Feature Selection for fMRI Classification Across Multiple Human Subjects

Heather Dunlop
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
hdunlop@cmu.edu

1 Introduction

With the use of functional Magnetic Resonance Imaging (fMRI) it is possible to safely and non-invasively observe neural activity in the brain. Three-dimensional images of the brain can be captured every second, producing large quantities of fMRI observations over the course of a session. This data has the potential to decode which parts of the brain are active in performing a particular task.

This paper investigates the use of fMRI data to develop a classifier to identify a subject’s cognitive state during a particular time interval. In particular, data from a set of subjects is used to decode the cognitive state of a new subject not used in the training process. This is a difficult task because each subject may produce different activation for a particular task and each has a different size and shape of brain.

2 Related work

Numerous studies have been done using fMRI data to detect cognitive states. In [1] different classifier training and feature selection methods are explored for developing classifiers for a single subject. Gaussian Naïve Bayes (GNB), Support Vector Machines (SVM) and k Nearest Neighbour (kNN) are used as learning methods. Regions of interest (ROIs), anatomically defined brain regions, are identified in each subject. The feature selection methods explored are selecting the most discriminating voxels, the most active voxels over the whole brain, the most active voxels in a ROI and the mean of the active voxels in each ROI. This study concluded that SVM and GNB consistently outperform kNN. Also, selecting the most active voxels as features outperforms the other feature selection methods.

This work is taken further in [2] to perform the same task but across multiple subjects. A set of subjects is used for training and a new subject is used to test the classifier. Again, GNB, SVM and kNN are used as learning methods. ROI mapping and Talairach co-ordinates are used to produce a representation of the brain across multiple subjects. This study found that averaging the 20 most active voxels in each ROI produced the best results.

Further work on developing classifiers across multiple subjects has been explored in [3]. Cross-subject classifiers are developed by clustering and achieve higher classification accuracies than in [2].
3 Methods

3.1 Learning

In this paper, machine learning methods are used to approximate classification functions that take a sequence of fMRI images from a contiguous time interval and predict the cognitive state. The cognitive state is the set of states to be discriminated.

We are attempting to create classifiers that will perform well when presented with a new subject. Thus, training is done with all but one subject and then that subject is used for testing. The average accuracy is calculated from these tests.

3.2 Region of interest mapping

A useful method identified in [1] for mapping across multiple subjects is the use of anatomically defined regions of the brain known as Regions of Interest (ROIs). Each subject has the same set of ROIs and an expert has assigned each voxel in each subject to one of these ROIs. Only the ROIs CALC, LIPL, LT, LTRIA, LOPER, LIPS and LDLPFC are used in the algorithms in this paper. Figure 1 shows the ROIs for eight slices of the brain for a particular subject. Each color represents a different ROI.

3.3 Voxel selection

The following methods were identified in [1] and [2] as useful in selecting voxels to include when creating a classifier:

- All – all voxels are used.
- Active(n) – the n most active voxels from across the brain are selected.
- ROIActive(n) – the n most active voxels in each ROI are selected

For single subject classifiers, any of the above voxel selection methods can be used. However, for multiple subject classifiers each subject has a different size and shape of brain. Thus, some mapping must be used that results in the same number of features selected for each subject. Active(n) and ROIActive(n) are the only methods of those listed that can be applied to multiple subject classifiers.
3.4 Feature selection

From examining some of the time series plots for voxels, it is apparent that there is a large variation in the characteristics of these plots. Each consists of the activation at 16 points in time – every 0.5 seconds for 8 seconds. Comparing these plots by Euclidean distance may not make sense. However, comparing time series using some of the characteristics of the plot might be useful.

A few time series plots are shown in Figure 2. The average activation is shown in red. Notice that the average magnitude of the signal is different for each and the frequency of oscillation is also vastly different. A feature that captures the average magnitude and the frequency of the signal might prove useful.

![Figure 2: Some Voxel Time Series Plots](image)

The following techniques are used for feature selection:

- Original – no additional feature selection is done; the entire time series is used as a feature.
- TimeAvg – the average voxel activation over the time series is used.
- DomFreq – the dominant frequency for each voxel is used. The dominant frequency is calculated by taking the Fourier transform of the time series and determining the frequency with the largest response.
- TimeDesc – the average voxel activation and the dominant frequency are used as two features for each voxel.

A GNB classifier is created with the features produced from one of the above methods after selecting the voxels using All, Active(n) or ROIActive(n), and labeled by the known cognitive state.

3.5 Clustering

A clustering algorithm was also attempted, but encountered some implementation issues. This algorithm involved accumulating a table of all voxel time series and clustering these to use each cluster mean as a prototype. Each voxel time series would then be compared to the prototypes to find the closest one. A histogram of this distribution over the entire data set or the most active voxels in the brain would then be computed. However, the Matlab implementation of hierarchical could not handle the large amount of data in this problem and k-means did not produce useful clusters. A clustering solution like this might still be worth experimenting with, but this paper does not examine that topic any further.
4 Experiments

Using the methods described in the previous section, classifiers were created and tested using leave-one-out cross validation. The resulting classifier accuracies are examined and a particular case study is presented in this section.

4.1 Case study

In this fMRI study, thirteen subjects performed a sequence of trials. During each trial they were shown a sentence and then a simple picture. In half of the trials the sentence was presented first, followed by the picture (SP). In the other half the picture was presented first, followed by the sentence (PS). This paper only examines the use of the SP data set. We are attempting to distinguish whether a subject is looking at a sentence or a picture, given a particular time interval. In this case, each time interval consists of 16 fMRI images. Since there are only two possible classes (sentence or picture), the expected accuracy from random guessing is 50%. Data for the thirteen human subjects is used in training and testing the classifiers.

