Generalized Sparse Classifiers for Decoding Cognitive States in fMRI

Bernard Ng[1], Arash Vahdat[2], Ghassan Hamarneh[3], Rafeef Abugharbieh[1]

[1]Biomedical Signal and Image Computing Lab, The University of British Columbia
[2]Vision and Media Lab, Simon Fraser University
[3]Medical Image Analysis Lab, Simon Fraser University
bernardn@ece.ubc.ca

Abstract. The high dimensionality of functional magnetic resonance imaging (fMRI) data presents major challenges to fMRI pattern classification. Directly applying standard classifiers often results in overfitting, which limits the generalizability of the results. In this paper, we propose a new group of classifiers, "Generalized Sparse Classifiers" (GSC), to alleviate this overfitting problem. GSC draws upon the recognition that numerous standard classifiers can be reformulated under a regression framework, which enables state-of-the-art regularization techniques, e.g. elastic net, to be directly employed. Building on this regularized regression framework, we exploit an extension of elastic net that permits general properties, such as spatial smoothness, to be integrated. GSC thus facilitates simultaneous sparse feature selection and classification, while providing greater flexibility in the choice of penalties. We validate on real fMRI data and demonstrate how explicitly modeling spatial correlations inherent in brain activity using GSC can provide superior predictive performance and interpretability over standard classifiers.

Keywords: elastic net, fMRI, sparse classifiers, spectral regression

1 Introduction

The application of pattern classification techniques for analyzing brain activity has attracted the attention of the functional magnetic resonance imaging (fMRI) community in recent years [1,2]. Departing from the standard univariate approach [3], pattern classification methods exploit the activity distribution of the entire brain to discriminate different cognitive states. The power of this whole-brain classification approach stems from the observation that even voxels with weak individual responses may carry important cognitive information when analyzed jointly [4]. However, the high dimensionality of fMRI data and the interpretability of the classification weights remain as major challenges to this class of approaches [1,2].

Under a pattern classification framework, the fMRI signal intensity (or summary statistic [3]) at each voxel is usually taken as a feature (variable), with each brain volume (or each subject) treated as a sample (observation) [5,6]. Since typical fMRI datasets consist of considerably more voxels (~tens of thousands) than brain volumes (~hundreds) and subjects (~tens), direct application of standard classifiers, such as

linear discriminant analysis (LDA) [7] or support vector machines (SVM) [5,8,9] where all the brain voxels are used as features, will likely result in overfitting [1,2]. To reduce the dimensionality of the feature vector, a common strategy is to restrict the feature set to only those voxels displaying significant activation or discriminant power [4-6,9]. Alternatively, principal component analysis (PCA) can be applied prior to classification [10]. However, neither of these strategies considers the collective discriminant information encoded by the voxel patterns, and thus may result in suboptimal feature selection [11-13].

Recently, powerful methods that simultaneously select discriminant voxels and estimate their weights for classification have been proposed [11-13]. These methods extend traditional classifiers by incorporating sparse regularization, which controls overfitting by encouraging zero weights to be assigned to irrelevant voxels. However, naively enforcing sparsity may lead to spatially-spurious classification weight patterns (e.g. weights assigned to isolated voxels), which limits interpretability and hence defeats the ultimate objective of fMRI studies [13]. To relax this highly-overlooked limitation, methods that include an additional ridge penalty to promote joint selection of correlated voxels have been proposed [12,13]. This refinement seems to produce less spatially-scattered weight patterns [13]. However, it is unclear whether indirectly modeling voxel correlations by means of a ridge penalty is sufficient for fully capturing the spatial correlations inherent in brain activity and hence jointly selects sparse sets of spatially-contiguous clusters as features. This important issue is investigated in this work.

