Across-Subject Classification of Single EEG Trials

Blair Bohannan, Jorge Herrera, Lewis Kaneshiro
{blairbo, jorgeh, lewiskan}@ccrma.stanford.edu

INTRODUCTION

The focus of this project is classification, across human subjects, of single trials of EEG recorded while subjects viewed images of human faces and inanimate objects. The data used in this project were originally collected in the Suppes Brain Lab at Stanford for use in another experiment comparing classification rates to Representational Dissimilarity Matrices [in preparation]. Classification for that experiment was done only within-subject (training and testing on one subject at a time), used only LDA classification, and was not implemented by anyone in our project group. Our present goal is to explore a variety of machine-learning techniques with our dataset in three scenarios: within-subject classification; training and testing on all subjects together; and training on nine subjects and testing on the tenth.

Our dataset consists of 124 channels of scalp EEG recorded at 1 kHz from 10 subjects while they viewed individual images (each trial consisting of an image shown onscreen for 500 ms). The original stimulus set, adapted from the stimuli used in [1], consists of 72 images grouped into 6 categories. For the current analysis, we are using only the 24 images from the Human Face and Inanimate Object categories. Each subject viewed each image 72 times; for 12 images per category, 2 categories, and 10 subjects, we have a total of 17,280 EEG trials in our dataset. Our dataset has been highpass filtered for DC offset removal, then lowpass filtered and downsampled by a factor of 16 for smoothing and data reduction. It has also already undergone ICA for removal of eye artifacts [2], and has been converted back to channel space. We have 124 channels and 32 samples (covering roughly the 500-ms interval of stimulus presentation) for each trial, giving a feature vector of length 3,968.

Figure 1: Human Face and Inanimate Object images.

INITIAL IMPLEMENTATIONS

One challenge of this project has been feature reduction and managing data complexity. We explored both supervised and unsupervised learning techniques, and considered different configurations of training and test sets (for instance, training and testing on everyone; training on one person and testing on another; and training on all but one and testing on the last).

We first attempted Naïve Bayes classification and PCA on the entire dataset, using MATLAB's `NaïveBayes` object and `princomp` function respectively. These attempts both resulted in immediate memory errors. We then attempted to increase and manage MATLAB memory, both through the command-line interface and by

launching a graphical interface using the memory manager. We continued to have memory failure issues.

Next, we attempted to reduce the dataset size, and we used a Naïve Bayes classifier on individual channels, all subjects combined, using 10-fold cross validation. This resulted in an accuracy rate for each channel and produced a working first-attempt classifier for the dataset. Using this method, we identified the best channel (96, with accuracy of 66.1%) and the worst channel (20, with accuracy of 50.0%). We expect that using more channels simultaneously (more features) should achieve accuracy higher than 66.1%.

In an effort to improve on this process, we used the same single-channel iteration method over all subjects using 10-fold cross validation with both a linear and quadratic discriminant analysis classification model (LDA and QDA respectively). These new classification models improved accuracy slightly over Naïve Bayes. The best classification accuracy using LDA on a single channel was 67.2%, again for channel 96. We also achieved accuracy of 66.3% using QDA for channel 96.

We then explored within-subject classification using Naïve Bayes and LDA by using one channel at a time. Both models produced accuracy rates similar to those described above.

In an attempt to reduce data dimensionality, we decided to select a subset of samples per channel. To identify the appropriate range of samples, we considered the channel that performed best by itself (96 for both NB and LDA), and plotted the averages across all subjects and trials for this channel alone, for each image category separately. We then took the absolute value of the difference between the averages for faces versus objects, and picked the range of samples with the greatest magnitude difference across the averages, which turned out to be samples 6-18 (corresponding to the time range of 80-272 ms after stimulus onset). Thus, to reduce our dataset size further, we used only samples 6-18 for all channels and re-ran the LDA and Naïve Bayes models.

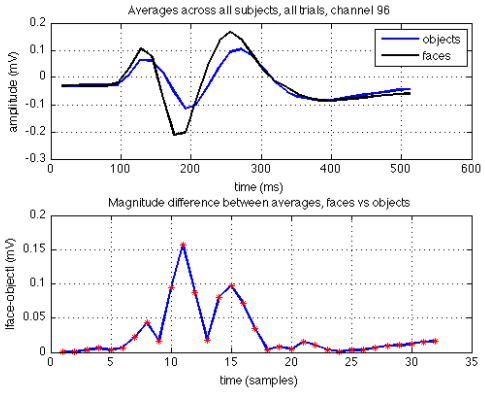

Figure 2: Top - Averaged EEG for each image category, across all subjects and trials (8,640 trials per category). Bottom - Magnitude difference between the averages.

