB (see below), which do not take the covariance structure into account. On the other hand, estimation of the covariance Σ requires more data, and for this reason, LDA could perform suboptimally on small data sets.

2.3.2.4. Support vector machine. Like LDA, the support vector machine is a linear classification algorithm [26]. However, the SVM does not assume a specific model of the data points but rather seeks to find the hyperplane that separates the two classes with maximum margin. This hyperplane is constructed entirely in terms of inner products between the labeled training data (in this case, the BOLD response over a set of voxels). While we used a linear inner product, one can also define inner products between high-dimensional features of the data using a kernel function (see Ref. [26], Chapter 2, for more details). We also investigated a nonlinear variant with a Gaussian kernel, which yielded similar results to the linear SVM, but was much more sensitive to parameter settings (results not shown).

We used a Matlab implementation of the SVM classifier provided by the Spider toolbox (http://www.kyb.tuebingen.mpg.de/bs/people/spider/). Unlike other methods, the support vector machine requires careful parameter tuning if it is to perform satisfactorily. Thus, it is necessary to choose C, a scalar that controls the tradeoff between regularization and classification error). Within each training data set, the parameters were cross-validated on 10 folds in order to obtain the best value of the parameter C. The search space of linear SVM was C=2 to the power of −5 to 5.

2.3.2.5. Gaussian naïve Bayes. The algorithm and software released by Mitchell's lab were used (Machine Learning Department, School of Computer Science, Carnegie Mellon University, http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/) [3]. In short, the learning method was a GNB classifier [11]. The responses conditioned on the stimuli were modeled as Gaussians, where it was assumed that each voxel was independent of the others (i.e., no between-voxel correlations once we conditioned on the stimuli). The means and variances of the Gaussians were computed from the within-class training data. A reweighting was also made for differences in the ratio of stimuli in each class, although in our case, this ratio was close to 1 and the correction was not very relevant. The decision boundary for classification was then the optimal boundary given the model, and the predicted class on test data was the most probable class under this model.

2.3.2.6. Correlation analysis. The responses in the training set were averaged to compute the mean responses for each category as templates. For prediction, we obtained the correlation coefficients between each test point (which, depending on the task, could be a single trial or the averaged activity of multiple trials) and each of the templates. Then, each test point was predicted to belong to class one if the correlation coefficient for Class one was bigger than for Class two, and to Class two otherwise. Correlation analysis, GNB and LDA are equivalent if the within-class covariance is an identity matrix for each class.

3. Results

The functional activity obtained with the current study will be published elsewhere. A short description is given here. Consistent with the previous literature [12], the objects evoke strong, widespread activity in early visual areas and the IT cortex, which is known to be a higher-order visual processing area in primates [27,28]. Since we are interested in how well the brain can identify object categories by their higher order properties instead of low-level features, we restricted our analysis to IT cortex. The IT cortex was segmented according to anatomical criteria. The functional activity was detected by performing an F test on the visually responsiveness regardless of category. Despite the fact that we used an event-related paradigm, the stimuli we used were able to significantly activate more than 800 voxels in the IT cortex (P<.05, FWE corrected).

We first report the performance of the classifiers with optimized preprocessing and on average across all tasks. Fig. 1 shows the performance of each classifier for different numbers of voxels. Performance for individual pairs of categories was

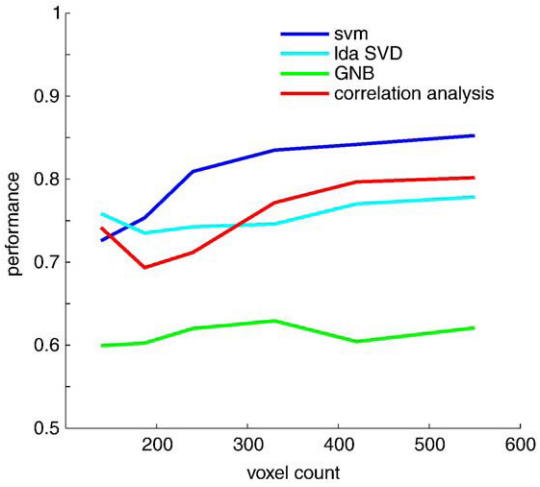

Fig. 1. Classifier accuracy as a function of dimensionality (number of voxels). Both the correlation analysis and SVM show a performance increase from 139 to 420 voxels. The performance of all classifiers saturates at around 400 voxels and remains at that level for higher voxel counts. SVM performs better than other methods after outlier removal, whereas correlation analysis and LDA perform better than GNB after outlier removal and SVD.

measured by repeatedly splitting the data into training and test sets and averaging classification performance on each test set. In our experiments, 100 such splits were used, and the data was split in a 70/30 training/test proportion.