In this paper, we propose a new group of classifiers, "Generalized Sparse Classifiers" (GSC), that permits more general penalties, such as spatial smoothness in addition to sparsity, to be seamlessly integrated. GSC constructs upon the realization that numerous standard classifiers can be reformulated and trained under a regression framework [14,15], which enables direct deployment of standard regularization techniques, such as least absolute shrinkage and selection operator (LASSO) and elastic net [16]. Building on this regression framework, we employ an extension of elastic net that facilitates higher flexibility in the choice of penalties. The implications of explicitly modeling spatial correlations in brain activity using GSC are explored.

2 Proposed Method

2.1 Problem Formulation

Given N M-dimensional feature vectors, x_i, forming the columns of a predictor matrix, X, our goal is to find the corresponding $N\times1$ response vector, l, containing the class label of x_i. In the context of fMRI, the feature vector usually comprises either signal intensities [5] or summary statistics [6] of M image voxels, and the N samples are either the brain volumes [5] or the subjects drawn from different populations [6]. The problem of fMRI classification can thus be posed as that of subspace learning for finding a mapping that well separates feature vectors of different classes.

Many algorithms, such as LDA, PCA, isomap, laplacian eigenmap, locally linear embedding, neighborhood preserving embedding, and locality preserving projection, have been proposed for subspace learning [17]. Despite differences in motivation, all

of these algorithms can in fact be unified under a graph embedding framework [17]. Specifically, if we let each voxel be a graph vertex with W_{ij} being the edge weights representing the degree of similarity between voxels i and j, all of the aforementioned algorithms can be reformulated into the following optimization problem [15]:

$$\max_{y} y^T W y \quad s.t. \quad y^T D y = 1, \tag{1}$$

where y_i is the projection of x_i onto the subspace defined by W, and D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. Varying W results in different algorithms [17] with the optimal y determined by solving the following generalized eigenvalue problem:

$$Wy = \lambda D y . \tag{2}$$

For ease of interpretation [1], we restrict our attention to linear classifiers, i.e. $y = X^T a$, so that the relative contribution of each voxel i can be directly discerned from a_i. However, naive estimation of a by substituting $y = X^T a$ into (1) and solving the corresponding eigenvalue problem $XWX^T a = \lambda XDX^T a$ [15] will likely result in overfitting due to the large number of voxels compared to the number of brain volumes and subjects [15]. To control overfitting, a popular strategy is to enforce sparsity on a [16]. In particular, in their seminal paper [14], Zou et al. proposed transforming PCA into a regression problem, where techniques, such as LASSO and elastic net, can be exploited. This regression approach provides an efficient means for obtaining sparse PCA, which has been successfully applied to a multitude of large-scale problems, such as gene expression analysis with tens of thousands of features [14]. To generalize beyond PCA, Cai et al. extended this approach to graph embedding under the name "spectral regression" [15], which we adopt in this paper.

2.2 Spectral Regression

Spectral regression decomposes classifier learning into two steps [15]: (i) Solve the eigenvalue problem (2) to find y. (ii) Find a such that $y = X^T a$. However, such a may not exist. Thus, one may have to relax the equality [15]:

$$\hat{a} = \arg\min_{a} \left(\left\| y - X^T a \right\|_2^2 + J(a) \right), \tag{3}$$

where $J(a)$ is a penalty for controlling overfitting. A widely-used $J(a)$ is the LASSO penalty, $\|a\|_1$, which shrinks a_i of irrelevant features to exactly zero [16]. The solution of the resulting problem can be efficiently computed using least angle regression (LARS) [16]. However, LASSO has two main drawbacks [16]. First, the number of non-zero a_i cannot exceed the number of samples. Second, for groups of mutually correlated features, LASSO tends to arbitrarily select only one feature within each group. To alleviate these limitations, Zou et al. proposed the elastic net approach [16]:

$$\hat{a} = \arg\min_{a} \left(\left\| y - X^T a \right\|_2^2 + \alpha \|a\|_2^2 + \beta \|a\|_1 \right), \tag{4}$$

where α and β control the amount of regularization. By augmenting X and y as below, (4) can be transformed into a LASSO problem [16]:

