VOXEL SELECTION IN FMRI DATA ANALYSIS: A SPARSE REPRESENTATION METHOD

Yuanqing Li, Zhuliang Yu

School of Automation
Southchina University of Technology
Guangzhou, 510640, China

Praneeth Namburi[†]*, Cuntai Guan*[♯]

[†]School of Electrical and Electronic Eng.
Nanyang Technological University
Singapore 639689
[♯]Institute for Infocomm Research
Singapore 119613

ABSTRACT

This paper proposes an iterative sparse representation-based algorithm for voxel selection in functional magnetic resonance imaging (fMRI) data. The output of the algorithm is a sparse weight vector, of which the magnitude of each entry represents the significance of its corresponding voxel with respect to mental tasks or stimulus. To demonstrate the validity of our algorithm and illustrate its application, we apply this algorithm to the Pittsburgh Brain Activity Interpretation Competition (PBAIC) 2007 fMRI data set for selecting the voxels which are the most relevant to the *tasks* of the subjects. Compared with three baseline methods, general linear model (GLM)-based statistical parametric mapping (SPM), correlation method and mutual information method, our method shows satisfactory performance for voxel selection.

Index Terms— Functional magnetic resonance imaging (fMRI), voxel selection, sparse representation, statistical parametric mapping (SPM), prediction.

1. INTRODUCTION

In functional magnetic resonance imaging (fMRI), an fMRI scanner measures the blood-oxygenation-level dependent (BOLD) signal at all points in a three dimensional grid of the brain. The cells within this three-dimensional grid are known as voxels. A typical fMRI data set is composed of the time series (BOLD signals) of tens of thousand voxels. Therefore, voxel selection plays an important role in fMRI data analysis because of: (i) heavy computation burden; (ii) uncorrelation (or redundancy) of a large number of voxel time series with respect to the stimulus/task presented to the subject. One class of voxel selection methods are based on statistical test/statistics that find brain regions with statistically significant response. A typical example is statistical parametric mapping (SPM) based on general linear model (GLM) [1]. The second class of voxel selection methods (e.g. Correlation method) are based on the correlation between the voxel time series and the time series of the task or stimulus [2, 3]. In this paper, we present a new sparse representation based method for voxel selection in fMRI data.

The sparse representation of signals can be modeled by

$$\mathbf{y} = \mathbf{A}\mathbf{w}, \tag{1}$$

where $\mathbf{y} \in R^N$ is a given signal vector, $\mathbf{A} \in R^{N\times M}$ is a basis matrix, $N < M$. When the model (1) is used for fMRI data analysis, $\mathbf{A}$ is a data matrix of which each column is a time series of a voxel, $\mathbf{y}$ is a transformed stimulus/task function which is obtained by convolving a stimulus/task function with a hemodynamical response function.

The task of sparse representation is to find a solution $\mathbf{w} \in R^M$ of (1) such that this solution is as sparse as possible. In many references such as [5, 6], a sparse solution is found by solving the following optimization problem.

$$\min ||\mathbf{w}||_1, \;\; s.\,t. \;\; \mathbf{A}\mathbf{w} = \mathbf{y}, \tag{2}$$

where 1-norm $||\mathbf{w}||_1$ is defined as $\sum_{i=1}^{M} |w_i|$. Problem (2) can be converted into a standard linear programming problem. Recently, it has been found that Model (2) has applications in feature selection and detection [7].

Two related methods are 1-norm support vector machine (sparse SVM) [8] and Lasso regularization [9], which have potential applications in feature selection including dimension reduction. Compared with 1-norm support vector machine and Lasso method, (2) has computational advantage especially when the number of the variables is extremely large.

GLM model is common used in fMRI data analysis. The sparse representation model (1) can bee seen as the opposite of the GLM Model, however there exists significant difference between the two models. In the sparse representation model, a lot of voxels are considered simultaneously, but

Correspondence to: Dr. Yuanqing Li, E-mail: auyqli@scut.edu.cn.
Yuanqing Li's work was supported by National Natural Science Foundation of China under Grants 60475004 and 60825306. Zhuliang Yu's work was supported by National Natural Science Foundation of China under Grant 60802068.

the task/stimulus conditions are considered separately. Conversely, voxels are considered separately, but the task/stimulus conditions are considered simultaneously in GLM model.

