Finding the Augmented Neural Pathways to Math Processing: fMRI Pattern Classification of Turner Syndrome Brains

Gary Tang
{garytang} @ {stanford} · {edu}

Abstract

The use of statistical and machine learning methods for multi-voxel pattern analysis (MVPA) has grown rapidly in the past two decades. Today, sophisticated pattern classification techniques applied to functional magnetic resonance images (fMRI) have allowed researchers to reliably discern different mental states with high accuracy. In this work, we perform MVPA using support vector machines (SVM) to study the neural activity associated to math processing in a clinical population with cognitive deficiencies in math. Namely, we studied the effect of training to rehabilitate Turner Syndrome patients and its ability to augment the neural pathways associated of math processing. Applying a modified form of the canonical recursive feature elimination (RFE) algorithm to spatial and spatial-temporal formulations of the data, we successfully classified the two groups with great predictive accuracy. Unfortunately, the resulting classifiers were unable to demonstrate that post-training images were more similar than pre-training images to a healthy individual with any statistical significance.

Introduction

Sufferers of mental or cognitive disorders have traditionally been diagnosed using symptom-based criteria. Even common disorders, such as depression, are not diagnosed in a rigorous scientific manner, but rather through a checklist of known related-symptoms defined in a medical manual. A rigorous and precise measure of cognitive function and activity has long been the holy grail for psychiatrists and neuroscientist. Today, many believe neuro-imaging, such as magnetic resonance imaging (MRI) and computed tomography (CT), to be the prophetic technology. In particular, functional magnetic resonance imaging (fMRI) has emerged as one of the most popular tools to map of neural activity to mental states. These images track the hemodynamic response (HR) of the brain as a correlate to neural activity and its associated mental state. For this project, we are concerned with the set of neural sequences that suggest a clinical brain is performing a math-related task. While it's generally understood how healthy brains perform math, it remains unclear how brains suffering from cognitive deficiencies process math. One such population are individuals with Turner Syndrome. This condition, affecting 1 in 2500 girls, leads to different types of cognitive impairment, including poorer than average math skills.

Problem Statement

Members of the Stanford School of Medicine conducted a study to rehabilitate the mathematical abilities of individuals with Turner Syndrome through a series of training exercises that taught students to mimic the way a healthy individual processes math. Testing before and after the program suggests a statistically significant improvement in math performance, but a univariate analysis of the fMRI data found no identifiable difference between pre-training and post-training brain activity. The overall goal of this project is to determine which, if any, regions of the brain can be used to discriminate between pre-training and post-training images, and by corollary, identify the possibly new, math-related, neuronal pathways that were generated as a result of the training.

Objective

We separate and define the objective into two components. Firstly, we wish find a reliable measure, in this case a classifier, with which we can claim that a difference exists between pre-training and post-training images. We must take steps to ensure that the classifier is well-generalized such that it can predict on persons from whom we have no prior data, what this paper refers to as *disjoint data*, which is acknowledged to be a much harder task[1] than simply predicting on data left out of the training set. The second

part is simply to apply that measure, either directly or indirectly, to determine if the post-training images look more similar to that of healthy individuals. The classifier chosen and the data objects that we are classifying is problem can be formulated in different ways which will be presented in later sections.

Materials

Data

The fMRI data comes from eleven Turner Syndrome patients and five control patients who were imaged while performing a block test. During this test, images were sampled every two seconds over 14 blocks, which span 420s in total. Each block is either a 30s math task or a 30s control task, where a control block follows a math block. Each image is composed of 31608, 4mm voxels. Each Turner Syndrome patient performed two block tests, corresponding to a pre-training date and a post-training data. The two block tests were separated by approximately six months.

Computational Tools

The images were processed by the MATLAB toolbox SPM[5] and a custom extension authored by Professor Fumiko Hoefft. The optimizer CVX[6] was used to solve the SVM optimization problem. It primarily employs a interior point method whose complexity is either constant or weakly proportional to the number of features. This feature is critical to our problem since the data from each image is converted to a single feature space with dimensions on the order $O(10^4)$. All additional programming development was performed inside the MATLAB environment

Approach

For this project we employ the support vector machine (SVM) learning algorithm for its robustness and proven record of effectiveness. More specifically, we employ a linear SVM since neither the author nor the existing literature[?] has found a significant performance gain in the use of nonlinear SVM. And while we mainly use the algorithm in its classical form, we explore different feature definitions, feature selection techniques, and pre-processing strategies that are best suited to handle our particular application.

