Decoding Cognitive Processes from functional MRI

Oluwasanmi Koyejo[1] and Russell A. Poldrack[2]

[1] Imaging Research Center, University of Texas at Austin
sanmi.k@utexas.edu
[2] Depts. of Psychology and Neuroscience, University of Texas at Austin
poldrack@utexas.edu

Abstract. The goal of cognitive neuroscience is to understand the the brain processes that underlie cognitive function. These brain processes are studied by examining neural responses to experimental tasks and stimuli. While most experiments are designed to isolate a single cognitive process, the resulting brain images often encode multiple cognitive processes simultaneously. Thus standard classification methods are inappropriate for decoding cognitive processes. We propose a multilabel classification approach to decoding, and present experimental results showing that this approach accurately predicts of the set of cognitive processes associated with an experimental contrast image.

1 Introduction

An important hypothesis in modern cognitive neuroscience is that brain function is decomposable into a set of elementary cognitive processes [17]. These cognitive processes represent the basis set of brain functions recruited for cognitive tasks. For example, recognizing a face may require the cognitive process of vision, working memory and retrieval, while the music comprehension may require, in addition to the shared cognitive processes of working memory and retrieval, additional cognitive processes of rhythm and intonation. Cognitive neuroscientists and other researchers measure these cognitive processes in the laboratory setting by developing experiments that allow (e.g., via cognitive subtraction) the isolation of a specific cognitive process from other recruited processes. For example, in the stop signal task [6], participants are instructed to respond as quickly as possible to a particular stimulus such as an image. On a few trials, the stimulus is followed by a stop signal which instructs the participants to withhold a response. Differences between the brain images measured during the stop signal task are often used to study the cognitive process of response inhibition or self control [1]. Unfortunately, despite careful selection of the stimuli and control tasks, the measured brain function often contains multiple cognitive processes [10].

Functional magnetic resonance imaging (fMRI) has enabled the non-invasive measurement of brain function in response to experimental stimuli at fine spatial scales. From initial studies that used classifiers to discriminate between different classes of visual objects [5] to more recent studies showing large scale classification across experiments [14], decoding from brain images has become an important research tool [9]. Decoding performance can be used as an indicator to test hypothesis about the cognitive content of the brain images, and once learned, the classifiers can be used for

additional verification of experimental studies by testing the measured images. Further, the weight vectors estimated by the model can be used to localize predictive voxels [4], or select regions of interest for additional processing. In addition to the scientific utility of decoding in general, the specific application to cognitive processes may help address additional scientific questions, such as which cognitive processes outlined in the literature represent true differences in brain function, and which merely reflect theoretical distinctions [12]. Despite these potential insights, direct decoding of cognitive processes from brain function has not been attempted before.

Our approach is inspired by the work in [14], where the cognitive processes associated with experimental contrast images were analyzed by projection onto the latent dimensions of a neural network classifier. It was also noted by [14] that brain images computed from each contrast are often associated with multiple cognitive processes, especially when those contrasts are relatively coarse. We propose a multilabel classification approach for decoding. Multilabel classifiers are designed to solve classification problems where each example may be associated with multiple labels, and are popular in several domains such as image processing and text processing [19] In the neuroimaging domain, the work most closely related to ours is Neurosynth [18], a text based meta-analytics platform. Neurosynth is trained using the text in published neuroimaging papers as labels, and using the corresponding reported significant voxels as "brain images". Trained using a the naïve Bayes model, Neurosynth can be applied to both encoding and decoding. [18] focused only on standard classification. However, in principle, the naïve Bayes classifier learns a set of probabilities that can be thresholded to predict multiple labels.