$$\hat{a} = \sqrt{1+\alpha}\,\arg\min_{a^*}\left(\left\|y^* - X^{*T}a^*\right\|_2^2 + \frac{\beta}{\sqrt{1+\alpha}}\left\|a^*\right\|_1\right), \quad (5)$$

$$X^* = (1+\alpha)^{-\frac{1}{2}}\begin{pmatrix} X \\ \sqrt{\alpha}I \end{pmatrix}, \quad y^* = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad (6)$$

where I is a $M \times M$ identity matrix. Since rank(X^*) = M, elastic net can potentially select all M features [16]. Zou et al. also showed that adding the ridge penalty, $\|a\|_2^2$, promotes sparse sets of correlated features to be jointly selected [16]. Moreover, (5) can be efficiently solved using LARS. Thus, elastic net enjoys the same advantages as LASSO, while relaxing LASSO's limitations. However, one may wish to model application-specific properties, in addition to feature correlations. We thus exploit an extension of elastic net that provides such flexibility as discussed in the next section.

2.3 Generalized Sparse Classifiers

To facilitate incorporation of domain-specific properties, such as spatial smoothness in addition to sparsity, into the classifiers listed in Section 2.1, we replace I in (6) with a general non-singular penalization matrix, Γ, which transforms (4) into the following optimization problem:

$$\hat{a} = \arg\min_{a}\left(\left\|y - X^T a\right\|_2^2 + \alpha\left\|\Gamma a\right\|_2^2 + \beta\left\|a\right\|_1\right). \quad (7)$$

We refer to classifiers built from (7) as GSC, which clearly inherit all desired characteristics of (5); namely sparse feature selection without the number of features being limited by the number of samples, and efficient classifier learning through LARS. To demonstrate the power of GSC, we construct a spatially-smooth sparse LDA (SSLDA) classifier by first solving for y in (2) with:

$$W_{ij} = \begin{cases} 1/m_t, & l_i = l_j = t \\ 0, & otherwise \end{cases}, \quad (8)$$

where m_t is the number of samples in class t and $D = I$ [15]. We then apply (7) with Γ being the spatial Laplacian operator to encourage spatial smoothness. SSLDA thus enables explicit modeling of the spatial correlations inherent in brain activity, and hence encourages sparse sets of spatially-contiguous clusters to be jointly selected as features. α and β in (7) are optimized using nested cross-validation [5,16].

3 Materials

The StarPlus data [17] were used for validation. Data from six healthy subjects were kindly made available by the authors of [5]. Each dataset comprised preprocessed voxel time courses within 25 regions of interest (ROIs). ROIs included calcarine

fissure, supplementary motor areas, left inferior frontal gyrus, bilateral dorsolateral prefrontal cortex, frontal eye fields, inferior parietal lobule, intraparietal sulcus, inferior temporal lobule, opercularis, posterior precentral sulcus, supramarginal gyrus, superior parietal lobule, temporal lobe, and triangularis. In each trial, subjects were required to look at a picture (sentence) followed by a sentence (picture) and 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. Each subject performed 40 trials. In half of the trials, the picture preceded the sentence, and vice versa. fMRI brain volumes were acquired at a TR of 500 ms. To account for delay in the hemodynamic response, only the 8 brain volumes collected 4 s after stimulus onset were used. We treated signal intensity of each voxel as a feature and each brain volume as a sample, resulting in 320 samples per class. Further details regarding the experiment and data acquisition could be found in [5,17].

4 Results and Discussion

Quantitative results obtained using the proposed SSLDA to discriminate brain volumes associated with a sentence from those associated with a picture are shown in Fig. 1(a). For comparison, we also applied LDA [7], linear SVM [5,8,9], sparse LDA (SLDA), and LDA with elastic net regularization (EN-LDA) to the StarPlus data. Five-fold cross validation was used to estimate predictive accuracy [5,16]. LDA resulted in the worse overall predictive accuracy, which was likely due to overfitting. Controlling overfitting using SLDA improved accuracy, but SLDA's constraint on the number of features might have limited its predictive performance compared to EN-LDA. Using linear SVM, which is also prone to overfitting, surprisingly outperformed SLDA. We again suspect this result to have arisen from SLDA's limitation on the number of features. Using our proposed SSLDA resulted in the best overall predictive performance with an average accuracy of 93.7% across subjects.