4.2 Results

The feature selection methods mentioned in the previous section were tested using the SP data set. Single subject leave-out-one-trial cross validation was performed as well as multiple subject leave-out-one-subject so that two can be compared.

For Active(n), multiples of 40 less than 1400 were tested for n and the best one is reported. For ROIActive(n), multiples of 2 less than 38 were explored.

Figure 3 shows how the best value of n was determined for ROIActive(n) using TimeDesc for multiple subject classification. The highest accuracy across all values of n was determined to occur for an n value of 18.

![Figure 3: Accuracy for Multiple Subject Classifiers using ROIActive(n) and TimeDesc](image)

Results for both single subject and multiple subject classifiers are displayed in Table 1. The value of n shown is the one that produced the greatest accuracy over all those tested.
### Table 1: Classifier Accuracies for SP Data Set

<table>
<thead>
<tr>
<th>VOXEL SELECTION</th>
<th>FEATURE SELECTION</th>
<th>SINGLE SUBJECT</th>
<th>MULTIPLE SUBJECT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>ACCURACY</td>
<td>n</td>
</tr>
<tr>
<td>All</td>
<td>Original</td>
<td>N/A</td>
<td>87.9%</td>
</tr>
<tr>
<td>All</td>
<td>TimeAvg</td>
<td>N/A</td>
<td>88.3%</td>
</tr>
<tr>
<td>All</td>
<td>DomFreq</td>
<td>N/A</td>
<td>53.7%</td>
</tr>
<tr>
<td>All</td>
<td>TimeDesc</td>
<td>N/A</td>
<td>57.1%</td>
</tr>
<tr>
<td>Active(n)</td>
<td>Original</td>
<td>1280</td>
<td>99.5%</td>
</tr>
<tr>
<td>Active(n)</td>
<td>TimeAvg</td>
<td>1120</td>
<td>99.5%</td>
</tr>
<tr>
<td>Active(n)</td>
<td>DomFreq</td>
<td>1240</td>
<td>85.1%</td>
</tr>
<tr>
<td>Active(n)</td>
<td>TimeDesc</td>
<td>960</td>
<td>87.2%</td>
</tr>
<tr>
<td>ROIActive(n)</td>
<td>Original</td>
<td>36</td>
<td>98.4%</td>
</tr>
<tr>
<td>ROIActive(n)</td>
<td>TimeAvg</td>
<td>38</td>
<td>96.0%</td>
</tr>
<tr>
<td>ROIActive(n)</td>
<td>DomFreq</td>
<td>36</td>
<td>79.2%</td>
</tr>
<tr>
<td>ROIActive(n)</td>
<td>TimeDesc</td>
<td>36</td>
<td>84.4%</td>
</tr>
</tbody>
</table>

### 4.3 Analysis

With a multiple subject classifier, the accuracy is expected to be lower than for a single subject classifier as the brain of each subject is very different – both in size and shape, but also in how that person thinks. It may be that slightly different regions of the brain are activated for a particular task for different subjects. However, we also have more test data to train a multiple subject classifier. The results show that single subject classifiers have much higher accuracy than multiple subject classifiers for most methods of feature selection.

One of the advantages of using TimeAvg, DomFreq or TimeDesc for feature selection is that it greatly reduces the amount of data that is used to build a classifier, simplifying the task at hand. Each time series plot consists of the activation at 16 points in time. Thus, TimeAvg and DomFreq reduce the feature set by 16 times and TimeDesc reduces it by 8 times. But this reduction in data is only important if the newly selected features can still be used for accurate classification.

Many of the time series plots have a great deal of oscillation as previously shown in Figure 2. Some of this oscillation might be due to noise instead of signal. Thus, it is possible that using the dominant frequency as a feature captures the noise instead of reducing it, thus reducing the accuracy of the classifier. Yet, for multiple subjects, results for DomFreq and TimeDesc show that useful classifiers are produced with this feature. However, the accuracy is not nearly as good using these features for single subject classifiers compared to other methods.

Another interesting result is that for ROIActive(n) on single subject classifiers using all feature selection methods, the accuracy improves as n gets larger right up to a value of 38. This is shown in Figure 4 for TimeAvg. However, larger values of n cannot be tested because some ROIs in some subjects do not have more than 38 voxels. Thus, it seems that there might still be room for improvement in the accuracy by using a larger number of active of voxels in the ROIs that contain larger numbers of voxels.
It is also apparent that the value of $n$ for $\text{Active}(n)$ and $\text{ROIActive}(n)$ that produces the greatest accuracy is highly dependent on whether we are developing a single subject or multiple subject classifier.

It should be noted that some of the accuracies reported for multiple subject classifiers are quite a bit lower than those reported in [2]. Some reasons for the difference could be the subtleties of normalization and the selection of active voxels.

5 Conclusions

Overall, the methods explored do not produce results significantly better or worse than in [1], [2] and [3].

Many of the methods explored are very successful in developing an accurate classifier for single subjects. However, the accuracy is lower for multiple subject classifiers. Using $\text{Active}(n)$ or $\text{ROIActive}(n)$ without any further feature selection produces the greatest accuracy for both types of classifiers. Averaging over the time series of each voxel, determining the frequency of oscillation of this time series or using both of these as features produce results comparable to each other, but are more successful with $\text{ROIActive}(n)$ than for $\text{Active}(n)$. This suggests that the average of the time plot and its frequency are useful characteristics in developing a classifier, but more information than these simple characteristics is required to create a successful classifier.

The results in this paper are only for the sentence then picture data set. A more complete analysis of the use of the features described should be performed using other data sets – the other sets available in from the same study, as well as other fMRI studies.

References