Our work is enabled by the recent availability of a large public fMRI database (OpenFMRI[3]) [11] and a large cognitive ontology labeled by domain experts (Cognitive Atlas[4]) [16]. We study the decoding of cognitive processes from brain function measured via functional magnetic resonance imaging (fMRI) contrasts. We focus on the subclass of multilabel classification methods known as label decomposition methods [19], where the multilabel classification problem is decomposed into multiple binary classification problems (Section 2). Our results provide experimental evidence that the cognitive processes can be decoded accurately (Section 3). Our results also suggest that the metrics used for evaluating the decoding performance should be selected carefully due to the label imbalance observed in in typical datasets.

Notation: We denote vectors by bold-face lower case letters $\mathbf{x}$ and matrices by bold-face capital letters $\mathbf{X}$. The set of real valued D dimensional vectors are denoted by $\mathbb{R}^D$. Label sets are denoted by script capital letters $\mathcal{S}$, and $|\mathcal{S}|$ evaluates the cardinally of the set $\mathcal{S}$.

2 Methods

Let $\mathbf{x}_n \in \mathbb{R}^D$ denote the n^{th} brain volume with voxels collected in a real valued D dimensional vector. The total number of brain volumes is represented by N. Each brain volume is associated with a set of labels $\mathcal{S}_n = \{s_1, \ldots s_K\}$ chosen from the full set

[3] openfmri.org

[4] www.cognitiveatlas.org

Table 1. Cognitive Concepts Sorted by Prevalence in Data.

Code	Concept
A	Vision
B	Action Execution
C	Decision Making
D	Orthography
E	Shape Vision
F	Audition
G	Phonology
H	Conflict
I	Semantics
J	Reinforcement Learning
K	Working Memory
L	Feedback
M	Response Inhibition
N	Reward
O	Stimulus-driven Attention
P	Speech
Q	Emotion Regulation
R	Mentalizing
S	Punishment
T	Error Processing
U	Memory Encoding
V	Spatial Attention

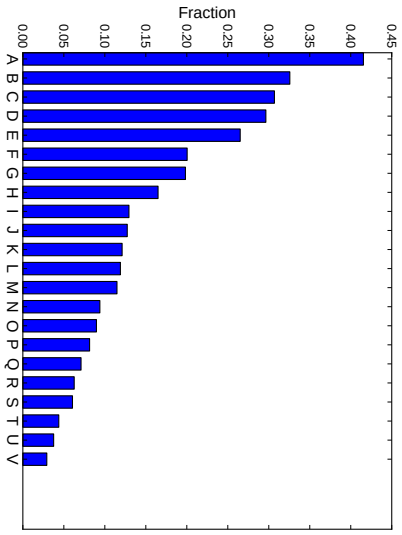


Fig. 1. Fraction of Data with each Cognitive Concept

of possible labels $\mathcal{L} = \bigcup_{n=1,\ldots,N} \mathcal{S}_n$ with $|\mathcal{L}| = L$. Multilabel classification involves estimating a predictive mapping $f : \mathbf{x}_n \mapsto \mathcal{S}_n$. There are several approaches in the literature for multilabel classification including label decomposition, label ranking, and label projection methods [19]. We focus on label decomposition methods due to their simplicity, scalability and ease of interpretation. Label decomposition methods separate the multilabel classification task into a set of binary classification tasks. A popular approach in this family is the *One-Vs-All* decomposition, where the multilabel classification is decomposed into binary classification tasks. Each binary classification model is trained to predict the presence or absence of each each label independently.

We experimented with the multilabel decomposition approach using the following base classifiers: (i) l_2 regularized support vector machine (**SVM**) [3] , (ii) l_2 regularized logistic regression (**Logistic**) [2] , and l_2 regularized squared loss classifier (**Ridge**) [2] . In each decomposed classification, we use linear models of the form $f(\mathbf{x}_n) = \mathbf{w}^\top \mathbf{x}_n$ where $\mathbf{w} \in \mathbb{R}^D$ is a real valued weight vector. In addition, we experimented with a baseline classifier where the set of predicted labels are determined based on prevalence in the training set (**Popularity**). Here, each label is drawn independently from a Bernoulli distribution with probability given by the fraction of examples where the label appeared in the training data. We also present results using a baseline that always predicts that all the labels are present (**All Classes**) i.e. $\mathcal{S}_n = \mathcal{L}$.

