Detecting Cognitive States Using Machine Learning

Xuerui Wang & Tom Mitchell
Center for Automated Learning and Discovery
School of Computer Science
Carnegie Mellon University
{xuerui,tom.mitchell}@cs.cmu.edu

October 9, 2002

Abstract

Very little is known about the relationship between the cognitive states and the fMRI data, and very little is known about the feasibility of training classifiers to decode cognitive states. Our efforts aimed to automatically discover which spatial-temporal patterns in the fMRI data indicate a subject is performing a specific cognitive task, such as watching a picture or sentence. We developed machine learning methods that can be used to discover such spatial-temporal fMRI patterns, across subjects and contexts, which support probalistic predictions about the cognitive states of the human subjects.

1 Introduction

The study of human brian function has received a tremendous boot in recent years from the advent of new brain imaging technique, functional Magnetic Resonance Imaging(fMRI), which dramatically improve our ability to collect data about brain activity in human subjects performing tasks such as reading, answering questions, comparing images, solving algebra problems, or driving simulated vehicles. A typical fMRI experiment can produce a three-dimensional image of the human subject's brain activation every half second, at a spatial resolution of 3mm, resulting in tens of millions of observations of local brain activation over the course of a single twenty minute experiment. Quickly, scientists found themselves drowning in a flood of data and in need of computer support to extract general principles from the millions of observed data points.

Whereas much work has been done to develop fMRI data analysis methods that average together data from multiple stimuli and episodes in order to determine which brain regions are involved on average in various cognitive tasks, very little is known about the feasibility of training classifiers

to decode cognitive states from single episodes. Our research aimed to develop a new category of computer tools to assist scientists in formulating theories of cognitive brian function from fMRI data. In this report, we developed software that will automatically detect the time intervals and specifi c brain regions in which a subject is performing unreported cognitive activities such as watching a sentence or picture. These detectors of cognitive activities and states have been developed using machine learning algorithms, and are used to parse and segment fMRI data from subsequent experiments in which these cognitive activities are otherwise unobservable, producing results such as "at the time being the subject appears with probability 0.9 to be watching a picture." Before, most of the similar researches done by our colleagues are based upon the fMRI data of a single human subject because of the well-known diffi culty, the subject-to-subject variation. The key point of our software is that it can be trained and applied across subjects and contexts, that is, the data we used are from an experiment performed by different subjects, or even different experiments, but we can use them as if they were from an experiment performed by a single subject.

Our classifi ers could be used as virtual sensors of hidden cognitive states, which would be of tremendous use for experimental research in cognitive science and in diagnosis of mental processes in patients with brain injuries. The results in this report will be of great help to support the study of cognitive brain function, such as learning transitions among states and discovering abstractions. Furtherly, it will help scientists discover, represent, and evaluate the correspondence between the components of their cognitive theory, and the growing database of experimental results from fMRI and other sources.

In Section 2, we simply introduced the dataset and the corresponding experiment we used in this report. And then we presented the methods we adapted in this report in Section 3. The results achieved across subjects and across contexts are detailed in Section 4 and Section 5, respectively. We also reported some interesting fi ndings in Section 6, which might trigger some new research directions. Finally, we gave our conclusions in Section 7.

2 The Star/Plus Experiment

The star/plus experiment was designed to engage several different cortical areas, in order to look at their interaction. The Regions of interst(ROIs) plausibly involved are known from several other studies. In this experiment, the subject fi rst sees a sentence(semantic stimulus) for 4 seconds, such as "The plus sign is above on the star sign.", then a blank screen for 4 seconds, and fi nally a picture(symbol stimulus) such as

$$\begin{array}{c} + \\ \hline * \end{array}$$

for another 4 seconds, during which the subject must press a button for "yes" or "no", depending on whether the setence matches the picture seen or not. Snapshots were made every 1/2 seconds. The subject is instructed to rehearse the sentence in his/her brain until the picture is presented

rather than try to visualize the sentence immediately. The second variant, for which we also used in this report, switches the presentations of sentences and pictures, and the instruction is to keep the picture in mind until the presentation of the sentence. In the later part of this report, we will note them as $SP\ dataset$ and $PS\ dataset$, respectively.

