Feature Selection on Raw and Wavelet-Transformed fMRI Data

Indrayana Rustandi

Computer Science Department, Carnegie Mellon University
email: indra@cs.cmu.edu

Introduction

For the classification of cognitive states in fMRI data, [1] found that feature/voxel selection can improve classification performance. They discussed several voxel selection methods, namely, selecting the n most discriminating voxels (Discrim), selecting the n most active voxels (Active), and selecting the n most active voxels in each region of interest (roiActive). In general, they found that the Active method works best on most of the datasets.

We present the results of applying three new feature/voxel selection methods on fMRI data from one subject in a category-discrimination study. Because of its touted ability to deal with nonstationarities, which are arguably present in the fMRI data, we also obtained a wavelet-transformed version of the data and ran one of the methods to select relevant wavelet coefficients.

Feature Selection

Relevant features aid in classification, but on the other hand, irrelevant features can hurt classification performance, and with more features we might need more training examples to get good performance. Therefore, choosing relevant features are essential in doing classification. This is true especially in the task of classifying fMRI where the features (in the ten thousands if we consider voxels as features) outnumber the examples (the number of trials in a typical fMRI experiment) by several orders of magnitude.

All the methods presented here are *filter* methods, namely they choose the features based on the characteristics of the data without taking into account the classification method used.

Information Gain This method was proposed as part of a feature selection scheme for microarray data [3]. Initially, a K-mixture Gaussian Mixture Model is learned using the EM algorithm. Then the voxel values are discretized based on their posterior probabilities being in any of the mixtures. This step basically partitions each voxel F_i into K components $E_1, \cdots, E_K$. We then calculate the information gain of F_i with respect to the partitions $S_1, \cdots, S_C$ induced by the C classes:

$$I_g = H(P(S_1), \cdots, P(S_C)) - \sum_{k=1}^{K} P(E_k) H(P(S_1|E_k), \cdots, P(S_C|E_k))$$

where H is the entropy function and P is the probability function. The n voxels with the highest information gain are then chosen.

Fisher's Class Separability or Fisher Criterion (Mean) This and the following method were inspired by [2]. For each voxel v_i, we calculate its Fisher's class separability:

$$\frac{\sum_{c=1}^{C} \pi_c (\mathrm{mean}_i(v_i) - mean_c(mean_i(v_i)))^2}{\sum_{c=1}^{C} \pi_c \mathrm{var}_i(v_i)},$$

where C is the number of classes, π_c is the empirical proportion of class c, $\mathrm{mean}_i(\cdot)$ and $\mathrm{var}_i(\cdot)$ are the mean and variance of values of voxel v_i, and $\mathrm{mean}_i(\cdot)$ is the mean over class c.

Fisher's Class Separability or Fisher Criterion (Median) This is similar to the version above, but using median instead of mean, and median absolute deviation instead of variance:

$$\frac{\sum_{c=1}^{C} \pi_c |\mathrm{med}_i(v_i) - med_c(med_i(v_i))|}{\sum_{c=1}^{C} \pi_c \mathrm{mad}_i(v_i)}.$$

One reason to use this method instead of the mean version is because it is more robust, i.e. it is more resistant to outliers.

Experiment

7 words each from categories tools and dwellings were presented to the subjects in 6 epochs. In each epoch, the words were presented in a random order. Each word was presented for 3 seconds, with an interstimulus interval of 7 to 8 seconds. For each word presentation, the subjects had to determine the category of the word by pressing one of two buttons. There is a fixation period at the end of each epoch. Brain images were acquired every 1 second (TR=1,000ms). There are 16 64x64 slices in each image.

Methods

Before the analysis, the data was subjected to time/slice correction, motion correction, filtering, detrending, and scaling preprocessing steps. Then we obtained a representative image for each trial by averaging images from timepoints 4 to 7 in the respective trial. Then we normalized the voxel values such that the mean of all voxels becomes 0 and the standard deviation becomes 1.

In the case of the wavelet-transformed data, we obtained the wavelet coefficients for each trial using the 3D discrete wavelet transform, applying the *db2* wavelet, over the averaged normalized image for the respective trial.

As a classifier, we used the *SVMlight* implementation of the support vector machine (SVM). We did a leave-one-pair-out cross-validation on the data, where the pair consists of one tool word and one dwelling word. The classification task is to classify whether a given image belongs to either the tools or the dwellings category. So it is a two-class classification. Random guessing will yield an accuracy of 0.50.

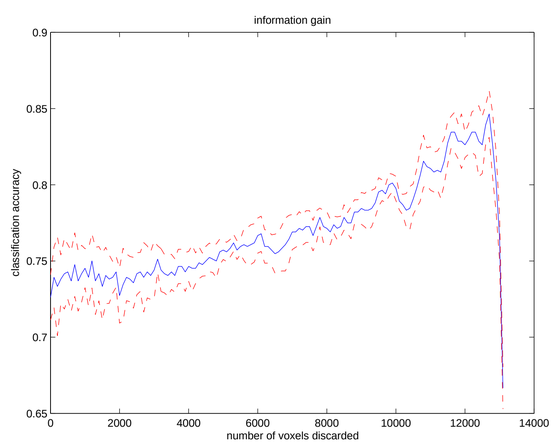

Figure 1: Information gain: classification accuracy vs number of voxels discarded. The blue line is the mean accuracy, and the red lines are standard deviations of the accuracies.

