Modeling Spatiotemporal Structure in fMRI Brain Decoding Using Generalized Sparse Classifiers

Bernard Ng
Biomedical Signal and Image Computing Lab
The University of British Columbia
Vancouver, Canada
bernardyng@gmail.com

Rafeef Abugharbieh
Biomedical Signal and Image Computing Lab
The University of British Columbia
Vancouver, Canada
rafeef@ece.ubc.ca

Abstract—**The curse of dimensionality constitutes a major challenge to functional magnetic resonance imaging (fMRI) classification. Coupled with the typically strong noise in fMRI data, prediction accuracy is often limited. In this paper, we propose exploiting the inherent spatiotemporal structure of brain activity to regularize the typically ill-conditioned fMRI classification problem. To impose a spatiotemporal prior, we employ a recent classifier learning formulation for building Generalized Sparse Classifiers (GSC). This formulation combines a generalized ridge term with the LASSO penalty, which is integrated into classifier learning to permit various general properties, such as spatial smoothness, to be modeled. Here, we exploit this flexibility of GSC to build a spatiotemporally-regularized sparse linear discriminant classifier, and contrast its performance on real fMRI data against a number of state-of-the-art classification techniques. Our results show that incorporating a spatiotemporal prior jointly improves prediction accuracy and result interpretation, which demonstrate the added value of such prior information in fMRI spatiotemporal classification.**

Keywords-brain decoding, fMRI, neuroimaging, sparse regularization, spatiotemporal classification

I. INTRODUCTION

Functional magnetic resonance imaging (fMRI) provides a non-invasive means for studying human brain function. The standard fMRI analysis approach examines each brain voxel in isolation [1], which completely ignores the inherent spatiotemporal structure of fMRI data. To remedy this limitation, pattern classification techniques have been explored [2,3], which enable all voxels to be jointly analyzed. Under the typical fMRI classification framework, signal intensity of each voxel is treated as a feature with brain volumes taken as samples [2,3]. Alternatively, one may concatenate multiple brain volumes within a trial of the same experimental condition and treat each concatenated set of brain volumes as a sample [4,5]. In either case, the classification task is to determine the experimental condition to which the brain volumes belong. However, the number of features (tens of thousands of voxels) typically far exceeds the number of samples (hundreds of brain volumes). Thus, direct application of standard classifiers, such as linear discriminant analysis (LDA) [6] and support vector machines [5], will likely result in overfitting [2,3].

The conventional approach to reduce feature dimension is to apply univariate analysis to select voxels that display significant activation or discriminant power [5]. However, this approach neglects the collective information encoded by the voxel patterns [7]. Recently, sparsity-enforcing techniques have become the dominant means for dimension reduction [7-13]. The key advantage of these techniques lies upon how the collective information of all voxels is jointly considered during feature selection. Also, enforcing sparsity provides a more clear-cut segregation of the relevant voxels (i.e. weights of irrelevant voxels are shrunk to exactly zero). However, merely encouraging sparsity without accounting for the inherent structure in the data may result in suboptimal predictive performance and spurious weight patterns [7].

To account for the intrinsic structure in fMRI data, group sparse techniques have been explored to jointly select spatially proximal voxels [11]. However, imposing group sparsity alone does not encourage spatially smooth weight patterns, which deviates from how spatially neighboring voxels tend to display similar level of brain activity [14]. Elastic net has also been employed for jointly selecting correlated voxels, but suffers from similar limitations [12]. More recently, Van Gerven et al. [13] proposed a Bayesian formulation for incorporating a spatiotemporal prior, where the authors opted to model uncertainty by estimating the posterior probabilities of the classifier weights as opposed to obtaining sparse classifier weights through finding the maximum a posterior solution.

In this paper, we focus on spatiotemporal classification, where multiple brain volumes within a trial are treated as a sample. To regularize this ill-conditioned problem (i.e. limited number of samples coupled with severe noise), we propose to exploit the spatiotemporal structure of fMRI data. In particular, we employ our recently proposed formulation for building Generalized Sparse Classifiers (GSC) [12] to incorporate a spatiotemporal prior into classifier learning. Brain activity is intrinsically a spatiotemporal process [4] and tends to be sparsely distributed in localized clusters [7]. These properties of brain activity can be jointly captured by simultaneously enforcing sparsity and spatiotemporal smoothness using GSC. We explore the implications, both spatially and temporally, in explicitly modeling the spatiotemporal structure of fMRI data in classifier learning in this work.