3 Experimental Results

We experimented with brain image data from the publicly available openfMRI database [11]. OpenfMRI contains pre-extracted z statistic contrasts for each subject computed using a generalized linear model. This data extraction was implemented using the FMRIB Software Library (FSL). Combining the whole brain data with the standard brain mask resulted in $D = 174,264$ extracted voxels. We extracted $N = 479$ contrast images for $K = 26$ contrasts. Further details on data preprocessing may be found in [11]. In addition to the brain volumes, we extracted a list of cognitive concepts associated with each experimental contrast. The list was curated starting from labels in the Cognitive Atlas [16] and refined by domain experts. The final set of cognitive concepts are provided in Table 1 and the fraction of examples containing each label is shown in Fig. 1. It is clear that some concepts are significantly more prevalent in the data than other concepts. For example vision is more than 20 times more prevalent than spatial attention. The data samples included an average of 3.5 labels per example with a maximum of 9 labels per example and a minimum of 1 label per example.

We evaluated the models using (label) *Accuracy*, *Precision*, *Recall*, *Hamming loss* and *F1score*, metrics commonly applied for evaluating multilabel classification [19]. Let $\mathcal{S}_n$ represent the true labels and $\mathcal{Z}_n$ represent the predicted labels. The metrics are computed as:

$$\text{Precision} = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mathcal{S}_n \cap \mathcal{Z}_n|}{|\mathcal{Z}_n|}, \qquad \text{Accuracy} = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mathcal{S}_n \cap \mathcal{Z}_n|}{|\mathcal{S}_n \cup \mathcal{Z}_n|},$$

$$\text{Recall} = \frac{1}{N}\sum_{n=1}^{N}\frac{|\mathcal{S}_n \cap \mathcal{Z}_n|}{|\mathcal{S}_n|}, \qquad \text{Hamming Loss} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}|\mathcal{S}_n \ominus \mathcal{Z}_n|,$$

$$\text{F1score} = \frac{1}{N}\sum_{n=1}^{N}\frac{2.0 * \text{Precision}_n \times \text{Recall}_n}{\text{Precision}_n + \text{Recall}_n},$$

where $\mathcal{A} \ominus \mathcal{B}$ represents the symmetric difference of set $\mathcal{A}$ and $\mathcal{B}$. Label *Accuracy* measures the average fraction of labels that are predicted correctly evaluated with respect to the total number of true and predicted labels, while *Precision* and *Recall* measure the fraction of labels that are predicted correctly with respect to the number of predicted labels and the number of true labels respectively. Thus, *Precision* measures the fraction of predicted labels that are relevant, and *Recall* measures the fraction of relevant labels that are predicted. It is worth noting that the *Recall* metric is not sensitive to false positives. In particular, we show that the **All Classes** model achieves perfect recall without any learning. In contrast, the *Hamming Loss* directly penalizes both false positives and false negatives, and *Precision* penalizes false positives via the scaling in the denominator term. The *F1score* combines *Precision* and *Recall* into a single score. Higher scores indicate superior performance for *Accuracy*, *Precision*, *Recall* and *F1score*, and the best possible score is 1. In contrast, lower cores indicate superior performance for *Hamming Loss*, and the best possible score is 0. Further details on the metrics are variable in [19].