Performance is best for the SVM, which achieves an accuracy of 85% (550 voxels, averaged across 25 trials, after outlier removal). A comparison of performance with chance level, as a function of the classification task, is given below. Both LDA and correlation analysis have an overall accuracy of 79%. GNB had lower performance scores, only achieving 62%.

The number of voxels used has a weak effect on the performance of the classifiers, at least across the range of voxels tested in this study (Fig. 1). While both the correlation analysis and SVM show a performance increase from 139 to 420 voxels, the performance of all classifiers saturates at around 400 voxels and remains at that level for higher voxel counts.

Fig. 2 shows the effect of multiple trial averaging on performance of individual pairwise classification tasks. Performance was again averaged over 100 random splits of the data into training and test sets but, this time, retaining 25 points of each class in the test set. When more than one trial was averaged, 25 points were generated in each category by drawing with replacement from this test pool. Note that the effective test sample size therefore decreases as the number of averaged trials increases.

Performance in the single-trial prediction tasks varies from 53% (GNB) to 55% (LDA) in the average across all tasks. To determine whether classifiers performed significantly higher than chance, we followed the shuffling procedure described in [22]. Specifically, we repeatedly permuted the labels relative to the classes, then obtained an empirical distribution of test errors as described above, and finally used the 95% quantile of this distribution as a significance threshold. The number of necessary permutations can be obtained by plotting the stability of the resulting threshold as a function of permutation number; we found 250 permutations to be sufficient.

Single trial performance was significantly above chance level ($P<.05$ permutation test [22]) on three of the six classification tasks (e.g., face vs. house, Fig. 2C). Raw classification rates improved markedly even after averaging across as few as three presentations and could be as high as 100% when averaging across 25 trials. However, this averaging reduces the number of independent test points in the data set, and results obtained from averaging were never significant if the single trial analysis was not.

The two different preprocessing steps (outlier removal and denoising by SVD) had a substantial impact on prediction accuracy (Table 1). Removal of outliers (even without SVD) greatly enhanced performance for most classifiers: for example, the mean accuracy of SVM across data sets (using 550 voxels and averaging across 25 trials) was enhanced by 12%, and the accuracy of correlation analysis increased from 69% to 78%. Denoising the data by SVD helped all algorithms besides SVM (and actually decreased SVM performance by 10%). For instance, the accuracy of GNB was enhanced by 10% when we used SVD and outlier removal.

It is not possible to discriminate houses and fractals in our experiments, since no method is significantly different from chance level. Curiously, both the correlation coefficient and