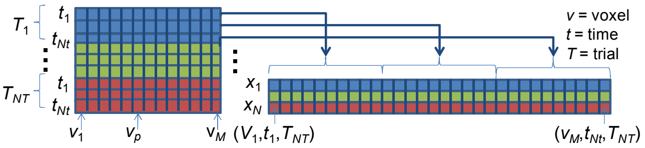

Fig. 1. Predictor matrix. Brain volumes within the same trial are concatenated and taken a single sample x_i. N_t is number of volumes within a trial and N_T is the number of trials for a particular experimental condition.

II. Methods

A. Problem Formulation

Given N M×1 feature vectors, x_i, forming the rows of an N×M predictor matrix, X, the goal of pattern classification is to find the corresponding N×1 label vector, l, containing the class label l_i of x_i. In context of spatiotemporal fMRI classification, we treat the signal intensity of each brain voxel p at time t_k within a trial as a feature, and all brain volumes within a trial of the same experimental condition as a sample x_i (Fig. 1). In contrast to spatial fMRI classification where each brain volume is taken as a sample, multiple brain volumes are concatenated into a single sample under the spatiotemporal setting. Our goal is thus to determine to which experimental condition l_i does each concatenated set of brain volumes x_i belong. Since brain activity is inherently a spatiotemporal process [4], exploiting the spatiotemporal structure of fMRI data during classifier learning can be beneficial. We employ the GSC formulation for integrating such prior information as presented next.

B. Generalized Sparse Classifier

To integrate application-specific properties beyond mere sparsity into classifier learning, we [12] proposed combining a generalized ridge term with the least absolute shrinkage and selection operator (LASSO) penalty [15]:

$$J(a) = \alpha\|\Gamma a\|_2^2 + \beta\,\|a\|_1, \tag{1}$$

where a is a vector containing the classifier weights, Γ is a matrix for modeling the associations between features, and α and β control the amount of regularization. The power of (1) stems from how the generalized ridge term enables penalization of differences in weights between associated features, which presumably should be similar to reflect their associations, and has been previously proposed for penalizing differences in successive weights of ordered features in the context of regression under the name, “Smooth LASSO” [16], and for modeling correlations between adjacent pixels in image analysis applications. Combining (1) with the typical misclassification loss results in the GSC formulation [12]:

$$\hat{a} = \min_a \|l_g - f(Xa)\|_0 + \alpha\|\Gamma a\|_2^2 + \beta\|a\|_1, \tag{2}$$

where $f(\cdot)$ maps Xa onto the label space, and l_g is a vector containing the ground truth labels. To efficiently minimize (2), we employ a two-step optimization strategy, where l_g are transformed into continuous variables to facilitate direct application of sparse regression solvers [12,17,18]:

Step 1. Learn the constraint-free optimal projection of the training data X using, e.g. graph embedding (GE) [19]:

$$Wy = \lambda Dy, \tag{3}$$

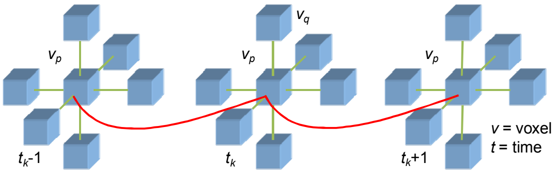

Fig. 2. Spatiotemporal neighborhood. 6-connected voxels and signal value of voxel p itself at adjacent time points are defined as neighbors of voxel p.

where y is the projection of X in the subspace of W [19]. W_{ij} models the intrinsic relationships between samples i and j, and D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. We note that the main advantage of GE is that it enables various subspace learning algorithms to be used as classifiers by simply varying W [19].

