Data-driven Selection of Relevant fMRI Voxels using Sparse CCA and Stability Selection

Kristian Nybo\textsuperscript{1}, John Shawe-Taylor\textsuperscript{2}, Samuel Kaski\textsuperscript{1,3}, Janaina Mourao-Miranda\textsuperscript{2}

\textsuperscript{1} Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University
\textsuperscript{2} Centre for Computational Statistics and Machine Learning, University College London
\textsuperscript{3} Department of Computer Science, University of Helsinki

Abstract. Functional MRI data is both high-dimensional and noisy, and consequently many machine learning methods overfit severely if applied without first reducing the dimensionality. We propose an algorithm to find the most relevant voxels automatically using a Sparse Canonical Correlation Analysis (SCCA) model, the statistical framework of stability selection, and a rich feature representation for the experimental task. We empirically validate the voxels chosen by our algorithm by applying it to fMRI data from 16 subjects listening to classical music classified as 'happy', 'sad' or 'neutral'. Our algorithm selects voxels based on automatically extracted music features, without access to the label information. We show that a linear Support Vector Machine classifier trained on a small subset of voxels selected using our method can match or beat an SVM trained on the full brain in the task of predicting the labels 'happy' and 'sad', while performing much better than an SVM trained on a similar randomized voxel set. We further show using Representational Similarity Analysis that the voxel subsets chosen by our method are related to voxels that are denoted as active by traditional Statistical Parametric Mapping.

Keywords: fMRI, feature selection, stability selection, sparse canonical correlation analysis

1 Introduction

Functional MRI data is notoriously high-dimensional and noisy. One sample can have upwards of 200,000 variables, data for one subject may consist of only a few hundred samples, and the meaningful activations that analysis methods try to find typically show up as 2% or 3% increase over the noise level in the BOLD signal \cite{2}. One traditional solution to this problem is to model each voxel independently, as is done in Statistical Parametric Mapping \cite{5}. Another solution is to limit the analysis to a small subset of voxels, regions of interest known a priori to be related to the experimental task.
Both solutions have been applied successfully, but they also have their drawbacks. Modelling each voxel independently may hide effects that are explained well by a group of voxels but only weakly explained by any single voxel. Choosing regions of interest based on neuroscientific prior information may exclude regions that are not yet known to be relevant for the task, and also requires us to solve the problem of locating those regions in each individual brain.

We propose a data-driven approach for discovering the task-relevant voxels without relying on explicit neuroscientific prior information or even the categorical labels (e.g. ‘happy music’, ‘sad music’) derived from a typical event or block experimental design. Instead we make use of an automatically extracted rich feature representation of the stimulus. By ‘rich’ we mean a real-valued, multi-dimensional set of features that we expect to capture every aspect of the experimental task that we are interested in.

2 Methods

2.1 Sparse CCA

At the core of our voxel selection algorithm is the Sparse Canonical Correlation Analysis (SCCA) model due to Hardoon and Shawe-Taylor [6]. A distinctive feature of this particular sparse CCA formulation is that one of the views is represented as a matrix $X \in \mathbb{R}^{M \times N}$ (primal representation), where each column is one sample vector, while the other view is represented as a kernel matrix $K \in \mathbb{R}^{M \times M}$ (dual representation).

SCCA tries to find sparse projection vectors $w \in \mathbb{R}^{N}$ and $e \in \mathbb{R}^{M}$ such that the projections $X^T w$ and $Ke$ are maximally correlated. This can be formulated as the convex optimization problem

$$\min_{w, e} \|X^T w - Ke\|_2^2 + \mu \|w\|_1 + \gamma \|e\|_1,$$

subject to the constraints $\|e\|_{\infty} = 1$ and $e \geq 0$ (element-wise). The optimization problem is solved using a fast iterative algorithm that alternates between optimizing $w$ and optimizing $e$. The constraint $\|e\|_{\infty} = 1$ is enforced partly by fixing a specific element $e_k$ of the vector $e$ to be 1 at all times. The so-called seed index $k$ is a parameter that needs to be set by the user, for example by trying all values of $k$ and choosing the one that gives the maximal correlation. The regularization parameters $\mu$ and $\gamma$ are determined automatically using an approach that has been shown to work well in practice; see [6] for more details.