In this paper, we design an algorithm based on the linear programming problem (2) for voxel selection in fMRI data analysis. To demonstrate the effectiveness of this method, it was applied to the fMRI data set of Pittsburgh Brain Activity Interpretation Competitions (PBAIC) 2007. After voxel selection with our method, we perform the prediction of experience based cognitive tasks from the fMRI data set of PBAIC 2007 as in [4]. The prediction results will be used in evaluation of our method. In our data analysis, we also compare our method with three baseline methods, GLM-SPM method, correlation method and mutual information method.

2. MATERIALS AND METHODS

2.1. Algorithm

In this section, we present our algorithm based on (2) for voxel selection. We do not directly use (2) mainly considering the following aspect: When N is not sufficiently large, $\mathbf{w}$ obtained by single optimization may not reflect the important features well. Even if N is sufficiently large, this problem still exists because of noise. Note that in this paper, $\mathbf{A}$ is an fMRI data matrix of which each column is a time series of a voxel and each row contains the data of one volume (or part of one volume), $\mathbf{y}$ is the convolution of a stimulus/task function and a hemodynamical response function. The following algorithm is designed to detect the parts in the rows of $\mathbf{A}$ (voxels) relevant to $\mathbf{y}$.

Algorithm 1:

Step 1: For $k = 1, \cdots$, do the following Steps 1.1 to 1.4.

Step 1.1: Randomly choose L rows from $\{\mathbf{a}_1, \cdots, \mathbf{a}_N\}$ to construct a L by M matrix denoted as $\mathbf{A}_k$, the corresponding L entries of $\mathbf{y}$ form a column vector denoted as $\mathbf{y}_k \in R^L$.

Step 1.2: Solve the following optimization problem (similar to (2)),

$$\min ||\mathbf{w}||_1, \ s.t., \ \mathbf{A}_k\mathbf{w} = \mathbf{y}_k. \tag{3}$$

The optimal solution of (3) is denoted by $\bar{\mathbf{w}}^{(k)}$.

Step 1.3: Let

$$\mathbf{w}^{(k)} = \frac{1}{k}\sum_{i=1}^{k} \bar{\mathbf{w}}^{(i)}. \tag{4}$$

Step 1.4: If $k > 1$ and $d(k) = ||\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)}||_2 < \alpha$, where α is a predefined small positive constant, set $\mathbf{w} = \mathbf{w}^{(k)}$ and go to Step 2. Otherwise go to Step 1.1.

Step 2: For a given positive θ, define $R = \{j| \ |w_j| > \theta, j = 1, \cdots, M\}$. Then R is our detected part of interest in all rows of $\mathbf{A}$.

In Algorithm 1, there are three parameters L, α and θ to be set in advance. We first set L much smaller than N and M such that the optimization problem (3) can be quickly solved (e.g. $L = 0.2N$). Since Algorithm 1 is convergent (the proof is omitted here), we can easily set a small α (e.g. $\alpha < 0.01$) to obtain a stable $\mathbf{w}$. θ can be chosen in different ways depending on the application. Here we present a probability method. Considering the entries of $\mathbf{w}$ are sparse, we assume that the probability distribution of the entries of $\mathbf{w}$ is Laplacian. Using all entries of $\mathbf{w}$ as samples, we estimate the mean, the variance and the inverse cumulative distribution function F^{-1} of this Lapcian distribution. We then define $R = \{i| \ |w_i| > \theta, i = 1, \cdots, M\}$, where θ is chosen as $F^{-1}(p_0)$, p_0 is a given probability (e.g. 0.975 in this paper). As will be shown in section 3, this method for determining θ is acceptable.