Step 2. Determine the classifier weights a such that y and Xa are as similar as possible under the desired constraints:

$$\hat{a} = \min_a \|y - Xa\|_2^2 + \alpha\|\Gamma a\|_2^2 + \beta\|a\|_1 . \tag{4}$$

The two-step strategy hence converts the classification problem (2) into a regularized regression problem (4). If we further transform (4) by augmenting X and y:

$$\tilde{X} = (1+\alpha)^{-\frac{1}{2}}\begin{pmatrix} X \\ \sqrt{\alpha}\Gamma \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \tag{5}$$

we obtain exactly LASSO regression problem [15]:

$$\hat{a} = \sqrt{1+\alpha}\,\min_{\tilde{a}} \|\tilde{y} - \tilde{X}\tilde{a}\|_2^2 + \beta\|\tilde{a}\|_1 , \tag{6}$$

which is a well-studied problem with a wealth of efficient solutions [20,21]. We used the spectral projected gradient technique due to its ability to handle large-scale problems (~40,000 features in our study, Section III) [21]. Parameters α and β were selected using nested cross-validation [5].

To investigate the implications in incorporating a spatiotemporal prior, we build a spatiotemporally smooth sparse LDA (STSLDA) classifier by first solving for y in (3) with:

$$W_{ij} = \begin{cases} 1/m_c, & l_i = l_j = c \\ 0, & otherwise \end{cases}, \tag{7}$$

where m_c is the number of samples in class c [18]. We then optimize (4) with Γ being a spatiotemporal Laplacian operator:

$$\Gamma_{p_k q_s} = \begin{cases} -1, q_s \in N_{p_k} \\ 0, otherwise \end{cases}, \quad \Gamma_{p_k p_k} = -\sum_{q_s \neq p_k} \Gamma_{p_k q_s}, \tag{8}$$

where N_{pk} is the “spatiotemporal” neighborhood of voxel p at time t_k as depicted in Fig. 2.

III. Materials

The publicly available StarPlus database [22] was used for validation. We provide here a brief description of the data and the experiment for convenience. Details can be found in [5,22].

In the StarPlus experiment, all subjects performed 40 trials of a sentence/picture matching task. In each trial, subjects were required to look at a picture (or sentence) followed by a sentence (or picture), and to decide whether the sentence (picture) correctly described the picture (sentence). The first stimulus was presented for 4 s followed by a blank screen for 4 s. The second stimulus was then presented for up to 4 s followed by a 15 s rest period. In half of the trials, the picture preceded the sentence, and vice versa.

fMRI brain volumes were acquired from 13 normal subjects at a TR of 500 ms [5] with 6 of the subjects' data made available online [22]. Each subject's dataset comprised voxel time courses within 25 ROIs that were chosen by neuroscience experts, resulting in approximately 5000 voxels per subject. Inter-subject differences in the number of voxels were due to anatomical variability. Motion-correction and temporal detrending were applied on the voxel time courses to account for head motions and low frequency signal drifts [5]. No spatial normalization was performed.

IV. Results and Discussion

In this study, we treated the intensity of each voxel at each time point within a trial as a feature, and all brain volumes within the same trial of each experimental condition as a sample. The classification task was thus to discriminate sets of brain volumes associated with a picture from those associated with a sentence. To account for the delay in the hemodynamic response function (HRF) [1], we only used the 8 brain volumes collected 4 s after stimulus onset within each trial. Each sample thus consisted of approximately 40,000 features (i.e. roughly 5000 voxels × 8 volumes, Section III) with 40 samples per class (i.e. 40 trials per class) for each subject. For comparison, we contrasted STSLDA against LDA [6], linear SVM [5], SLDA, ENLDA, and SSLDA [12]. Ten-fold cross validation was used to estimate prediction accuracy [5].

A summary of the quantitative and qualitative results are shown in Fig. 3 and Fig. 4, respectively. Axial view of the classifier spatial weight patterns corresponding to 6 s after stimulus onset (i.e. when HRF typically peaks) was plotted. Only 1 exemplar subject was included in Fig. 4 due to space limitation. To study the temporal dynamics of the classifier weights, we reshaped the $N_t \cdot M \times 1$ classifier weight vector back into a $N_t \times M$ matrix and applied principal component analysis (PCA). The resulting principal temporal mode is displayed adjacent to its associated classifier spatial weight pattern in Fig. 4. The displayed time window corresponds to 5 s to 8 s after stimulus onset (Section III). The canonical HRF [1] (within the same time window) was overlaid for comparison.