Challenges and Strategies to Analyzing fMRI

Overfitting

If the ultimate goal for neuroscientists/psychiatrists to use classifiers for diagnosis, they need to have the ability to generalize well. But for scientists who study fMRI, the problem is further compounded by the fact that the training size m is much smaller than the feature space dimension n. An SVM applied under these conditions will tend to overfit the training data, resulting in poor generalization. And while soft-max regularization helps alleviate this issue, many believe it is essential to explicitly remove features, pruning noisy and uninformative features.[1,2] Therefore, in this project we apply a modified form of the canonical recursive feature elimination (RFE) that seeks to further improve the generalization. Details can be found in later in the report.

Data Quality I

While much of the fMRI data is noisy and uninformative, this is consequence *after* the data is in a usable form (e.g. greyscale numbers from an image). That is to say, some factors affecting the quality of the data are outside the control of the statistician or the analyst. The ability of capturing a proper fMRI is a tenuous one. Patients often move, the fMRI machines experience drift, etc. In order to obtain a useful fMRI volume, it must first be motion corrected, repaired, transformed onto a template, and smoothed. Not only will this affect the signal that is ultimately sent to the SVM, but often times the images are simply not salvageable and their inclusion into a training set will adversely affect the performance of the classifier. One example is when the "pre-pre-processing" (before our pre-processing, see Data Quality II) results in a lot of redundant data. In general, the SVM is agnostic to redundant data so long as they are consistent (i.e. the labeling of redundant data is the same). But manual inspection indicated that the redundant data is not consistent. Evidenced by poorer prediction rates in a preliminary study, these datasets were left out when producing the final results. Because we posit that each individual is statistically representative of the population we seek to classify, we did not preclude the use of an individual's data obtained if data obtained at a different time is usable (e.g. if subject A had a poor test 1 but a usable test 2, that data is included in our pooled set of data).

Data Quality II

Once useful a useful training set is selected, the data often remains inundated with noise and other additive signals (e.g. drift). We address this with several strategies filtering/detrending, masking and the use of principal components.

De-trending

Because of the data acquired by the fMRI machine experiences a significant amount of drift, the drift is regressed out by a quadratic regression model. This could also be performed using a high pass filter, but using this particular form of de-trending reduces the possibility of removing any informative modes.

Low Pass Filtering

It is well known within the neuroscience community that fMRI signal is very noisy.[1] The author's own preliminary study indicatesthat classification is much more difficult without the use of a low pass filter.We used a discrete cosine transform (DCT) to decompose the modes and perform the filtering. Because the signal of interest has a theoretical frequency of about 0.03hz, and motivated by a survey of the spectral distribution of our data, we use a 0.1hz cut off threshold. It is important to note that the survey is best performed on voxels that lie inside the region of interest (ROI). Otherwise, a random survey of uninformative nodes can erroneously infer that there is no underlying signal or similarly, there is no discernable difference between noise and signal.

Science-Driven Feature Reduction: Masking

The use of a mask immediately eliminates features that are definitively known to lack information (e.g. grey matter). A common practice, however, is also to mask the regions of the brain that are hypothesized *a priori* to carry no informative information. In this study, we try both approaches: a minimalist mask, and a mask that includes only the ROI. The results, as expected, are strongly dependent on the choice of mask since most, and also in our case, fMRI analyses are performed with very few samples. As most learning and statistical methods rely on the assumption that the data is representative of the population, when applied to small samples algorithms often find a solution that is strongly dependent on the specific training data and unrepresentative of reality, namely, science.

Principal Component Analysis (PCA)

The principal components highlight those features whose variance is high, likely indicating a signal or an informative voxel. This has the effect of further de-noising the data matrix before going it into the SVM algorithm. If filtering is applied, one should be careful to examine the eigenvalues or squared singular values on a scree plot to ensure that the proper number of principal components are taken, or rather, that enough are taken. In this project, we essentially take all principal components, only removing those that are likely the result of numerical round off (e.g. $\lambda_i > 10^{-6}$). It also has the effect of removing redundant data (by virtue of the fact that the maximum number of principal components does not exceed the rank of the matrix in our case) which is good from both a classification point of view (conflicting redundant data) and from a computational point of view since we now are required to process less data while extracting an equivalent amount of information. Applying this within RFE has the effect of subset-matrix-specific de-noising. Furthermore, because RFE is such a computationally intensive procedure, it is even more important to have an efficient algorithm. It is important to remark that PCA is not a feature reduction method; the complete set of original features is retained. PCA is applied at each step within the RFE procedure (i.e. a new PCA is done on each resulting subset)