This task engages several brain areas, such as visual cortex for reading and seeing the sentence, Broca's area for language processing, the Intra Parietal Sulcus for spatial visualization, motor cortex for pressing the button, etc. The exact way in which the areas coordinate varies across subjects, based on their mental and verbal abilities and also on the strategy followed(e.g. mentally rehearse the sentence until you see the picture Vs. making up a picture as you read the sentence and then match it with the picture being displayed). The subject were instructed to try to rehearse the sentence and delay making a mental picture of its semantic content until the picture was presented. Similarily, in the picture first variant, the subjects should try not to express the picture as a sentence while they wait for the sentence presentation.

The subject does this affirmative task 10 times where the sentence and picture are consistent, a similar task where the sentence contains a negative description 10 times, and a control task where he or she looks at a fixed point in the screen. The order in which these repetitions are performed is randomized and each repetition is called a $trial$. Not surprisingly, we call these 3 kinds of trials $affirmative$ trial, $negative$ trial and $fixation$ trial in the later part of this report.

Based upon the observations that the activity of an voxel will last about 8-10 seconds after stimulus, in $SP\ dataset$, we divided the whole course of activity of a voxel into two segments: the first 16 snapshots as $Sentence$ segment, and the next 16 snapshots as $Picture$ segment, and vice versa in $PS\ dataset$. Our classifier could predict, with probability, whether a subject is looking at a picture or a sentence, given an unlabeled segment.

3 Methods

In the fMRI field, a very basic difficulty arises from the fact that the data are very high dimensional (an fMRI image contains 10,000-20,000 voxels) and training data is relatively sparse (in many cases we have only a few dozen training examples). Therefore, we also need discover useful abstractions of the fMRI signals to reduce the apparent dimensionality of the learning task. On the other hand, the subject-to-subject variation makes difficult using the data from different subjects uniformly partially because the number of voxels in an ROI will definitely vary across subjects. Normally, two basic abstractions are used: one is the mean, i.e., averaging the activity values of all voxels in an ROI to get an "averaged" voxel, and the other one is the top n active voxels in an ROI under t-test. Through these ways, we can have the same number of voxels in an ROI for different subjects, which means that we will have the same number of features in the training examples from different subjects and on this basis, we can take advantage of the common Machine Learning methods, such as Support Vector Machine, and Naive Gaussian Bayes Classifier, K Nearest Neighbour, Logistic Regression, and so on. In this report, we will provide our results using Naive Gaussian Bayes Classifier based upon the average abstraction.

Another basic difficulty arises also from the subject-to-subject variation. Without surprise, the responses of different subjects to some particular stimuli are same, to some extent, in pattern, but might differ greatly in intensity. Even for a single subject, the response of him/her is not fixed, but usally assumed normal distributed at corresponding time points. Obviously, the data from different subjects are not directly comparable. Now we can apply a common technique in machine learning: Normalization. But for different machine learning methods, we will definitely do normalization a little differently. In this report, for Naive Bayes Classifier, the features in training examples are assumed to be independent, we simply normalized the data of each ROI in each trial of each subject subjects into $[0, 1]$, that is,

$$Y_t = \frac{X_t - min_t X_t}{max_t X_t - min_t X_t} \quad t = 1, ..., 32$$

$$where\ X_t's\ and\ Y_t's\ are\ the\ data\ before\ and\ after\ normalization,\ respectively.$$

but for other classifiers, we have to do some particular process, such as making the data at particular time points have the same intertrial mean and variance for different subjects.

Sanity check is necessary almost for all fMRI data analyses. In both of our datasets, the data from 3 subjects are with extraordinary large or small values(even 100 times greater or smaller than the normal values). A possible reason for this might be related with the scaling factor in the proprecessing step. More possibly, those 3 subjects performed very badly in the Star/Plus experiment. After sanity check, we have totally 13 "good" subjects. We made some selections of subjects according to the "goodness" of subject.