All models were trained using 5-fold double loop cross validation. The inner loop was used for model selection, and the outer loop was used to estimate the generalization performance. The l_2 regularization parameters for all models was selected from the

set $\{10^2, 10^{1.5}, 10^1, \ldots, 10^{-2.5}, 10^{-3}\}$. We used the *Hamming Loss* metric for model selection. We evaluated the use of the other metrics for parameter selection and found the results to be qualitatively equivalent. In addition to performance comparisons, we were interested in evaluating the statistical significance of the results. Hence, we computed an empirical null distribution by randomly permuting the labels 1000 times and retraining the model. Note that the empirical null distribution was estimated separately for each trained model, so the statistical significance is model dependent. We computed statistical significance using a threshold of $p = 10^{-3}$, suggesting high confidence in rejecting the hypothesis that the statistically significant results were the result of chance. Performance comparisons between models were evaluated using the mean and variance of the scores over the cross validation sets.

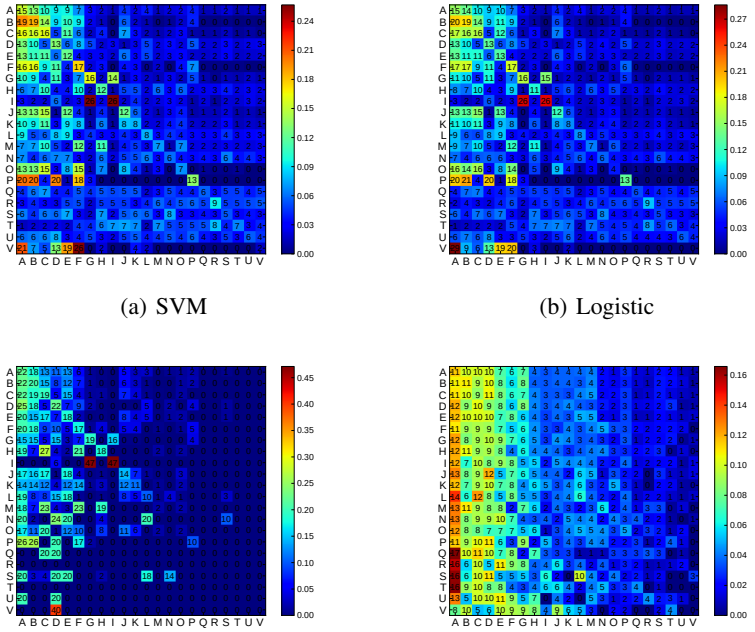
Experimental performance results are presented in Table 2. Examining the effect of label imbalance, we found that the baseline **All Classes** model achieved perfect *Recall* without any learning suggesting caution in interpreting evaluation metrics such as *Recall* with such severely imbalanced labels. We found that the performance of **SVM** and **Logistic** were almost identical. **Ridge** was comparable to **SVM** and **Logistic** in terms of *Precision*, but performed worse in terms of *Accuracy* and *Recall*. However, **Ridge** outperformed all other models in terms of *Hamming Loss*. This behavior was investigated further by evaluating confusion matrices shown in Fig. 2. Our results showed that the accurate *Hamming Loss* performance of **Ridge** was due to a lower false positive rate at the expense of overall accuracy. In particular, we found that both **SVM** and **Logistic** were more prone to false positive errors. This empirical observation is consistent with the observation that the *Hamming Loss* penalizes false positives directly. Thus, we conjecture that in severely imbalanced datasets, multilabel classifiers tuned for *Hamming Loss* will have lower false positive rates, possibly at the expense of label *Accuracy* and *Recall*.

4 Conclusion

The decoding of cognitive processes is an important first step towards evaluating and verifying the latent processes the brain employs to complete various tasks. We have provided experimental evidence that cognitive labels can be accurately decoded from brain function using a multilabel classification approach. We also studied some of the trade-offs between accuracy and false positive rate that arise due to the imbalance of the labels. We intend to continue further verification of the decoding performance by evaluating various multilabel classification methods in the literature [19]. This will also aid in understanding the trade-offs between different methods in the specific application to neuroimaging data. In addition, we plan to incorporate structured regularizers such as the total variation regularization [7], or Bayesian models for structured sparsity [8] that are may be able to localize the sources of classification performance, improving the interpretability of our results.