Using LDA resulted in the worst average predictive performance over subjects, which was likely due to overfitting. Also, the resulting classifier spatial weight patterns appeared randomly-distributed. Using SVM obtained similar accuracy level, and the spatial weight patterns again seemed random. Reducing overfitting by enforcing sparsity using SLDA improved accuracy. However, the spatial weight patterns appeared overly-sparse, which deviates from how brain activity is distributed in localized spatial clusters [7]. Using ENLDA further increased prediction accuracy, but the spatial weight patterns were still overly-sparse with few local clusters. Explicitly modeling spatial correlations using SSLDA resulted in higher accuracy and spatially smoother weight patterns, thus demonstrating the value of incorporating prior knowledge in fMRI classification, both in terms of predictive performance and result interpretability. These benefits of exploiting prior knowledge is further exemplified using our proposed STSLDA. By incorporating a spatiotemporal prior, STSLDA achieved the best overall predictive performance with an average accuracy of 90%. Also, STSLDA obtained the lowest variability in prediction accuracy across subjects, thus demonstrating stable performance albeit the considerable inter-subject variability often seen in fMRI studies [14]. Moreover, STSLDA provided spatially smooth weight patterns that conform well to prior neuroscience knowledge. Specifically, classifier weights were correctly assigned to voxels within the visual cortex, temporal lobe, and inferior temporal lobe, which are known to be involved in picture/sentence discrimination [23].

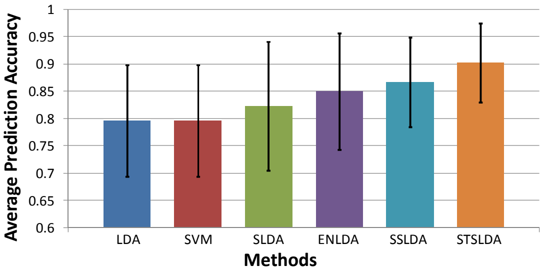

Fig. 3. Prediction accuracy comparisons. STSLDA outperformed all other contrasted methods with an average accuracy of 90%.

Temporally, STSLDA provided smoother temporal profiles than its sparsity-enforcing counterparts. The overall shape of STSLDA's temporal profiles also moderately resembles that of the canonical HRF. Our results thus suggest that analyzing classifier weights in the manner described above may provide an alternative means for HRF estimation. However, further validation with more data is required to confirm this finding.

V. Conclusions

We proposed integrating a spatiotemporal prior into classifier learning for handling the curse of dimensionality and the strong noise often seen in fMRI data. Exploiting GSC, we built a spatiotemporally-regularized sparse LDA classifier, which improved prediction accuracy over the widely-used LDA and linear SVM classifiers as well as a modest number of sparse LDA variants. More neurologically sensible weights, both spatially and temporally, were also obtained. Our results thus demonstrate the value of modeling data structure in fMRI spatiotemporal classification.

References

[1] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C.D. Frith, and R.S.J. Frackowiak, "Statistical Parametric Maps in Functional Imaging: A General Linear Approach," Hum. Brain Mapp., vol. 2, pp. 189-210, 1995.

[2] K.A. Norman, S.M. Polyn, G.J. Detre, and J.V. Haxby, "Beyond Mindreading: Multi-voxel Pattern Analysis of fMRI Data," Trends Cogn. Sci., vol. 10, pp. 424-430, 2006.

[3] J.D. Haynes and G. Rees, "Decoding Mental States from Brain Activity in Humans," Nat Rev. Neurosci., vol. 7, 523-534, 2006.

[4] J.M. Miranda, K.J. Friston, and M. Brammer, "Dynamic Discrimination Analysis: A Spatial-temporal SVM," NeuroImage, vol. 36, pp. 88-99, 2007.