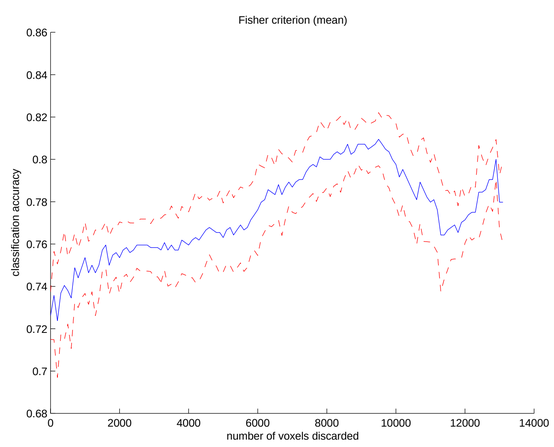

Figure 2: Fisher criterion (mean): classification accuracy vs number of voxels discarded.

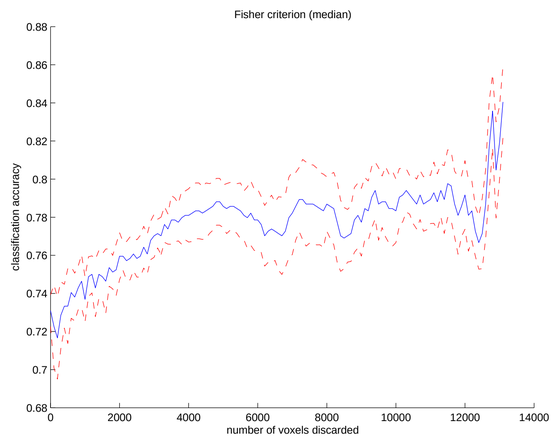

Figure 3: Fisher criterion (median): classification accuracy vs number of voxels discarded.

Results

Raw fMRI Data We wanted to know whether choosing a subset of the voxels using the three methods above improves classification performance. Another question is what's the optimal n when we have to choose n voxels for classification. For this purpose, we iterated all three methods over several different ns to get an idea on which n the peak classification performance is.

Figure 1 shows the classification accuracy vs the number of voxels *discarded* (total number of voxels minus the number of voxels *retained*). It shows an increasing trend of accuracy as the number of voxels retained decreases, and reaching its peak at slightly below 0.85 when around 500 voxels are retained.

The same results for both kinds of the Fisher criterios are shown in Figures 2 and 3, for the mean and the median versions respectively. Both figures show similar increasing trends in classification accuracy as the number of voxels retained decreases. However, for the mean version of the Fisher criterion, the peak is when around 3700 voxels are retained, with accuracy of around 0.81. On the other hand, for the median version, the peak is reached when about only 100 voxels are retained with accuracy of around 0.84.

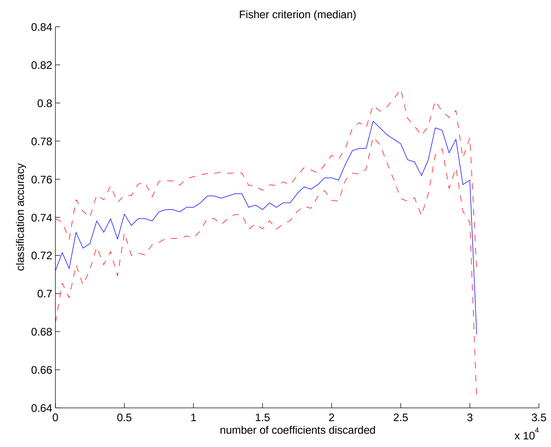

Figure 4: Fisher criterion (median) on wavelet coefficients: classification accuracy vs number of coefficients discarded.

Wavelet-Transformed fMRI Data Figure 4 shows the results when applying the median version of the Fisher criterion on the wavelet-transformed data. Because of the boundary of the image, we obtain around 30000 wavelet coefficients compared to around 13000 voxels originally. As the number of coefficients retained decreases, a similar increasing trend of classification accuracy is also seen. The peak is reached when around 7500 coefficients are retained, with accuracy of about 0.79.

Discussion

Although this is an ongoing study and only the data from one subject was analyzed, we can already see the effect of how the three methods presented help improving classification accuracy. However, one problem that the three methods share is that we still have to guess or try all different possibilities to find out how many voxels give the highest classification accuracy; all three methods cannot provide us this information.

At this stage, the results on the wavelet-transformed data do not justify transforming the fMRI data into the wavelet domain. However, there are a lot of parameters when doing the wavelet transform that still need to be investigated. A couple of examples would be choosing the type of wavelet and the number of vanishing moments.

References

[1] Tom M. Mitchell et al. Learning to decode cognitive states from brain images. *Machine Learning*, 57:145–175, 2004.

[2] Naoki Saito and Ronald R. Coifman. Local discriminant bases and their applications. *Journal of Mathematical Imaging and Vision*, 5:337–358, 1995.

[3] Eric Xing, Michael I. Jordan, and Richard M. Karp. Feature selection for high-dimensional genomic microarray data. In *ICML*, 2001.

Carnegie Mellon

CNBC Center for the Neural Basis of Cognition