2.2 Voxel selection with SCCA and stability selection

When applying SCCA, we use the primal representation $X$ for the fMRI scans and the dual representation $K$ for the music features. The basic idea is to run SCCA and then take as relevant the voxels that have non-zero entries in the sparse weight vector $w$. We improve upon this simple concept by incorporating...
stability selection as introduced by Meinshausen and Bühlmann [10]. We randomly subsample 10% of the voxels and 66% of the samples and train SCCA on those. We repeat this 1000 times, and keep count of how many times each voxel received a non-zero weight relative to how many times it was included in the random subsample. We interpret this ratio as an empirical probability of relevance: if a voxel is almost always assigned a non-zero weight no matter what other voxels are available, it is more likely to be relevant than a voxel that only rarely gets a non-zero weight. We include in our set of relevant voxels all voxels whose probability exceeds a certain threshold $p_{\text{thres}}$. We discuss setting this threshold in Section 3.3.

Due to the subsampling, each iteration of the stability selection procedure effectively trains SCCA on a different data set, and so the optimal seed index $k$ could be different each time. An exhaustive search on each iteration would be prohibitively expensive. Instead we first clustered the music feature sample vectors using $K$-means with 20 clusters. For each of these index clusters, we then performed stability selection and SCCA on the full data set, as described in the previous paragraph, so that on each iteration a new seed index $k$ was chosen randomly from the corresponding cluster. Thus we obtained 20 different sets of empirical voxel probabilities; to create a single stable voxel set, we included every voxel whose probability exceeded $p_{\text{thres}}$ in at least one of the 20 sets.

3 Experiments

3.1 The data set

We applied our algorithm to a data set originally introduced in [11]. 16 healthy subjects listened to 20 extracts of orchestral classical music, each 30 seconds long. Based on a separate pilot study, each piece was known to be either ‘happy’, ‘sad’ or ‘neutral’. The scanning interval was 2 seconds, so we had altogether 300 samples for each subject. For details on data pre-processing, see [11]. We applied a voxel mask to remove the eyes and the skull, leaving 219,727 voxels per scan. We divided the data set into 150 training samples and 150 test samples; the same random, balanced partition was used for each subject. Voxel selection was performed only on the training data.

3.2 Music features

We automatically extracted a 26-dimensional set of features from CD-quality music (encoded in .WAV format) using the MIR toolbox for MATLAB [8]. The feature set we used consists of all the low-level features (all features but pulse clarity, key clarity and tonal centroid) in the feature set recently used by Alluri et al. to study fMRI measurements from subjects listening to tango [1]. Music feature vectors within each 2-second window corresponding to one fMRI scan were then averaged to obtain one music feature sample for each fMRI scan. Finally the music samples were convolved with a standard hemodynamic response function.
3.3 SVM classification of fMRI scans as 'happy' or 'sad'

Our goal is to retrieve voxels that are relevant to the task, that is, listening to music. Our method has two possible points of failure: first, the set of music features might not be rich enough to capture relevant aspects of the music; second, our algorithm might fail to find the voxels that are relevant to the features. We can address both points at once by trying to predict the music class labels (which our algorithm does not use) from the selected voxel sets. Presumably a good set of features would capture at least some of the aspects of music that interact to create the subjective experience of 'happy' or 'sad' music. Thus if the selected voxels are good for predicting the labels, this would suggest that the voxels are indeed relevant to the music.