[5] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to Decode Cognitive States from Brain Images," Mach. Learn., vol. 57, pp.145-175, 2004.

[6] J.D. Haynes and G. Rees, “Predicting the Orientation of Invisible Stimuli from Activity in Human Primary Visual Cortex,” Nat. Neurosci., vol. 8, pp. 686-691, 2005.

[7] M.K. Carroll, G.A. Cecchi, I. Rish, R. Garg, and A.R. Rao, “Prediction and Interpretation of Distributed Neural Activity with Sparse Models,” NeuroImage, vol. 44, pp. 112-122, 2009.

[8] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, “Sparse Estimation Automatically Selects Voxels Relevant for the Decoding of fMRI Activity Patterns,” NeuroImage, vol. 42, pp. 1414-1429, 2008.

[9] S. Ryali, K. Supekar, D.A. Abrams, and V. Menon, “Sparse Logistic Regression for Whole-brain Classification of fMRI Data,” NeuroImage, vol. 51, pp. 752-764, 2010.

[10] V. Michel, E. Eger, C. Keribin, and B. Thirion, “Multi-Class Sparse Bayesian Regression for Neuroimaging Data Analysism” in MICCAI Workshop on Mach. Learn. Med. Imaging, vol. 6357, pp. 50-57, 2010.

[11] M. Van Gerven, A. Takashima, and T. Heskes, “Selecting and Identifying Regions of Interest Using Groupwise Regularization,” in NIPS Workshop on New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis, 2008.

[12] B. Ng, A. Vahdat, G. Hamarneh, and R. Abugharbieh, “Generalized Sparse Classifiers for Decoding Cognitive States in fMRI,” in Proc. MICCAI Workshop on Mach. Learn. Med. Imaging, vol. 6357, pp. 108-115, 2010.

[13] M. Van Gerven, B. Cseke, F.P. de Lange, and T. Heskes, “Efficient Bayesian Multivariate fMRI Analysis Using a Sparsifying Spatio-temporal Prior,” NeuroImage, vol. 50, pp. 150-161, 2010.

[14] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.B. Poline, “Dealing with the Shortcomings of Spatial Normalization: Multi-subject Parcellation of fMRI Datasets,” Hum. Brain Mapp., vol. 27, pp. 678-693, 2006.

[15] R. Tibshirani, “Regression Shrinkage and Selection via the LASSO,” J. Royal Stat. Soc. Series B, vol. 58, pp. 267-288, 1996.

[16] M. Hebiri, “Regularization with the Smooth-Lasso Procedure,” preprint, 2008.

[17] L. Grosenick, S. M. Greer, and B. Knutson, “Interpretable Classifiers for fMRI Improve Prediction of Purchases,” IEEE Trans. Neural Sys Rehab. Engin., vol. 16, pp. 539–548, 2008.

[18] D. Cai, X. He, and J. Han, “Spectral Regression: A Unified Approach for Sparse Subspace Learning,” in Proc. IEEE Int. Conf. Data Mining, pp. 73-82, 2007.

[19] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, and S. Lin. Graph Embedding and Extension: A General Framework for Dimensionality Reduction. IEEE Trans. Pat. Ana. Machine Intell., vol. 29, pp. 40-50, 2007.

[20] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, “Least Angle Regression,” Ann. Stat., vol. 32, pp. 407-499, 2004.

[21] E. van den Berg and M.P. Friedlander, “Probing the Pareto Frontier for Basis Pursuit Solutions,” SIAM J. Sci. Comput. vol. 31, pp. 890-912, 2008.

[22] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/.

[23] R. Vandenberghe, C. Price, R. Wise, O. Josephs, and R.S.J. Frackowiak, “Functional Anatomy of a Common Semantic System for Words and Pictures,” Nature, vol. 383, pp. 254-256, 1996.