In addition to providing better predictive performance, SSLDA also produced more neurologically sensible classification weight patterns, as shown in Fig 2. We only show a representative slice of four exemplar subjects due to space limitation. LDA (Fig. 2(a)) resulted in spatially-scattered weight patterns with larger weights randomly-distributed across the brain. These weight patterns substantially deviate from the widely-accepted conjecture of how brain activity is spatially distributed in localized clusters [19], as opposed to being randomly-scattered across voxels. Similar spatially-spurious weight patterns were observed with linear SVM (Fig. 2(b)), despite achieving higher predictive accuracies compared to LDA. These results thus illustrate the important, yet highly overlooked, fact that higher predictive accuracies do not necessarily translate to more neurologically interpretable weight patterns, which is the primary objective of fMRI studies [13].

SLDA (Fig. 2(c)) resulted in overly sparse weight patterns, which was partially alleviated with EN-LDA (Fig. 2(d)). However, promoting joint selection of correlated voxels appeared inadequate to generate spatially-smooth patterns. We suspect this irregularity in weight patterns was due to voxel correlations being obscured by noise.

Explicitly modeling spatial correlations using SSLDA (Fig. 2(e)) produced smoother patterns than EN-LDA with the weights forming spatially-contiguous clusters. Also, larger weights were more consistently assigned to localized areas within brain regions implicated for discriminating sentences from pictures; namely the temporal lobe (green dashed circles), the inferior temporal lobule (blue dotted circles), and the calcarine fissure (red circles) around which the visual cortex lies [20].

To quantify the improvement in spatial continuity, we divided each subject's brain into B bins and used the spatial distribution metric (SDM) employed in [13]:

$$SDM = H / H_0, \; H = -\sum_{b=1}^{B} p_b \log p_b, \; p_b = Q^{-1} \sum_{i \in b} |a_i|, \tag{8}$$

where $Q = \|a\|_1$ and $H_0 = \log \|a\|_0$. A bin size of 3×3×3 was used [13]. SDM ranges from 0 to 1, where 0 corresponds to a_i concentrated within one bin and 1 corresponds to a_i evenly distributed across the bins [13]. SSLDA achieved the lowest SDM among the sparse regularization techniques tested (Fig. 1(b)), thus demonstrating that, in addition to improving predictive accuracy, explicitly modeling spatial correlations provides more spatially-contiguous weight patterns than indirectly modeling voxel correlations with EN-LDA. We note that SDM is not applicable for LDA and SVM since weights were assigned to all voxels.

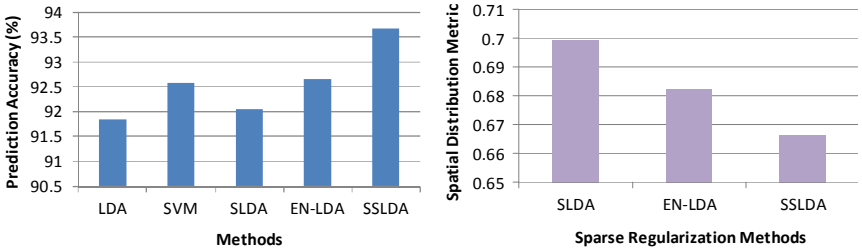

Fig. 1. Quantitative results on StarPlus data. (a) SSLDA resulted in the best overall predictive accuracy and (b) the lowest SDM among the sparse regularization methods. This suggests that SSLDA correctly assigns more weights to localized clusters, as opposed to isolated voxels.