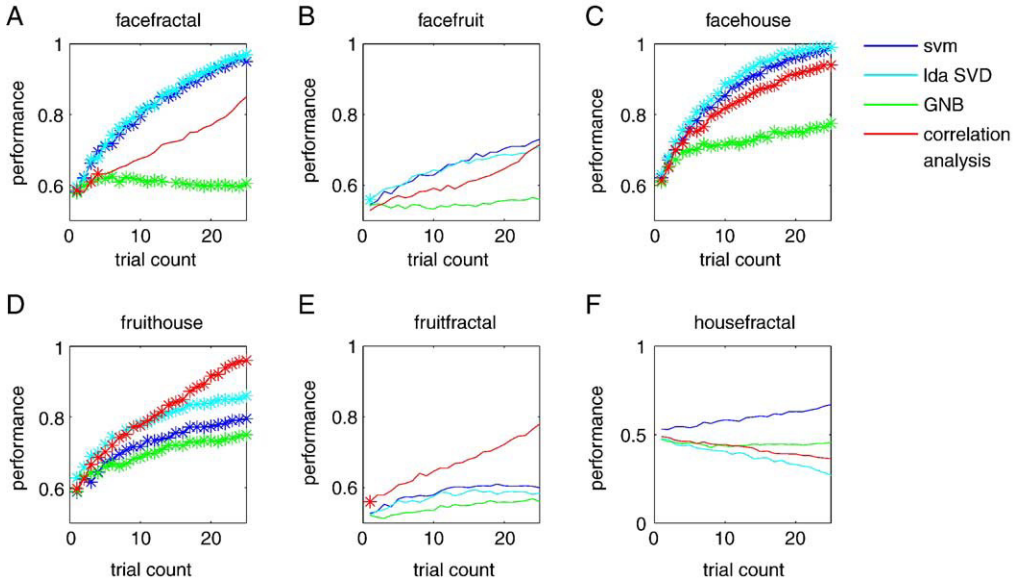

Fig. 2. The effect of multiple trial averaging on performance for individual pairwise classification tasks. Each plot indicates the performance of each classifier for different stimulus pair combinations, as a function of number of average trials. (A) Face vs. fractal trials; (B) face vs. fruit trials; (C) face vs. house trials; (D) fruit vs. house trials; (E) fruit vs. fractal trials; (F) house vs. fractal trials. In most of the cases trial averaging improves performance. Asterisks indicate where performance is significantly above chance level, with P value less than .05 (see results section for details of the permutation test for statistical significance).

Table 1
Mean discrimination performances of different classifiers with different preprocessing procedures (comparison of effects of preprocessing)

Outlier removal	No				No			
SVD	No				Yes			
	SVM	LDA	GNB	Corr	SVM	LDA	GNB	Corr
Face fractal	0,8100		0,5331	0,7300	0,9000	0,9700	0,6100	0,7550
Face fruit	0,7550		0,5167	0,6400	0,7150	0,6850	0,5600	0,7050
Face house	0,9000		0,5743	0,8400	0,9750	1,0000	0,6850	0,9500
Fruit house	0,7250		0,5575	0,8300	0,8600	0,8700	0,6750	0,9500
Fruit fractal	0,5500		0,5209	0,7000	0,6400	0,5600	0,5600	0,7100
House fractal	0,6600		0,4729	0,4300	0,3750	0,3650	0,4350	0,3950
Mean	0,7333		0,5292	0,6950	0,7442	0,7417	0,5875	0,7442
Outlier removal	Yes				Yes			
SVD	No				Yes			
	SVM	LDA	GNB	Corr	SVM	LDA	GNB	Corr
Face fractal	0,9650		0,5550	0,8400	0,9400	0,9650	0,6100	0,9200
Face fruit	0,7900		0,5300	0,7100	0,7100	0,7750	0,5400	0,8300
Face house	1,0000		0,5250	0,9900	0,9750	1,0000	0,8500	0,9900
Fruit house	0,9350		0,5150	0,9700	0,9100	0,9500	0,7700	0,9900
Fruit fractal	0,7100		0,5450	0,8500	0,5800	0,6600	0,5450	0,8000
House fractal	0,7150		0,5150	0,3500	0,4150	0,3200	0,4100	0,2800
Mean	0,8525		0,5308	0,7850	0,7550	0,7783	0,6208	0,8017

LDA was not performed without SVD.

LDA perform below chance (50%) for the house-vs.-fractal task, although the difference is not statistically significant. We return to this point in the Discussion.

4. Discussion

We have provided the first systematic study of the most commonly employed multivariate classification methodologies in an awake monkey imaging setting at high field. We have demonstrated that it is possible to infer which category the subject was viewing by classifying single trials or mean patterns in imaging data acquired with an event-related paradigm. After selecting visually responsive voxels in the IT cortex, linear SVM, correlation analysis and LDA can classify the patterns. GNB can separate the categories to some extent but performs much worse than reported in other studies [3].