Using this limited range of samples, running LDA with 10-fold cross validation using all channels at once, within-subject for all ten subjects, we generated an average accuracy of 73.4%. Using the same process with a Naïve Bayes model, we generated a slightly lower accuracy of 64.9%. For LDA and Naïve Bayes, we have the following confusion matrix, expressed as probabilities over the set of all trials:

	Predicted Object	Predicted Face
Actual Object	LDA = 37% NB = 34%	LDA = 13% NB = 16%
Actual Face	LDA = 13% NB = 19%	LDA=37% NB = 31%

Figure 3: Conditional probability matrix for LDA and Naïve Bayes attempts, using samples 6-18.

Finally, to do an initial assessment of the validity of the model across subjects, we computed the Precision and Recall values for all 10 subjects independently. Subjects who had higher Precision also tended to have higher Recall.

LIBLINEAR SVM OVER PCA

WITHIN-SUBJECT CLASSIFICATION

We proceeded by performing SVM classification using a linear kernel (implemented using LIBLINEAR). As a solution to our MATLAB memory issues, we performed all memory-intensive computations using Stanford MATLAB resources (cm-matlab.stanford.edu with 16GB memory and a Suppes Brain Lab machine with 20GB memory, both running 64-bit MATLAB). We also attempted SVM classification on smaller sample subsets (samples 6-18) in an effort to improve efficiency.

We considered performing PCA on all channels, as opposed to using a sample subset in the classification. We performed both spatial and temporal PCA, both to improve efficiency and as a noise reduction technique in an attempt to improve classification rates. By spatial PCA we mean that we are finding Principal Components across the samples (time points), while temporal PCA refers to finding Principal Components across the channels[1].

Data were pre-processed to normalize per-channel mean and variance across all subjects prior to running PCA as a standard step to ensure that the first principal component describes the direction of maximum variance. To further increase efficiency, we selected Principal Components whose respective eigenvalues, sorted in descending order, had a normalized cumulative sum reaching a pre-determined threshold (such as 0.80 or 0.90). The threshold of 0.8 compressed our data size by a factor of 5.26 for spatial PCA and 6.69 for temporal PCA.

Outlining our data sampling process for classification model training and testing, we randomly partitioned our 10-subject dataset into an 80% training/ 20% testing split. To establish a baseline accuracy rate for comparison to across-subject classification, we first performed within-subject classification, using SVM with 10-fold cross validation and attained the following accuracies (mean accuracy 82.4%):

S1	S2	S3	S4	S5
83.4	84.0	84.5	81.2	88.4
S6	**S7**	**S8**	**S9**	**S10**
81.2	86.5	73.0	76.5	85.2

Figure 4: Within-subject accuracies using SVM with spatial PCA.

ENSEMBLE METHODS

To conduct classification across subjects, we initially took an ensemble voting classification method, building 9 SVM models using individual subjects' spatial PCA data. Testing was performed by converting test-subject data into each other individual's training model PCA space, obtaining 9 sets of predicted labels, and taking the majority vote to produce a final ensemble classification. As expected for pairwise test/training individuals' classification, SVM models produced accuracies near 100% for same training and test subjects, while other pairwise classifications performed poorly, with mean accuracy of only 61.3% (Figure 5). However, after combining these weak learners and taking the majority vote of the predicted

[1] To the best of our understanding, this follows the spatial/temporal distinction made with ICA, though there appears to be some debate on the matter: http://sccn.ucsd.edu/~scott/tutorial/questions.html - TemporalICA

label, the resulting ensemble classifier produced satisfactory accuracy levels between 64-75% (Figure 6).

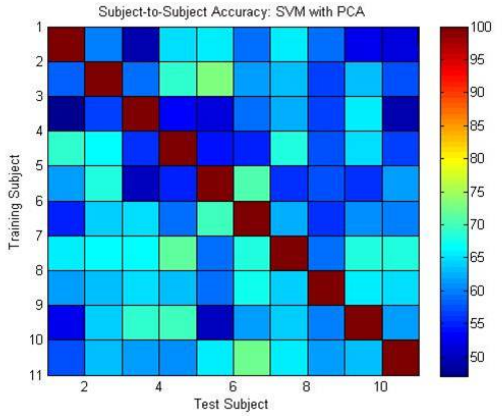

Figure 5: Pairwise accuracies for SVM with PCA.

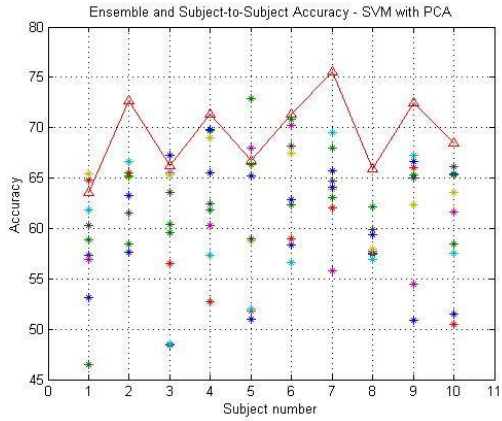

Figure 6: Individual weak learner accuracies (*) versus majority vote (red line).