Table 2. Mean (var) of Multilabel Performance Metrics. *Note*: Higher scores are better for all metrics apart from Hamming Loss, where lower scores are better. * - represents models where all metrics are significant wrt. the permutation based null distribution for the model.

	Accuracy	Precision	Recall	F1score	Hamming Loss
SVM*	0.43 (0.03)	**0.53 (0.03)**	0.68 (0.03)	0.51 (0.03)	0.21 (0.01)
Logistic*	**0.44 (0.03)**	**0.53 (0.02)**	0.68 (0.03)	**0.52 (0.03)**	0.21 (0.01)
Ridge*	0.34 (0.02)	0.47 (0.02)	0.37 (0.02)	0.39 (0.02)	**0.09 (0.00)**
Popularity	0.12 (0.01)	0.21 (0.02)	0.18 (0.03)	0.18 (0.02)	0.24 (0.01)
All Classes	0.15 (0.00)	0.15 (0.00)	**1.00 (0.00)**	0.25 (0.00)	0.85 (0.00)



(a) SVM (b) Logistic

(c) Ridge (d) Popularity

Fig. 2. Avg. of Normalized Confusion Matrices. Row (True process) and Column (predicted process) **Note**: the matrix entries are scaled $\times 10^2$. Confusion matrices are normalized per row for each validation fold. Intensity levels are computed separately for each row. Note that the normalized rows need not sum to 1 in multilabel classification. Cognitive concepts are coded as capital letters $A, \ldots, V$ (see Table 1).

Bibliography

[1] Aron, A.R., Poldrack, R.A.: Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. The Journal of Neuroscience 26(9), 2424–2433 (2006)

[2] Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)

[3] Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 121–167 (1998)

[4] De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E.: Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58 (2008)

[5] Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293(5539), 2425–2430 (2001)

[6] Logan, G.D., Cowan, W.B., Davis, K.A.: On the ability to inhibit simple and choice reaction time responses: a model and a method. Journal of Experimental Psychology: Human Perception and Performance 10(2), 276 (1984)

[7] Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for fmri-based prediction of behavior. Medical Imaging, IEEE Transactions on 30, 1328–1340 (2011)

[8] Park, M., Koyejo, O., Ghosh, J., Poldrack, R.R., Pillow, J.W.: Bayesian structure learning for functional neuroimaging. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2013)

[9] Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: A tutorial overview. NeuroImage 45, S199–S209 (2009)

[10] Poldrack, R.A.: Subtraction and beyond: The logic of experimental designs for neuroimaging. In: Hanson, S.J., Bunzl, M. (eds.) Foundational Issues in Human Brain Mapping, pp. 147–160. MIT Press, Cambridge, MA (2010)

[11] Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., Milham, M.P.: Towards open sharing of task-based fMRI data: The OpenfMRI project. Frontiers in Neuroinformatics (2013)

[12] Poldrack, R.A.: Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72(5), 692–697 (2011)

[13] Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., Devlin, J.T., Cumba, C., Koyejo, O., Milham, M.P.: Toward open sharing of task-based fmri data: the openfmri project. Front Neuroinform 7, 12 (2013)

[14] Poldrack, R.A., Halchenko, Y.O., Hanson, S.J.: Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychological Science 20, 1364–1372 (2009)

[15] Poldrack, R.A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D.S., Sabb, F.W., Bilder, R.M.: The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Front Neuroinform 5, 17 (2011)

[16] Poldrack, R.A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D.S., Sabb, F.W., Bilder, R.M.: The cognitive atlas: Towards a knowledge foundation for cognitive neuroscience. Frontiers in Neuroinformatics 5 (2011)
[17] Posner, M.I., Petersen, S.E., Fox, P.T., Raichle, M.E.: Localization of cognitive operations in the human brain. Science 240(4859), 1627–31 (Jun 1988)
[18] Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D.: Large-scale automated synthesis of human functional neuroimaging data. Nature methods 8(8), 665–670 (2011)
[19] Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering 99(PrePrints), 1 (2013)