Since GNB performs so much worse than the other methods, we first provide a more detailed breakdown of our observations for SVM, correlation and LDA for individual tasks, and then return to GNB in subsequent paragraphs. We see from Fig. 2 that, in all tasks requiring the discrimination of faces from other categories, LDA and SVM perform equivalently and better than other approaches (although in the face-vs.-fruit category, the performance not good enough to be statistically significant, and no conclusions should be drawn). Correlation does well when distinguishing fruit from house, being both significantly above chance and performing better at high-average trial counts. That said, when distinguishing fruit vs. fractal, performance is almost never significantly above chance: thus, even though correlation performance is highest, we cannot conclude from our data that any classifier is satisfactory. If the LDA modeling assumptions are correct, we would always expect LDA to outperform correlation, since the former takes into account the covariance structure of the data. Correlation can be more accurate, however, if the data do not conform to the Gaussian assumptions made in LDA, or if one does not have sufficient data to obtain a good estimate of the covariance. In both cases, normalization by the covariance matrices can be misleading and detrimental to classification performance. We emphasize that this is not the case when faces are to be distinguished from fractals: in this case, it is clear that LDA and SVM are more resistant to noise, since they require averaging over fewer test images than correlation to obtain a given level of performance.

We now address the problem of distinguishing houses from fractals (Fig. 2F). In this case, the performance of correlation and LDA actually decreases when the test images are averaged over more instances, while linear SVM performance improves; that said, the performance difference from chance is never statistically significant for any method. The assumptions made in both correlation and LDA appear to have failed, which can occur when a relatively small number of unusual points cause the means and/or covariances to be unrepresentative of the data at large. As SVM does not assume a specific model of the data, but rather directly maximizes the margin between the classes, it may be more robust to these unrepresentative points. Generally speaking, SVMs are able to make use of information distributed across a large number of voxels, each of which is only weakly informative. To confirm whether this advantage truly holds for the house-vs.-fractal task, more data would be needed to verify above-chance performance.

We now return to the performance of GNB relative to the other methods. To our surprise, GNB did not perform as well as reported in the previous study [3] (although a previous comparison of GNB and SVM (in [29], Table 1, last two rows) also reported relatively poor performance for GNB, albeit for a whole-brain study in humans without feature selection). Besides the fact that our data were acquired with an event-related paradigm, while most other data were obtained with block design, there may be additional reasons for the poor performance of GNB. Although GNB also performs implicit feature selection (as does SVM), the estimation procedure suffers from a suboptimal assumption. The learning process of GNB assumes that the features (voxels, in our case) are independent conditioned on the stimulus, which is not true of fMRI data. Fig. 3 shows an example scatterplot (Fig. 3A) of the activation of two voxels individually for two classes (face and house) and a histogram of the within-class correlation coefficient of all pairs of voxels (Fig. 3B). Over 35% of the voxel pairs show significant positive correlation within stimulus classes, and 15% show significant negative correlation (Fig. 3B, $P<.05$, obtained using repeated shuffling of all individual voxel activations across trials). It is known that the neighboring voxels are not independent due to the intrinsic point spread function of the imaging device, and the correlation between voxels due to physiological effects (such as blood supply of brain tissue from a common capillary bed), even without spatial smoothing. Therefore, it is important to take the covariance between features into account when constructing a Gaussian classifier.