Data Definition

The problem of classifying post-training and pre-training math processing can be formulated in different ways. This study choose two particular formulations: instantaneous volume classification and time series classification. Traditional classification of fMRI uses single volumes (images) as samples and has strongly demonstrated to be a reliable way to classify differing cognitive states.[4] Alternatively, we can classify pre-training and post-training images via an ensemble of volumes (15 math volumes + 15 control volumes), i.e. a time series. Since each image represents a snapshot of the given activity (i.e. math task or control task) at some point within the 30s interval. One could imagine that the neural activity at the end of the interval looks very different from at the beginning of the interval. Furthermore, we know that the brain signal within a block follows the hemodynamic response, which is a lagging indicator. Therefore, we expect a single "representative image" of the math task or the control task to have a large variance. Alternatively, the collection images, strewn into one 30x31608 feature vector, representing "duty cycle" is

likely to have a lower variance and more amenable to classification. Results for both formulations are presented in this paper. This technique has also been used to classify the time dependent changes in fMRI images in order to capture temporal features.[3] Although we built the implementation to conduct such a study, it was not explored further at time of writing.

Classifier Formulation

The most intuitive form to achieve our objective is to use the pre-training images representing the math task and comparing them to the post-training images representing the math task. The resulting discriminating weight vector (i.e. the separating hyperplane parameter w) would then be used to classify the control images representing math. That is, if the training did indeed augment the neural patterns to be more similar to a healthy individual, we expect the weight vector to classify the control images with the label corresponding to the post-training images. Alternatively, we can classify through a more indirect means. We can classify the control and math images in the pre-training set and obtain a classifier w_{pre}^{mc} and similarly for the post training set w_{post}^{mc} and using those two weight vectors, determine which weight vector can classify the control data (math + task) the best. Results for both formulations can be found later in this paper.

Specification of Our Classification Procedure

We follow the soft-max form of the linear SVM.[?] Specifically, we chose a soft-max regularization parameter $C = 1000$. Although a nonlinear kernel was tried in the study, it did not demonstrate provable performance improvements over the linear form, which is consistent with existing literature in fMRI classification. Furthermore, the discriminating weight vector (i.e. the support hyperplane) resulting from a nonlinear kernel is much hard to interpret as it relates to functional localization. The data is separated into three sets for each time the classification is run. Firstly, we create one set B that includes all the data belong to one (or two) individual(s); this will be our disjoint set and will remain constant through the RFE procedure. We then take all remaining data and call this set A. For each classification, set A is randomly broken into a testing set $A_{testing}$ and a training set $A_{training}$ in a 30/70 split. Cross validation (k-fold, loocv, etc) will be limited to the training set *only* and prediction rates are calculated for each of the three sets.

Generalization Weighted Recursive Feature Elimination

As discussed earlier, the ability to maximize generalization at all steps will be critical to infer on disjoint data. Therefore, we begin with the classical RFE algorithm,[2] whose procedure can be found in detail in the literature, and add modifications to improve the generalization. Firstly, the in our application of the RFE procedure, we remove a *percentage* of the worst ranking features as opposed to a single feature. Naturally this is for computational purposes since we are working with a very large features space. We begin our modification with where we perform our ranking. In the author's experience, the RFE process has a very large variance with respect to the reduced subsets that it creates. Therefore, within the RFE algorithm, we perform a k-fold cross validation. Furthermore, we seek to enhance the generalization of our classifier by weighting the weight vectors v_i (corresponding to the ith step in the cross validation procedure) in the cross validation process by the ability of v_i to predict on the set $A_{testing}$. A similar process can be applied to the disjoint set B but experience shows that this has no effect on the resulting weight vector because the disjoint set B is almost always predicted poorly to the same degree. One possible intepretation is for that particular iteration, you want the separating hyperplane to fit your test data well. Another, possibly better interpretation, is that you are picking separating hyperplanes that generalizes to all the data very well, that is, both training and testing. Since training will always be low, we expect that the generalization can be improved if we weight against he performance of the test set i.e. when we weight it against hte performance on the entire set. This idea isnt limited to fMRI but rather this is helpful whenever you have many more features than samples and overfitting is expected to be a problem. Since we are essentially applying a mean over the cross validation set, it is important to have this weight factor since some particular data sets will be conditioned such that the optimizer is unable to converge.of interest (ROI) mask.