5 Conclusion

In this paper, we proposed a new group of classifiers, "Generalized Sparse Classifiers," for performing large-scale classification problems such as those seen in fMRI studies. By adopting the spectral regression framework and extending the elastic net, GSC enables simultaneous sparse feature selection and classification with greater flexibility in the choice of penalties. Explicitly modeling the spatial correlations in brain activity using GSC resulted in higher predictive accuracy than state-of-the-art classifiers, while generating more neurologically plausible classifier weight patterns. Our results thus suggest that incorporating prior knowledge into classification models can jointly improve predictability and interpretability, which is crucial in medical imaging applications.

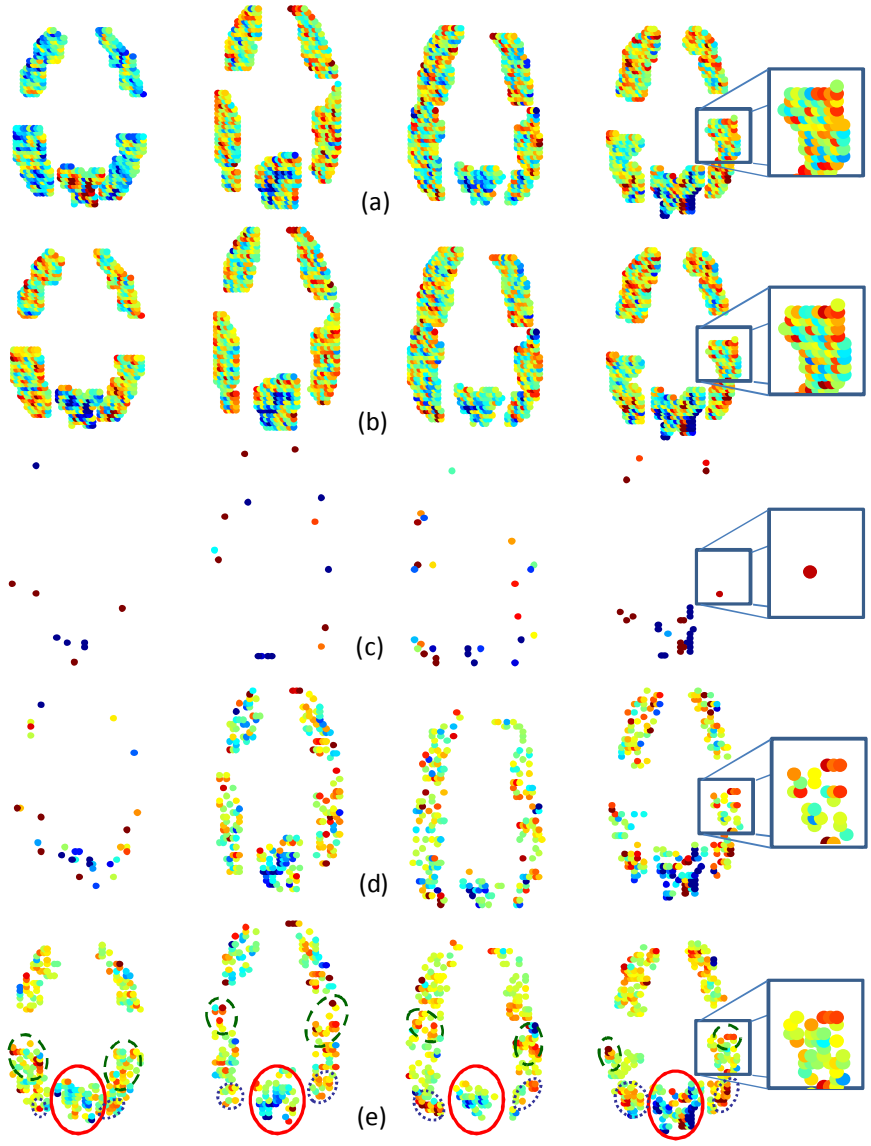

Fig. 2. Classifier weights of methods tested. Red (blue) indicates large positive (negative) weights. LDA (a) and linear SVM (b) resulted in randomly-distributed weight patterns. SLDA (c) generated overly sparse weights, partially overcome by EN-LDA (d). SSLDA (e) produced weight patterns comprising spatially contiguous clusters, which conforms to how brain activity is known to distribute across the brain in localized clusters. Also, SSLDA provides spatially smoother weight patterns than EN-LDA. Moreover, SSLDA consistently assigned larger weights to brain regions (circled) implicated in discriminating sentences from pictures.