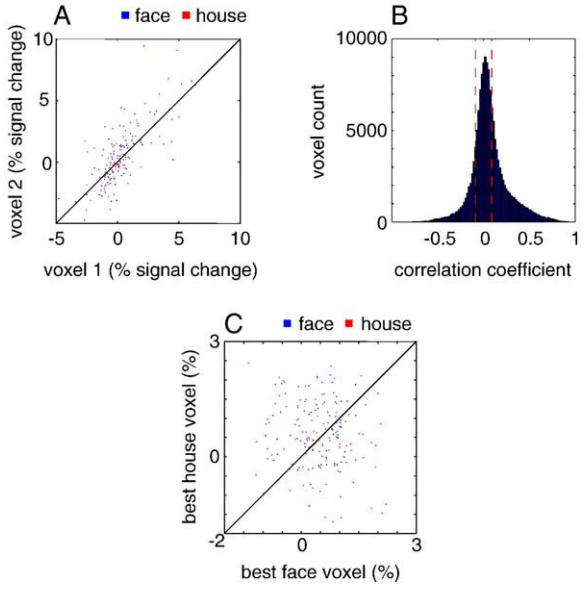

Fig. 3. Correlation between voxel-pairs. To examine why GNB did not perform well with our data set, we plotted the amplitudes of face and house trials extracted from two neighboring voxels selected at random (A) and the distribution of the within class correlation coefficients of all pairs of voxels, averaged across all four classes (B). We computed the correlation coefficient between all pairs out of 550 voxels, yielding 150,975 pairs in the plot. The red vertical dashed lines indicate the significant positive and negative correlation coefficient levels at $P=.05$, obtained using repeated shuffling of all individual voxel activations across trials. Thirty-five percent of the pairs are significantly positively correlated and 15% are significantly negatively correlated. (C) The amplitudes of face and house trials extracted from the most face-selective voxel (x-axis) and house-selective voxel (y-axis). The two voxels are not significantly correlated ($r=-0.02$).

Even in the absence of such correlation, however, we have found that the linear SVM can outperform GNB on our data. To illustrate this, we chose the two voxels that show greatest activation for faces and houses, respectively, and performed face vs. house classification using these voxels alone. Linear SVM was able to discriminate faces from houses with 70% accuracy (on test points averaged over 27trials), whereas the performance of GNB was 62%. In this case, the two voxels were not significantly correlated ($r=-0.02$, $P=.43$, Fig. 3C).

Classification performance in the single-trial prediction task was poor (although due to the greater effective data set size, performance nonetheless exceeded chance level in many cases). The noisy fMRI data makes it difficult to perform trial-by-trial analysis. It is possible, however, to improve the performance by selecting more discriminative voxels. In the present study, only the overall responsiveness was taken into account in voxel selection. The reliability of the response of a voxel to visual stimulus does not necessarily imply category-selective information.

It is known that classifiers will perform better as more informative dimensions (i.e., voxels) are added [5,23,30]. In most of the previous human imaging studies, classifier performance reaches a maximum at about 100 voxels. In our study, performance leveled off at around 400 voxels. This could simply be due to the fact that with higher resolution (as in the present study), the same brain volume corresponds to more voxels. Voxel size at lower field (such as 3 T) is typically 3×3×3 mm, which corresponds to four to five voxels in the present study. In addition, smaller voxels have a higher chance of containing a more homogeneous neural population; therefore, the individual voxels could contain more discriminative information. However, high-resolution fMRI without sufficient signal-to-noise ratio could decrease the discrimination ability of each voxel. Since reliability is also one of the most important concerns to obtain high classification performance, we acquired the data at high field (7 T). It would be useful to perform a systematic study of classification vs. field strength in the future.

We conclude with a suggestion for a general-purpose classification strategy in future fMRI studies, based on our observations. This strategy is also informed by considerations of computing time, since both correlation and LDA have a substantially lower cost than SVM (this is to a great extent due to the cross-validation needed to select the parameter C). First, denoise the data by outlier removal and SVD. Second, classify the data using both LDA and correlation and retain the classifier with highest performance, as long as the performance is significantly better than chance. In cases where such a classification does not yield values significantly above chance level, try an SVM with

outlier removal alone (i.e., without SVD denoising) to classify the data.

Further work could include extending the analysis to other brain regions with different paradigms. For example, voxel selection is limited to the IT cortex in the present study. Using a similar analysis, we could determine the predictability of visual categories in other brain areas. Further classification algorithms should also be added to the comparison, for instance (kernel) logistic regression. It is also of interest to examine whether even more information could be extracted from fMRI data through more sophisticated feature selection, including kernel approaches [23,25], and to expand the classifier comparison to a greater number of algorithms. SVD might not be optimal for dimensionality reduction, since it only accounts for second order moments. Higher-order methods such as independent component analysis could also be compared. In general, we expect that tools from machine learning will yield insight into neural representations of information that are not accessible via more classical statistical methodologies.

Acknowledgments

The authors thank Dr. J. Goense, Dr. A.S. Tolias and T. Steudel for their help in obtaining the data, and Dr. N. Kriegeskorte and A. Ecker for valuable suggestions. The work was supported by the Max-Plank Society, and in part by the IST Program of the European Community, under the FP7 Network of Excellence, ICT-216886-NOE.