Note that noise is not explicitly reflected in (3). However, the weight vector $\mathbf{w}$ generally is affected by noise. Through the average operation in (4), the effect of noise can be reduced. Considering that each $\bar{\mathbf{w}}^{(i)}$ is a regression coefficient vector between training data matrix $\mathbf{A}_i$ and $\mathbf{y}_i$, thus $\mathbf{w}^{(k)}$ reflects the connection between training data matrix $\mathbf{A}$ and $\mathbf{y}$. Therefore, the magnitude of $\mathbf{w}$ represents the significance of the ith column of $\mathbf{A}$ with respect to $\mathbf{y}$. Furthermore, our simulations and data analysis results show that a large fraction of entries of $\mathbf{w}^{(k)}$ are close to zero, i. e. $\mathbf{w}^{(k)}$ is still sparse. Based on the sparsity of $\mathbf{w}$, the voxels which are the most correlated to the stimulus/task function can be selected.

2.2. Voxel selection in functional MRI data

In this section, we apply Algorithm 1 to the fMRI data of PBAIC 2007 [10] for voxel selection. The fMRI data was collected by Siemens 3T Allegra scanner with imaging parameters TR and TE being 1.75s and 25ms respectively. Three subjects' data were available in the competition. Each subject's data consists of three *runs*. Each run consists of 500 volumes of fMRI data of which each volume contain $64 \times 64 \times 34$ voxels (voxel size: $3.2 \times 3.28 \times 3.5\, mm^3$). The preprocessed data provided by the competition are used in this paper, in which the feature data was preprocessed by convolving the raw feature vectors with the double gamma hemodynamic response filter (HRF) produced by the SPM. Through a mask preprocessing, the total number of voxels in the brain is ≈ 32000. Thus the fMRI data for each run is represented by a matrix consisting of around 32000 columns (voxels) and 500 rows (time points). When the scans were obtained, the subject was performing several tasks (e.g. listen to instructions, pick up fruits) in a virtual reality (VR) world. The *ratings* for these tasks were computed and form the ground truth. Only the tasks for the first two runs were distributed at www.braincompetition.org. Therefore, here we use only data from the first two runs for analysis. We present detailed results mainly for four tasks: (i) The *Hits* task, times when sub-

ject correctly picked up fruit or weapon or took picture of a pierced person; (ii) the *Instructions* task, which represents the task of listening to instructions from a cell phone in the virtual world; (iii) the *Faces* task, times when subject looked at faces of a pierced or unpierced person; (iv) the *Velocity* task, times when subject was moving but not interacting with an object. For more detailed description of the data, refer to [10]. The goal of the competition was to predict the task functions of the third run using the fMRI data.

Using the Instructions task as an example, we now describe our data analysis method. The preprocessed fMRI data downloaded from the competition website (www.braincompetition.org) is first filtered temporally by the filter $\frac{1}{4}[1, 2, 1]$. Next, the data is smoothed spatially by convoluting each volume with a 3 voxel $\times$ 3 voxel $\times$ 3 voxel gaussian kernel. We then perform 2-fold cross-validation as follows. In the first fold, we use Run 1 data to calculate the Pearson correlation between the time series of each voxel and the transformed Instructions task function. The voxels with high absolute value of this correlation are chosen to form a set of voxels, $\mathcal{N}$. Then, our algorithm is used for a second selection of voxels to obtain $R \subset \mathcal{N}$. In Algorithm 1, $\mathbf{A} \in R^{500\times|\mathcal{N}|}$, of which each column is a time series of a voxel in $\mathcal{N}$, $\mathbf{y} \in R^{500}$ is a transformed task function. The parameters in Algorithm 1 are set as follows. The number of iterations is fixed to 600, L is set to 25 and θ can be chosen as described in section II-A. Ridge regression is used on the time series of voxels $\in R$ to predict the transformed Instructions task function of Run 2. Prediction accuracy is measured as the Pearson correlation between the actual transformed task and the predicted task. In the second fold, we use Run 2 data for training and predict the transformed Instructions task function for Run 1. For the purpose of comparison, we use three baseline methods, GLM-SPM method, Pearson correlation and mutual information, to replace our method for selection of voxels and perform the 2-fold cross-validation as described above.