In the Star/Plus experiment, our psychology colleagues thought that 7 ROIs(CALC, LDLPFC, LIPL, LIPS, LOPER, LT, and LTRIA) were most possibly involved in the Star/Plus task. But base upon our experiment results, we found that some ROIs are not selective at all. We made some selections of ROIs by hand according to our observations, too.

How to measure a classifier? We use cross-validation because we only have limited training examples. Two kinds of cross-validations are used in this report. Leave-one-example-out cross-validation is used to roughly measure the power of the classifier, and leave-one-subject-out cross-validation is used to measure whether a classifier trained for some subjects is still useful for a new subject. Normally, the accuracy will be better under leave-one-example-out cross-validation than under leave-one-subject-out cross-validation, because we use more data under leave-one-example-out cross-validation. Since we have equal numbers of $Picture$ examples and $Sentence$ examples, the random guess probability to predict whether a segment is $Picture$ or $Sentence$ is 0.5. If the accuracy of our classifiers were much better than 0.5, our classifiers work!

4 Results Across Subjects

In this section, based on the average abstraction, we provided the accuracies of the Naive Gaussian Bayes Classifier trained across subjects for $SP\ dataset$ and $PS\ dataset$ separately.

Note:

A. 7 ROIs: {'CALC' 'LDLPFC' 'LIPL' 'LIPS' 'LOPER' 'LT' 'LTRIA'}
B. 4 ROIs: {'CALC' 'LIPL' 'LIPS' 'LOPER'}
C. 1 ROI: {'CALC'}
D. 13 subjects: {'04799' '04805' '04820' '04847' '04958' '05005' '05018' '05093' '05131' '05675' '05680' '05695' '05710'}
E. 9 subjects: {'04805' '04820' '04847' '04958' '05093' '05131' '05675' '05680' '05710'}
F. 8 subjects: {'04805' '04820' '04847' '04958' '05093' '05675' '05680' '05710'}
G. 4 subjects: {'04847' '05675' '05680' '05710'}
H. 2 subjects: {'04847' '05710'}
I. Average Accuracy: the mean value of all single subject accuracies.
I. The accuracies in parenthesis are the corresponding accuracies after simply normalizing the first 32 snapshots of each ROI in each trial of each subject into [0,1].

Subject	*7 ROIs*	*4 ROIs*	*1 ROI*
04799	75%(90%)	80%(93%)	42%(45%)
04805	88%(88%)	93%(95%)	82%(90%)
04820	88%(97%)	95%(100%)	90%(90%)
04847	100%(100%)	100%(100%)	100%(100%)
04958	88%(97%)	88%(97%)	93%(93%)
05005	65%(82%)	62%(80%)	70%(80%)
05018	72%(88%)	72%(90%)	60%(68%)
05093	90%(95%)	90%(95%)	93%(90%)
05099	50%(50%)	50%(50%)	50%(50%)
05131	82%(80%)	85%(82%)	88%(88%)
05393	50%(50%)	50%(50%)	72%(75%)
05643	50%(50%)	50%(50%)	50%(50%)
05675	93%(95%)	93%(95%)	97%(95%)
05680	97%(95%)	100%(97%)	97%(95%)
05695	62%(65%)	75%(70%)	53%(60%)
05710	100%(100%)	100%(100%)	100%(100%)

Table 1: Accuracies for single subject in $SP\ dataset$

Table 1 and Table 2 gave the accuaries for single subject in $SP\ dataset$ and $PS\ dataset$ under leave-one-example-out crossvalidation. The Naive Bayes Classifier performed much better for $SP\ dataset$ than for $PS\ dataset$. In virtue of that the segment is naturally a time course, we

Subject	*7 ROIs*	*4 ROIs*	*1 ROI*
04799	62%(68%)	45%(57%)	62%(72%)
04805	78%(80%)	80%(85%)	80%(75%)
04820	75%(88%)	75%(85%)	68%(68%)
04847	85%(95%)	88%(95%)	95%(93%)
04958	68%(95%)	75%(95%)	72%(82%)
05005	57%(68%)	55%(68%)	62%(72%)
05018	50%(72%)	60%(80%)	68%(53%)
05093	72%(80%)	80%(85%)	80%(80%)
05099	50%(50%)	50%(50%)	50%(50%)
05131	75%(88%)	72%(88%)	75%(72%)
05393	50%(50%)	50%(50%)	72%(72%)
05643	50%(50%)	50%(50%)	50%(50%)
05675	70%(80%)	72%(78%)	75%(70%)
05680	55%(72%)	60%(80%)	78%(82%)
05695	60%(65%)	62%(78%)	65%(68%)
05710	78%(90%)	82%(90%)	82%(93%)