To obtain a two-class problem, we removed the neutral samples from the training and test data set defined in Section 3.1. Separately for each subject, we then trained linear Support Vector Machines (SVM) [4] using libSVM [3] on three different sets of variables:

1. the full brain;
2. the subset of stable voxels chosen by applying our algorithm to on the training data;
3. and a random set of voxels generated by taking set #2 and moving each contiguous cluster of voxels to a random location in the brain (overlap allowed).

In training the SVM, we had two parameters to set: the SVM regularization parameter $C_{\text{SVM}}$ and the stability probability threshold $p_{\text{thres}}$ for our voxel selection method. We selected the parameters using 5-fold cross-validation on the training data, with $C_{\text{SVM}}$ taking values from $10^{-9}$ to $10^{-2}$ and $p_{\text{thres}}$ taking values from 0 to 0.8. As the utility function maximized under cross-validation we used $((\text{SVM classification accuracy in percent}) - \sigma(\% \text{ of total voxels used})$. Here $\sigma$ is simply a non-negative hyperparameter specified by the user to control the trade-off between classification accuracy and the number of voxels used. Here we set $\sigma = 7$ to favor sparsity over small performance improvements.

The classification accuracy for the SVM trained on all the voxels, averaged over subjects, was 82.10%. The corresponding accuracy for the SVM trained on stable voxels was 79.05%, and the stable voxels comprised on average 0.10% of the full voxel set (around 200 voxels). The accuracy for the randomly translated clusters was 65.60% (averaged over 10 different random sets).

In addition to selecting the stability threshold $p_{\text{thres}}$ separately for each subject using cross-validation as above, we also tried a wide range of values for $p_{\text{thres}}$, using the same value for all subjects, to investigate our method’s sensitivity to this parameter. Our findings are summarized in Table 1. We obtained a classification rate that was slightly better than the full-brain classification rate for a large range of values of $p_{\text{thres}}$, but for very small values — corresponding to a large set of voxels — the randomized voxel clusters performed equally well.

It is surprising that an SVM trained on a random set containing 15% of the voxels can on the average perform slightly better than a classifier trained on the full brain. The fact that we obtain better classification rates by fixing $p_{\text{thres}}$ for
all subjects suggests that the cross-validation procedure may have overfit the training data. On the other hand, the cross-validation procedure did achieve a larger difference between the respective classification rates of stable and random voxels than any of the fixed parameter values.

<table>
<thead>
<tr>
<th>Stability threshold $p_{\text{thres}}$</th>
<th>0.02</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM accuracy (stable), %</td>
<td>83.53</td>
<td>82.85</td>
<td>82.77</td>
<td>83.19</td>
<td>81.50</td>
<td>77.28</td>
<td>68.50</td>
</tr>
<tr>
<td>SVM accuracy (random), %</td>
<td>82.56</td>
<td>75.57</td>
<td>74.66</td>
<td>72.09</td>
<td>69.30</td>
<td>64.54</td>
<td>57.92</td>
</tr>
<tr>
<td>Voxels used, % of full</td>
<td>15.95</td>
<td>6.40</td>
<td>2.79</td>
<td>0.96</td>
<td>0.20</td>
<td>0.043</td>
<td>0.0057</td>
</tr>
</tbody>
</table>

3.4 Representational Similarity Analysis

For additional validation, we studied the voxel sets chosen by our method using the Representational Similarity Analysis (RSA) framework introduced by Kriegeskorte et al. [7]. This experiment too was performed separately for each subject.

In the RSA framework each model is encoded as a representational dissimilarity matrix (RDM), where the entry $(i,j)$ is the dissimilarity between samples $i$ and $j$ as represented by the model in question. We can think of the stable voxels (set #2 in Section 3.3) as one model of the brain and of the random voxels (set #3) as another; the dissimilarity between samples $i$ and $j$ under a ‘voxel set model’ is $1 - r_p(i,j)$, where $r_p$ is the Pearson correlation of fMRI scans $i$ and $j$, computed over the relevant set of voxels. Once we have an RDM for each model, we can compare the models by computing a second-order RDM, where each entry is a dissimilarity between two model RDMs. The dissimilarity between two model RDMs is $1 - r_s$, where $r_s$ is the Spearman correlation computed over the upper triangles of the matrices, excluding the identically zero diagonals.