3. RESULTS AND DISCUSSIONS

In this section, we demonstrate the validity of our algorithm for voxel selection by analyzing the prediction accuracy. We compare Algorithm 1 with GLM-SPM method, Pearson correlation method and mutual information method for voxel selection.

Prediction. First, we compare the ability of each method in choosing the *most relevant* voxel. In Table 1, we present the prediction accuracies (averaged over two folds) for $N_R = 1$ for the transformed instruction task. Hereafter, N_R denotes the number of voxels of R, the set of selected voxels. From table 1, we can see that the voxel selected by Algorithm 1 is the most correlated to the transformed task function across runs. For the other 3 tasks, similar conclusion was be obtained.

Furthermore, we test if Algorithm 1 is consistently better than GLM-SPM method, Pearson correlation and mutual information based methods for voxel selection. Let $\mathbf{b} = [1, 2, 4, \cdots, 300]$. For each i ($i = 1, \cdots, 151$), we set $N_R = b_i$ (the number of selected voxels), then predict the four transformed task functions for all subjects and average the results over two folds of cross validation. Fig. 1 shows the plots of average prediction accuracy with respect to $\mathbf{b}$ for the four methods and three subjects and four tasks. From this figure, the average prediction accuracy using the voxels selected by Algorithm 1 is consistently superior to those of Pearson correlation and mutual information based methods. In several cases e.g. shown in the subplot in the first row and the second column of Fig. 1, the performance of Algorithm 1 is comparative to that of GLM-SPM method; while in the other cases e.g. shown in the subplot in the second row and the third column, the performance of Algorithm 1 is significantly better than that of GLM-SPM method.

Table 1. Average Prediction accuracy over two folds obtained with one voxel and Instructions task for various subjects and methods

	Sub 1	Sub 2	Sub 3
Algorithm 1	0.3227	0.5518	0.5337
GLM-SPM	0.2233	0.3620	0.1417
Correlation	-0.0503	0.1213	0.1520
Mutual Information	-0.0209	-0.0450	-0.0396

We also analyze the effectiveness of choosing R using θ as described in Section II-A. Example, for instruction task, the number of voxels in R obtained using Algorithm 1, N_{th} are: *Subject 1,* 22 (fold 1), 29 (fold 2). The corresponding prediction accuracy (averaged over two folds) for Subject 1 are 0.8151. The prediction accuracy (averaged over two folds) for the four tasks and three subjects are marked with a '*' in Fig. 1. Even though N_{th} does not correspond to the best prediction accuracy, it can lead to a satisfactory result.

Localization. Now we analyze the effectiveness of Algorithm 1 in localization from a biological perspective. Each of the four tasks evaluated here can be related to activity in specific region(s) of the brain. For example, the hits and velocity events are expected to be correlated with activity in the motor cortex, especially the part for planning actions, which is the supplementary motor cortex. Instructions task is expected to be correlated with activity in language processing area, the auditory cortex. Now we choose a representative case, which is for task 2 (Instructions), run 2 and Subject 3, to show our results of localization. For the other cases, we also have biologically reasonable conclusion. For Instructions task, 25 voxels are selected using Algorithm 1 with their distribution shown in Fig. 2. We can see that most of these voxels are in the *appropriate* areas of the brain. For instance, Voxels from the auditory cortex are shown in Slices 14, 16, 19, 20 etc.

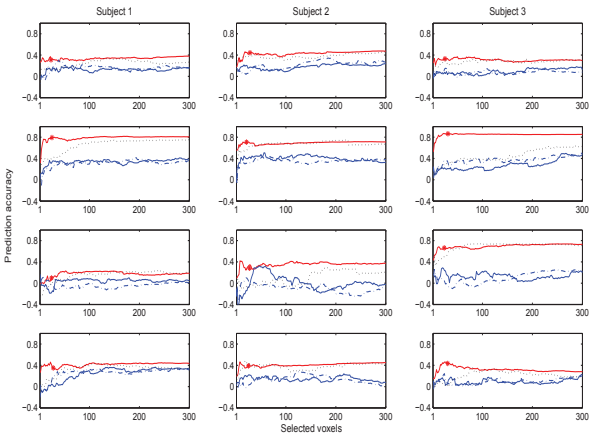

Fig. 1. Prediction accuracy curves obtained by four methods. In each subplot, red solid line: Alg. 1; black dotted line: GLM-SPM method; blue solid line: Correlation method; blue dash-doted line: Mutual information method. The four rows correspond to four tasks (Hits, Instructions, Faces, Velocity) respectively.