References

[1] Friston KJ, Worsley KJ, Frackowiak RS, Mazziotta JC, Evans AC. Assessing the significance of focal activations using their spatial extent. Hum Brain Mapp 1994;214(1):7.
[2] Haynes JD, Rees G. Decoding mental states from brain activity in humans. Nat Rev 2006;7(7):523–34.
[3] Mitchell TM, Hutchinson R, Just MA, Niculescu RS, Pereira F, Wang X. Classifying instantaneous cognitive states from FMRI data. AMIA Annual Symposium proceedings/AMIA Symposium; 2003. p. 465–9.
[4] LaConte SM, Peltier SJ, Hu XP. Real-time fMRI using brain-state classification. Hum Brain Mapp 2007;28(10):1033–44.
[5] Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 2003;19(2 Pt 1):261–70.
[6] Aguirre GK, Zarahn E, D'Esposito M. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. Neuron 1998;21(2):373–83.
[7] Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature 1998;392(6676):598–601.
[8] Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 1997;17(11):4302–11.
[9] Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science (New York, NY) 2001;293(5539): 2425–30.
[10] Carlson TA, Schrater P, He S. Patterns of activity in the categorical representations of objects. J Cogn Neurosci 2003;15(5):704–17.
[11] Ng AY, Jordan M.I. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. NIPS; 2001.
[12] Tsao DY, Freiwald WA, Knutsen TA, Mandeville JB, Tootell RB. Faces and objects in macaque cerebral cortex. Nat Neurosci 2003;6(9): 989–95.
[13] Gati JS, Menon RS, Ugurbil K, Rutt BK. Experimental determination of the BOLD field strength dependence in vessels and tissue. Magn Reson Med 1997;38(2):296–302.
[14] Ugurbil K, Toth L, Kim DS. How accurate is magnetic resonance imaging of brain function? Trends Neurosci 2003;26(2):108–14.
[15] Yacoub E, Duong TQ, Van De Moortele PF, et al. Spin-echo fMRI in humans using high spatial resolutions and high magnetic fields. Magn Reson Med 2003;49(4):655–64.
[16] Pfeuffer J, Merkle H, Beyerlein M, Steudel T, Logothetis NK. Anatomical and functional MR imaging in the macaque monkey using a vertical large-bore 7 Tesla setup. Magn Reson Imaging 2004;22(10): 1343–59.
[17] Keliris GA, Shmuel A, Ku SP, et al. Robust controlled functional MRI in alert monkeys at high magnetic field: effects of jaw and body movements. Neuroimage 2007;36(3):550–70.
[18] Goense J, Ku SP, Merkle H, Tolias AS, Logothetis NK. fMRI of the temporal lobe of the awake monkey at 7T. Neuroimage 2008.
[19] Gruetter R. Automatic, localized in vivo adjustment of all first- and second-order shim coils. Magn Reson Med 1993;29(6):804–11.
[20] Logothetis NK, Guggenberger H, Peled S, Pauls J. Functional imaging of the monkey brain. Nat Neurosci 1999;2(6):555–62.
[21] Frahm J, Haase A, Matthaei D. Rapid three-dimensional MR imaging using the FLASH technique. J Comput Assist Tomogr 1986;10(2): 363–8.
[22] Golland P, Fischl B. Permutation tests for classification: towards statistical significance in image-based studies. Inf Process Med Imaging 2003;18:330–41.
[23] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:26.
[24] Keerthi SS, Lin CJ. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 2003;15(7):1667–89.
[25] Song L, Bedo J, Borgwardt KM, Gretton A, Smola A. Gene selection via the BAHSIC family of algorithms. Bioinformatics (Oxford, England) 2007;23(13):490–8.
[26] Schoelkopf B, Smola A. Learning with kernels (chapter 2). Cambridge (MA): MIT Press; 2002. p. 187–404.
[27] Desimone R, Schein SJ, Moran J, Ungerleider LG. Contour, color and shape analysis beyond the striate cortex. Vision Res 1985;25 (3):441–52.
[28] Desimone R, Ungerleider LG. Multiple visual areas in the caudal superior temporal sulcus of the macaque. J Comp Neurol 1986;248(2): 164–89.
[29] Pereira F, Gordon G. The support vector decomposition machine. Proc of ICML, Pittsburgh, PA; 2006.
[30] Haynes JD, Rees G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 2005;8(5): 686–91.