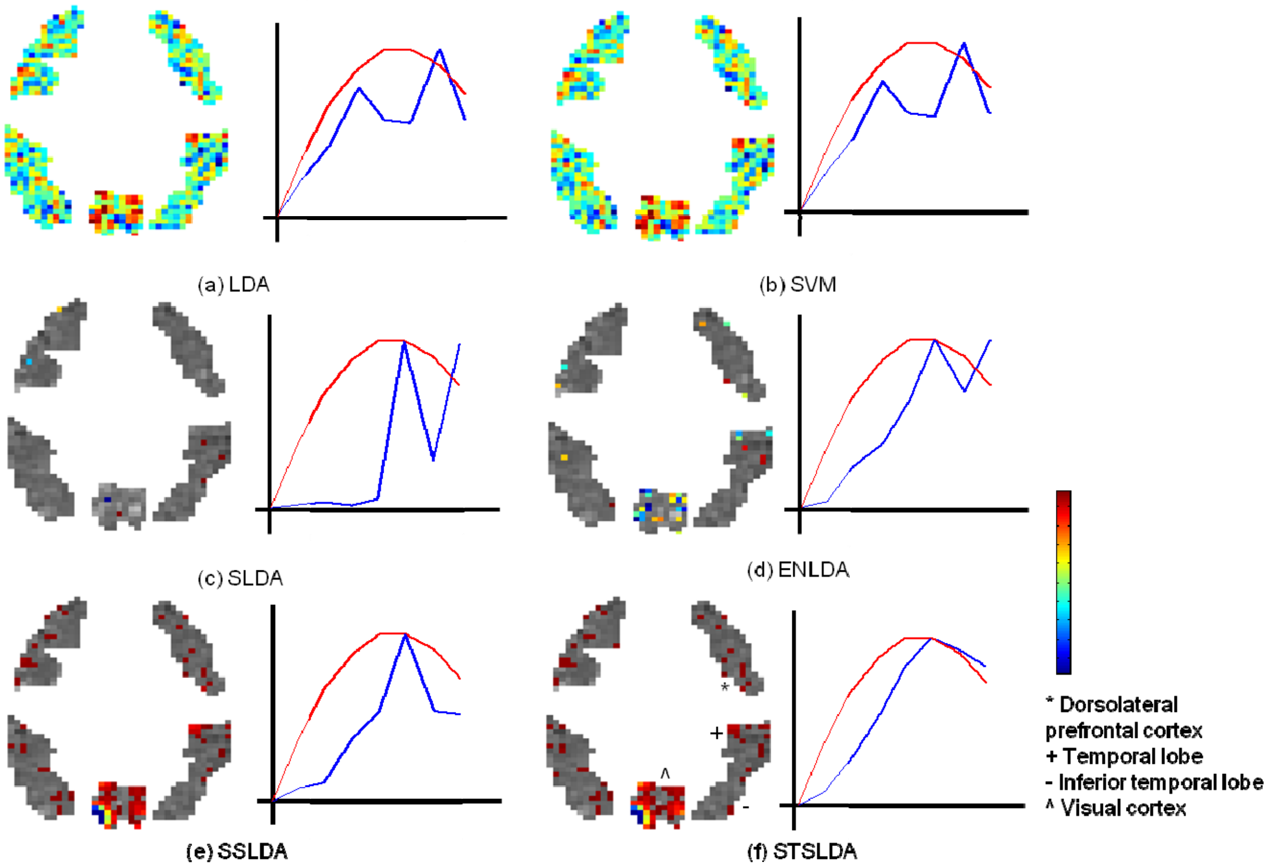

Fig. 4. Classifier weights of contrasted methods. Spatial weight patterns 6 s after stimulus onset are displayed. Red indicates positive weights and blue indicates negative weights. Adjacent to the spatial weight patterns are temporal profiles of the weights (blue curve) extracted using PCA (Section IV) and the canonical HRF overlaid (red curve). The displayed time window corresponds to 5 s to 8 s after stimulus onset. (a) LDA and (b) linear SVM resulted in randomly-distributed spatial weight patterns. (c) SLDA generated overly-sparse spatial weight patterns, which was only mildly improved with (d) EN-LDA. (e) SSLDA and (f) STSLDA produced spatially smooth patterns, which better reflect how spatially proximal voxels tend to display similar level of brain activity. STSLDA also generated smoother temporal profiles than its sparse LDA counterparts with moderate resemblance to the shape of the canonical HRF.