References

1. Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond Mindreading: Multi-voxel Pattern Analysis of fMRI Data. Trends Cogn. Sci. 10(9), 424--430 (2006)
2. Haynes, J.D., Rees, G.: Decoding Mental States from Brain Activity in Humans. Nat Rev. Neurosci. 7(7), 523--534 (2006)
3. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D., Frackowiak., R.S.J.: Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Hum. Brain Mapp. 2(4), 189--210 (1995)
4. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Aschouten, J.L., Pietrini, P.: Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science 293(5539), 2425--2430 (2001)
5. Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F.,Wang, X., Just, M., Newman, S.: Learning to Decode Cognitive States from Brain Images. Mach. Learn. 57, 145--175 (2004)
6. Damon, C., Pinel, P., Perrot, M., Michel, V., Duchesnay, E., Poline, J.B., Thirion, B.: Discriminating between Populations of Subjects based on fMRI Data Using Sparse Features Selection and SRDA Classifier. In: MICCAI Analysis of Functional Medical Images Workshop, pp. 25--32 (2008)
7. Haynes, J.D., Rees, G.: Predicting the Orientation of Invisible Stimuli from Activity in Human Primary Visual Cortex. Nat. Neurosci. 8(5), 686--691 (2005)
8. Cox, D., Savoy, R.: Functional Magnetic Resonance Imaging (fMRI) "Brain Reading": Detecting and Classifying Distributed Patterns of fMRI Activity in Human Visual Cortex. NeuroImage 19(2), 261--270 (2003)
9. Balci, S.K., Sabuncu, M.R., Yoo, J., Ghosh, S.S., Gabrieli, S.W., Gabrieli, J.D.E., Golland, P.: Prediction of Successful Memory Encoding from fMRI Data. In: MICCAI Analysis of Functional Medical Images Workshop, pp. 97--104 (2008)
10. Carlson, T.A., Schrater, P., He, S.: Patterns of Activity in the Categorical Representations of Objects. J. Cogn. Neurosci. 15, 704--717 (2003)
11. Yamashita, O., Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse Estimation Automatically Selects Voxels Relevant for the Decoding of fMRI Activity Patterns. NeuroImage 42, 1414--1429 (2008)
12. Ryali, S., Supekar, K., Abrams, D.A., Menon, V.: Sparse Logistic Regression for Whole-brain Classification of fMRI Data. NeuroImage 51, 752--764 (2010)
13. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and Interpretation of Distributed Neural Activity with Sparse Models. 44, 112-122 (2009)
14. Zou, H., Hastie, T., Tibshirani, R.: Sparse Principal Component Analysis. J. Comp. Graph. Stat. 15(2), 265--286 (2006)
15. Cai, D., He, X., Han, J.: SRDA: Spectral Regression: A Unified Approach for Sparse Subspace Learning. In: Int. Conf. Data Mining, pp. 73--82 (2007)
16. Zou, H., Hastie, T.: Regularization and Variable Selection via the Elastic Net. J. Royal Stat. Soc. B67, 301--320 (2005)
17. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin., S.: Graph Embedding and Extension: A General Framework for Dimensionality Reduction. IEEE Trans. Pat. Ana. Machine Intell. 29(1), 40--50 (2007)
18. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/
19. Thirion, B., Flandin, G., Pinel, P., Roche, A., Poline, J.B.: Dealing with the Shortcomings of Spatial Normalization: Multi-subject Parcellation of fMRI Datasets. Hum. Brain Mapp. 27, 678--693 (2006)
20. Vandenberghe, R., Price, C., Wise, R., Josephs, O., Frackowiak, R.S.J.: Functional Anatomy of a Common Semantic System for Words and Pictures. Nature 383, 254--256 (1996)