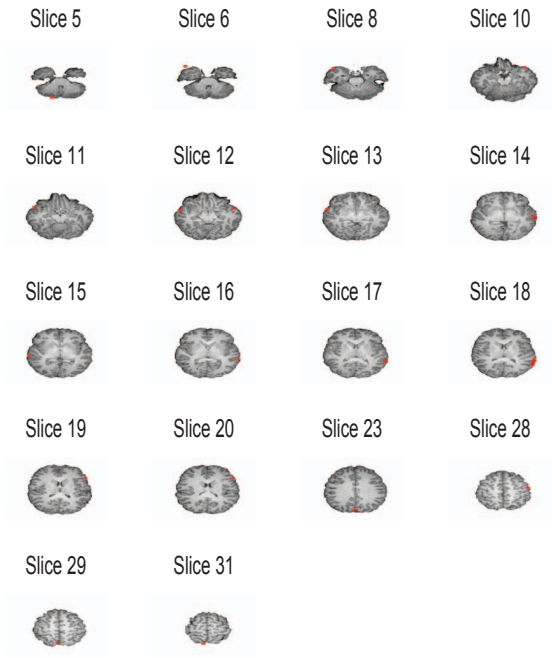

Fig. 2. Distribution of 25 selected Voxels (highlighted in red) corresponding to the first 25 highest weights calculated by Algorithm 1 for Instructions task, Run 2 and Subject 3. Slices are numbered from inferior to the superior parts of the brain.

4. CONCLUSIONS

In this paper, we presented an iterative detection algorithm based on sparse representation. This algorithm may be used for feature selection, localization, novelty detection, etc. Here, we presented one application for voxel (feature) selection in fMRI data analysis. The validity of our method was shown through the comparison with three baseline methods, GLM-SPM method, Pearson correlation method and mutual information method, in our data analysis.

Acknowledgement: The authors are grateful to the organizers of Pittsburgh Brain Activity Interpretation Competition 2007 for providing the fMRI data.

5. REFERENCES

[1] K. J. Friston, A. P. Holmes, K. Worsley, J. B. Poline, et al, "Statical parameter maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, pp. 189-210, 1995.

[2] D. D. Cox, R. L. Savoy, "Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage* 19, 261-270, 2003.

[3] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning* 57, 145-175, 2004.

[4] F. Meyer, G. J. Stephens, "Locality and low-dimensions in the prediction of natural experience from fMRI," *Proc. of the Twenty First Annual Conference on Neural Information Processing Systems*, Canada, 2007.

[5] D. L. Donoho, & M. Elad, "Maximal sparsity representation via l_1 minimization," *the Proc. Nat. Aca. Sci.* vol. 100, pp. 2197-2202, 2003.

[6] Y. Li, A. Cichocki and S. Amari, "Sparse representation and blind source separation," *Neural Computation*, vol. 16, no. 6, 2004.

[7] E. Kidron, Y.Y. Schechner, M. Elad, "Cross-modal localization via sparsity," *IEEE Transactions on Signal Processing*, Vol. 55, no. 4, pp. 1390 - 1404, 2007.

[8] J. Bi, P. Bennett, M. Embrechts, C. M. Breneman, M. Song, "Dimensionality reduction via sparse support vector machines," *Journal of Machine Learning Research*, vol. 3, pp. 1229-1243, 2003.

[9] R. Tibshirani, R., "Regression selection and shrinkage via the LASSO," *Journal of the Royal Statistical Society, Series B (Methodological) 58(1), 267-288, 1996.*

[10] Schneider, W., Siegle, G. Pittsburgh Brain Activity Interpretation Competition 2007 Guide Book: Interpreting subject-driven actions and sensory experience in a rigorously characterized virtual world. http://www.braincompetition.org