Algorithm

Algorithm *Generalization Weighted RFE*(X, S, R, k, PCA)
Given a dataset X, a starting feature set S, a removed feature set R, the number of folds in cross validation k, and the option to perform PCA within iterations

1. Begin with set $S = [allfeatures]$ and $R = []$
2. Do PCA if desired
3. Randomly generate $A_{training}$, $A_{testing}$

4. **for** i=1:k Perform k-fold or loocv cross validation Perform SVM Obtain weight vector w_i
5. Now weight obtain a weight-averaged weight vector $\bar{w} = \frac{\sum_i^k (1-\epsilon_i)}{\sum_i^k \epsilon_i}$ where ϵ_i is the testing error associated to weight vector w_i
6. Perform ranking on $\bar{w}$
7. Update S, and R

Results

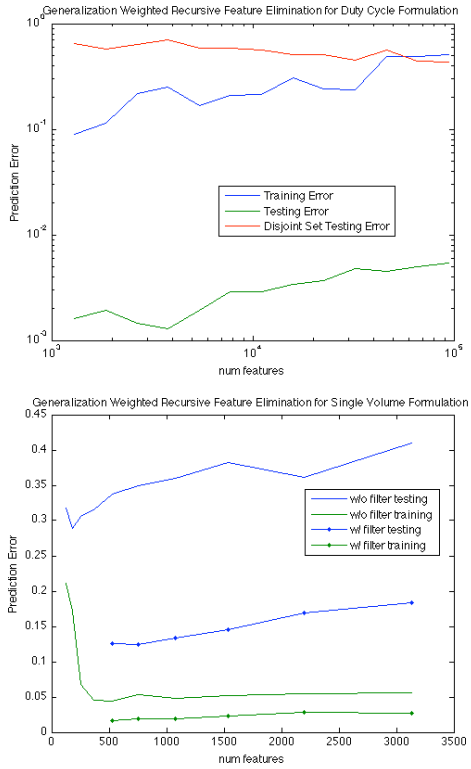

Figure 1: Prediction Rates For Single Volume and Duty Cycle Formulations.

Prediction Rates

Overall, we can confidently classify pre-training and post training data with a great deal of accuracy, with prediction rates averaging 87%. We can also make other observations from the figures. The top figure shows that the disjoint set B does not change as we proceed in the RFE process, which leads us to believe that generalizing outside an *individual's* data remains very difficult with the number of samples we have for the study. The bottom figure shows the effect of filtering on the prediction rate and demonstrates that filtering is a key part to getting quality data.

Comparing Against Healthy Brains

Data Set	Prediction Control Data
Pre, MathvCT, single	0.48
Post,MathvCT, single	0.45
Pre,ROI,MathvCT, single	0.523
Post,ROI,MathvCT, single	0.4913
Pre,ROI,Filter,MathvCT,single	0.5
Post,ROI,Filter,MathvCT,single	0.503
DutyCycle	0.5238 (CT=Post)
DutyCycle, ROI	0.5714 (CT=Post)
Pre Math v Post Math	0.487 (CT=Post)

From the table above, we can see that we have a very difficult time classifying the control data, as we'd expect from our experience trying to classify the disjoint set. Despite our attempts to maximize the generalization, we were unable to discern from the data whether or not the Turner Syndrome brains were more similar to the healthy brains after training.

Conclusion

The goal for our project was to determine if we could discern pre-training and post-training fMRI volumes and if we could, whether or not we could claim that the post-training volumes look more similar to the pre-training volumes. Although we can firmly state that there is indeed a difference between pre-training and post-training data, our attempts to generalize those results to data that were *disjoint* from our training were ultimately unsuccessful. Analysis of fMRI data is strongly hindered by the lack of data and until more data can be acquired, the holy grail of diagnosing mental states with fMRI classifiers remains just that.

References

[1] Norman, A. Kenneth, et. al. "Beyond Mind Reading: multi-voxel pattern analysis of fMRI data." *TRENDS in Cognitive Science.* doi:10.1016/j.tics.2006.07.005.

[2] Guyon, I., Vapnik, V. et al. "Gene Selection for Cancer Classification using Support Vector Machines." *Machine Learning*, Kluwer Academic Publishers, 2002, p389-442.

[3] Mourao-Miranda, Janaina, Friston, Karl J., and Michael Brammer. "Dynamic Discrimination Analysis: A spatial-temporal SVM." *NeuroImage* doi:10.1016/j.neuroimage.2007.02.020.

[4] Cox, David and Robert Savoy. "Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex." *NeuroImage.* doi:10.1016/S1053-8119(03)00049-1.

[5] Wellcome Trust Center for Neuroimaging. http://www.fil.ion.ucl.ac.uk/spm/software/spm5/, Statistical Parametric Mapping. UCL Institute of Neurology. Dec. 2005.

[6] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming (web page and software). http://stanford.edu/ boyd/cvx, June 2009.