We compare the stable and random voxels with voxel sets based on a traditional SPM analysis, where a univariate linear model was fit separately for each voxel, using the categorical variables ‘happy’, ‘sad’ and ‘neutral’ as the regressors. For each category, we chose the $N_2$ voxels with the largest absolute weights for that category, where $N_2$ is the number of voxels in the set of stable voxels chosen by our algorithm. Thus we obtained three voxel sets, one corresponding to each of the three emotional categories. To compare the five voxel sets, we computed a second-order RDM from the five model RDMs. We averaged the second-order RDMs across subjects to obtain one RDM. Table 2 shows the first two rows of the averaged second-order RDM, corresponding to the stable voxels (set #2, $p_{\text{thres}} = 0.05$) and the random voxels (set #3). We see that the stable voxels are clearly closer to the SPM-derived voxels than are the randomized voxels.

Again we tried a range of values for $p_{\text{thres}}$. As before, for values smaller than 0.05, stable voxels and random voxels were equivalent. For values greater than
0.05, the results were similar to Table 2, but the entries in the three last columns increased, and the difference between the two rows decreased. This is expected; one set of voxels is chosen based on the music features, the other based on the class labels, and in any case we would not expect perfect correlation between SPM regression weights and our voxel relevance probabilities.

Table 2. First two rows of the representational dissimilarity matrix comparing the five different voxel sets in Section 3.4. Stable voxels computed using $p_{\text{thres}} = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>Stable</th>
<th>Random</th>
<th>Happy $\beta$</th>
<th>Sad $\beta$</th>
<th>Neutral $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable voxels</td>
<td>0</td>
<td>0.17</td>
<td>0.28</td>
<td>0.24</td>
<td>0.28</td>
</tr>
<tr>
<td>Random voxels</td>
<td>0.17</td>
<td>0</td>
<td>0.41</td>
<td>0.41</td>
<td>0.40</td>
</tr>
</tbody>
</table>

4 Discussion

We have introduced a novel algorithm for automatically retrieving relevant voxels using sparse canonical correlation analysis, stability selection and a rich feature representation for the experimental stimulus. In the SVM experiment we managed to match or beat the full-brain classifier using a tiny fraction of the voxels. Because the voxel selection method did not use the category labels in choosing the voxels, this indicates that we must have managed to retrieve voxels that carry some information about the music. The RSA analysis provides some corroborating evidence by suggesting a connection with voxels that are indicated as relevant to the categorical labels by traditional SPM analysis.

We are not the first to propose data-driven selection of voxels in general. Recursive Feature Elimination [9] has been used specifically in the context of SVM classification. Yamashita et al. [13] selected voxels relevant for a classification task using a sparse logistic regression classifier. Varoquaux et al. [12] recently introduced a voxel selection algorithm based on LASSO and stability selection. SPM [5] itself can be interpreted as a voxel selection method. What all these methods have in common, however, is that they were designed for discrete categorical labels.

Selecting voxels based on labels is natural if we specifically want to discriminate between pre-defined categories; our original intention was rather to develop a general data-driven dimensionality reduction method for fMRI data that could be used as a preprocessing step for more complicated models that can’t handle high dimensionality. From this perspective the experiments here were only preliminary sanity checks, and the real test lies ahead.

Acknowledgements The authors thank Prof. Steve Williams (Dep. of Neuroimaging, Institute of Psychiatry, KCL) for providing the dataset. JMM was supported by the Wellcome Trust under grant no. WT086565/Z/08/Z. This work has been partly supported by the Academy of Finland (the COIN Center of Excellence) and the AivoAALTO project of Aalto University.
References