Table 2: Accuracies for single subject in $PS\ dataset$

tried to improve the accuracy by trying to capture the trend of it. We tried to use some statistics, such as $Var(X_{t-h}, ..., X_{t+h})$, etc.) instead of using the activity value at a single time point as a feature. We also tried the difference at lag h(i.e., $X_t - X_{t-h}$) and smoothing techniques(i.e., $mean(X_{t-h}, ..., X_{t+h})$). But all these methods made no improvement in accuracy. After checking the actual time courses, we found that even using eyeballs, it is harder to distinguish $Picture$ segment and $Sentence$ segment for $PS\ dataset$ than for $SP\ dataset$. It might be the true reason why the accuracy is lower for $PS\ dataset$.

Selected Subjects	*Average Accuracy*	*Leave-1-example-out*	*Leave-1-subject-out*
13 subjects	85%(90%)	80%(88%)	76%(86%)
9 subjects	92%(94%)	85%(93%)	81%(91%)
8 subjects	93%(96%)	87%(95%)	82%(94%)
4 subjects	98%(98%)	94%(94%)	85%(85%)
2 subjects	100%(100%)	95%(93%)	55%(60%)

Table 3: Accuracies for multiple subjects in $SP\ dataset$ (Using 7 ROIs)

For $SP\ dataset$, the accuracies of the classifiers trained across subjects are provided in Table 3(using 7 ROIs), Table 4(using 4 ROIs) and Table 5(using 1 ROI). For $PS\ dataset$, the accuracies are reported in Table 6(using 13 subjects). We can find that the simple normalization improved the accuracy of the classifier by about 10%. They showed the selection of subjects and ROIs changed the accuracy greatly, too. The effect of normalization is better for $PS\ dataset$ than for $SP\ dataset$, but the accuracy for $PS\ dataset$ is still be lower than for $SP\ dataset$ because of

Selected Subjects	*Average Accuracy*	*Leave-1-example-out*	*Leave-1-subject-out*
13 subjects	87%(92%)	82%(88%)	80%(88%)
9 subjects	93%(96%)	89%(95%)	89%(95%)
8 subjects	95%(97%)	89%(96%)	89%(96%)
4 subjects	98%(98%)	96%(98%)	95%(98%)
2 subjects	100%(100%)	100%(100%)	80%(100%)

Table 4: Accuracies for multiple subjects in $SP\ dataset$ (Using 4 ROIs)

Selected Subjects	*Average Accuracy*	*Leave-1-example-out*	*Leave-1-subject-out*
13 subjects	82%(84%)	84%(84%)	83%(83%)
9 subjects	93%(93%)	93%(94%)	92%(94%)
8 subjects	94%(94%)	94%(96%)	93%(95%)
4 subjects	99%(98%)	97%(98%)	97%(99%)
2 subjects	100%(100%))	100%(100%)	96%(100%)

Table 5: Accuracies for multiple subjects in $SP\ dataset$ (Using 1 ROI)

Selected ROIs	*Average Accuracy*	*Leave-1-example-out*	*Leave-1-subject-out*
7 ROIs	68%(80%)	65%(80%)	64%(80%)
4 ROIs	70%(82%)	68%(81%)	67%(81%)
1 ROI	74%(75%)	73%(78%)	72%(77%)

Table 6: Accuracies for multiple subjects in $PS\ dataset$ (Using 13 subjects)

Selected ROIs	*Leave-1-example-out*	*Leave-1-subject-out*
7 ROIs	70%(71%)	69%(70%)
4 ROIs	74%(75%)	73%(75%)
1 ROI	75%(75%)	75%(75%)

Table 7: Accuracies for multiple subjects in both datasets (Using 13 subjects)

the large difference in single subject accuracies. What is most exciting is that the accuracy under leave-one-subject-out cross-validation is even comparable to the average value of single subject accuracies.