LEAVE-ONE-OUT SVM

We then took the approach of building SVM models using LIBLINEAR on 9 subjects, systematically generating the models using training examples of raw channel data and spatial PCA for separate models, and testing on the 10th subject. Based on the LOOSVM (leave-one-out SVM) accuracies, we observed two subjects (S8, S10) who consistently under-performed as test datasets. We chose to remove these two subjects' datasets from the training model to produce a 7-subject LOOSVM model for comparison.

These LOOSVM classifiers achieve accuracy in 9-subject channel space ranging from 54-66%, 9-subject spatial PCA space from 55-75%, and, after removing 'outliers' based on the previous LOOSVM accuracy measures, 7-subject spatial PCA accuracy of 63-78%.

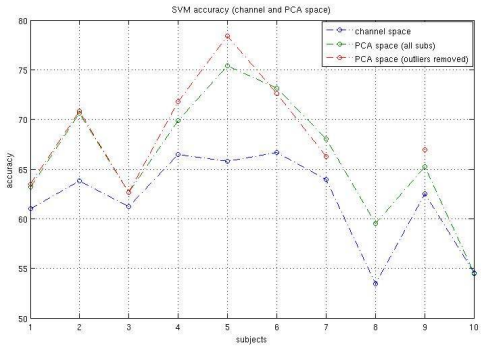

Figure 7: Accuracy rates for LOOSVM in channel space (blue); using spatial PCA with all subjects (green); and spatial PCA with S8 and S10 removed (red).

As an additional exercise, we also performed the above classifications on the smaller sample range (samples 6-18). The rates proved to be comparable to those in Figure 7, so there appeared to be no added benefit of using only a subset of the samples in addition to a subset of Principal Components, especially as the PCA reduction had already decreased our processing time and improved efficiency.

TEMPORAL PCA

Finally, we attempted the same LOOSVM classification using temporal rather than spatial PCA, with data from all ten subjects. Here we also experimented with different thresholds (0.8, the spatial PCA threshold, and also 0.9) for the number of Principal Components to choose; results are shown below. It appears that spatial PCA produces higher accuracies for

some subjects, while temporal PCA works better for others.

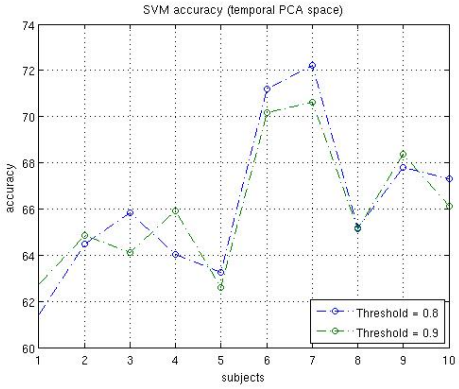

Figure 8: Accuracy rates for LOOSVM using temporal PCA and variable thresholds.

CONCLUSION

We observed the highest across-subject testing accuracies using the 7-subject LOOSVM classifier in spatial PCA space, with surprisingly comparable measures achieved using a 9-subject ensemble majority vote classifier built on weak learner pairwise SVM classifiers. These classification methods approach the accuracy achieved by within-subject classification. Our project implies the ability to train classifier models on training subjects completely separate from testing subjects.

We selected training outliers by training a LOOSVM model and testing on individual training subjects. We then improved overall accuracy by removing those training subjects whose data tested relatively poorly, prior to building the final LOOSVM model. The relative success of 9-subject ensemble majority vote accuracies, compared to the individual pairwise accuracies, suggests underlying diversity between single-subject SVM classifiers.

It is interesting to note that different subjects' datasets classified better on different attempts described in this paper. For example, S5, the top classifier in the within-subjects scenario, was also the highest classifying dataset in spatial PCA, but not in the ensemble method or in temporal PCA. In contrast, S8 was among the lower-classifying subjects for within-subject classification (significantly the lowest for this case), ensemble methods, and spatial PCA, but did slightly better with temporal PCA.

Our classification results suggest that to some extent, processing of face and object categories can be generalized across human subjects and applied to new subjects' data on even the single-trial level. It would be of further interest to us to better quantify the nature of the useful spatial and temporal features that contribute to successful across-subject classification.

REFERENCES

[1] Kriegeskorte, Nikolaus, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerry Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron* **60** (December 2008) 1-16.

[2] Jung, Tzyy-Ping, Colin Humphries, Te-Won Lee, Scott Makeig, Martin J. McKeown, Vicente Iragui, and Terrence J. Sejnowski. Extended ICA Removes Artifacts from Electroencephalographic Recordings. *Advances in Neural Information Processing Systems 10*, M. Jordan, M. Kearns, and S. Solla (Eds.), MIT Press, Cambridge MA (1998) 894-900.