5 Results Across Contexts

The accuracies trained across contexts and subjects are reported in Table 7. Here, the training examples are from different contexts($SP\ dataset$ and $PS\ dataset$), and leave-1-subject-out cross-validation will leave the test subject out from both of $SP\ dataset$ and $PS\ dataset$. Obviously, although the accuracies here are lower than for same context, the accuracy above 70% is still acceptable because of the great difference among contexts. A fact we have to pay attention to is that matching the sentence and picture happened in the second segment no matter which kind of segment it is. This will confuse the classifier since the classifier doesn't know whether a matching process took place in a segment. This might be another reason why the accuracy decreased when we train the classifier across contexts.

6 Interesting Findings

The activity intensity is generally higher in $negative$ trials than in $affirmative$ trials. It is rather reasonalbe, because we can imagine a subject' brain will need more activity to give an answer when he or she met some inconsistence. Based upon this observation, is it possible to do classification between $affirmative$ trials and $negative$ trials?

To some extent, but not generally, the activity intensity is kind of higher in $SP\ dataset$ than in $PS\ dataset$. We conjecture that it is more difficult for a subject to remember a sentence than a picture in a short period, which matches our intuition. Is it reasonable and feasible to train a classifier to detect which context the segments came from?

As mentioned in Section 5, we conjecture that an additional workload in the second segment led a lower accuracy of a classifier. Is there a more exact way to define $Picture$ segment and $Sentence$ segment?

Normalization improved the accuracy of the classifier trained in same context, but was not helpful when across context. Can we get higher accuracy by normalizing the segments instead of the whole time course? In our experiment results, the answer is no. Actually, this way will greatly reduce the difference between $Sentence$ segment and $Picture$ segment, however, that difference is the basis where we can do classification.

In Naive Bayes Classifier, the independency of features are assumed, that is, it doesn't take advantage of the temporal nature of fMRI data. We have reason to believe that some temporal models(such as Hidden Markov Models) will give better accuracy for our learning task. But it is not

easy to extend the temporal models to the field of fMRI data analysis.

The abstraction we mainly used is averaging all voxels in an ROI into an "averaged" voxel. It is too coarse to some extent. There will definitely exist more useful abstractions that not only reduce the apparent dimensionality of our learning task, but take more information as well. Can we extract more useful abstractions automatically using some machine learning methods(such as Artificial Neural Network), even beyond the limitation of ROI?

7 Conclusions

It is feasible to train classifiers to decode an interesting category of cognitive states, but a variety of machine learning research is needed to extend these capabilities, and such classifiers could be used as virtual sensors of hidden cognitive states, which would be of tremendous use for experimental research in cognitive science and in diagnosis of mental processes in patients with brain injuries. Our learning algorithms for training classifiers when used over multiple human subjects are with satisfying accuracy. The accuary of our classifier is still exciting when used across contexts. Our classifier can explicitly represent the scientist's hypotheses, and continuously evaluate the fit of these hypotheses to data gathered over time from multiple experiments, and suggest refinement to these hypotheses resulting in improved fit to these multiple datasets.

8 Acknowledgement

Many thanks to Francisco Pereira, Stefan Niculescu, Vladimir Cherkassky, Joel Welling, and Marcel Just.

9 References

1. Tom Mitchell, and Marcel Just, Scientific Data Mining to Understand Human Brain Function, March 1, 2002

2. Tom Mitchell, etc., Machine Learning of fMRI Virtual Sensors of Cognitive States, September 16, 2002

3. Francisco Pereira, fMRI, the Star/Plus Experiment and Our Toolbox, May 8, 2002

4. William Eddy, etc., The Challenge of Functional Magnetic Resonance Imaging, Journal of Computational and Graphical Statistics, Volume 8, Number 3, Page 545-558

5. Nicole Lazar, etc., Statistical Issues in fMRI for Brain Imaging, International Statistical Review(2001), Volume 69, Number 1, Page 105-127

Contents