Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies

Polina Golland[1] and Bruce Fischl[2]

[1] Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge, MA.
polina@ai.mit.edu
[2] Athinoula A. Martinos Center for Biomedical Imaging
Massachusetts General Hospital, Harvard Medical School, Boston, MA.
fischl@nmr.mgh.harvard.edu

Abstract. Estimating statistical significance of detected differences between two groups of medical scans is a challenging problem due to the high dimensionality of the data and the relatively small number of training examples. In this paper, we demonstrate a non-parametric technique for estimation of statistical significance in the context of discriminative analysis (i.e., training a classifier function to label new examples into one of two groups). Our approach adopts permutation tests, first developed in classical statistics for hypothesis testing, to estimate how likely we are to obtain the observed classification performance, as measured by testing on a hold-out set or cross-validation, by chance. We demonstrate the method on examples of both structural and functional neuroimaging studies.

1 Introduction

Image-based statistical studies typically compare two or more sets of images with a goal of identifying differences between the populations represented by the subjects in the study. For example, the shape of subcortical structures or the cortical folding patterns in the scans of schizophrenia patients would be compared to a set of brain images of matched normal controls to test a particular hypothesis on the effects of the disease on the brain. The analysis starts with a feature extraction step that creates a numerical representation of the example images in a form of feature vectors, followed by statistical analysis that generates a description of differences between the two groups and an estimate on how well the detected differences will generalize to the entire population. The high dimensionality of the feature space and the limited number of examples place this problem outside the realm of classical statistics. The solutions used today in most studies simplify the analysis either by reducing the description of the anatomy to a single measurement, such as the volume of the hippocampus, or by considering every feature separately, for example, comparing cortical thickness at every location on the cortical surface independently of its neighbors. The former sacrifices localization power, while the latter ignores potential dependencies among the features, making integration of the individual statistical tests

C.J. Taylor and J.A. Noble (Eds.): IPMI 2003, LNCS 2732, pp. 330–341, 2003.

very difficult. An alternative approach demonstrated recently in several studies of neuroanatomy and function is to train a classifier function for labeling new examples [2,9,10,13,14]. It is based on a presumption that if a classifier function can label new examples with better than random accuracy, the two populations are indeed different, and the classifier implicitly captures the differences between them. The training algorithm does not have to assume independence among features and therefore can discriminate between the two groups based on the entire ensemble of highly localized features[1].

This approach to statistical analysis uses the classifier performance as a measure of robustness of the detected differences, or of dissimilarity of the populations in question. We can estimate the expected accuracy of the classifier by testing it on a separate hold-out set, as the average test error is an unbiased estimator of the expected performance, but the small size of the test set is typically insufficient for the variance of this estimator to be low enough to lead to a useful bound on how close we are to the true expected error [12]. Furthermore, the total number of available examples might be so small that we have to resort to cross-validation, such as bootstrap or jackknife procedures, in order to estimate the expected accuracy [4]. Unfortunately, the cross-validation trials are not independent and therefore do not allow variance estimation at all without extensive modeling of the dependence of errors in the cross-validation trials[2]. Thus, neither testing nor cross-validation provides satisfactory estimates on how close the observed test error to the true expected error of the trained classifier. Furthermore, the high dimensionality of the data often renders many theoretical bounds based on the complexity of the data or that of the classifier useless.

In this paper, we demonstrate how permutation tests [11] can provide a weaker guarantee, namely that of statistical significance. This approach effectively reformulates the question on the differences between the populations, as measured by the classifier performance, in the traditionally used framework of hypothesis testing. Intuitively, it provides a guarantee on how likely we were to obtain the observed test accuracy by chance, only because the training algorithm identified some pattern in the high-dimensional data that happened to correlate with the membership labels as an artifact of a small data set size. Permutation tests were originally developed in statistics for testing whether the observed differences between two data sets, as measured by a particular statistic, are likely to occur under the null hypothesis that assumes that the two distributions that generated the data are identical. Since the test does not assume a generative model for the data to derive the distribution of the statistic under the null hypothesis, but rather estimates it empirically, it is applicable to a wide range of problems. We apply permutation tests to the classification setting by using the

[1] Examples of training algorithms used in this domain include Fisher Linear Discriminant, its generalization to two arbitrary Gaussian distributions, Support Vector Machines and others.

[2] While the mean of the set of cross-validation error measurements is an unbiased estimate of the expected error, the variance of the set is an overly optimistic estimate of the unknown variance of the cross-validation estimator for most training algorithms.

estimated classifier accuracy as a statistic that measures how different the two classes are. The null hypothesis is that the selected family of classifiers cannot learn to predict labels based on the given training set. The test estimates the statistical significance of the classifier by estimating the probability of obtaining the observed classification performance under the null hypothesis.

In the next section, we provide the necessary background on hypothesis testing and permutation tests. In Section 3, we explain how to estimate statistical significance of a classifier function, followed by several experimental examples in Section 4 and discussion and conclusions. Before we proceed, it is worth mentioning that permutation tests have been used previously in neuroimaging studies to estimate statistical significance of individual voxels or clusters of voxels, both in anatomical studies [1,16] and for signal detection in fMRI [15]. We explain the differences between such use of permutation tests and the work presented in this paper in Section 3.1. We believe permutation testing is a little known statistical tool that many other researchers in this community will find useful in their work.

2 Background. Hypothesis Testing

In two-class comparison hypothesis testing, the differences between two data distributions are measured using a data set statistic, which is a function

$$\mathcal{T} : (\mathbb{R}^n \times \{-1, 1\})^* \mapsto \mathbb{R},$$

such that for a given data set $\{(\mathbf{x}_k, y_k)\}_{k=1}^l$, where $\mathbf{x}_k \in \mathbb{R}^n$ are observations and $y_k \in \{-1, 1\}$ are the corresponding class labels, $\mathcal{T}(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_l, y_l)$ is a measure of how similar the subsets $\{\mathbf{x}_k | y_k{=}1\}$ and $\{\mathbf{x}_k | y_k{=}-1\}$ are. The null hypothesis typically assumes that the two classes have identical data distributions,

$$\forall \mathbf{x} \ : \ p(\mathbf{x}|y{=}1) = p(\mathbf{x}|y{=}-1).$$

The goal of the hypothesis testing is to reject the null hypothesis at a certain level of significance α, which defines the maximal acceptable probability of false positive (declaring that the classes are different when the null hypothesis is true). For any value of the statistic, the corresponding *p-value* is the highest level of significance at which the null hypothesis can still be rejected. In classical statistics, the data are often assumed to be one-dimensional ($n = 1$), but any multi-variate statistic $\mathcal{T}$ can be used in this fairly general framework.

For example, in the two-sample t-test, the data in the two classes are assumed to be generated by one-dimensional Gaussian distributions of equal variance. The null hypothesis furthermore assumes that the distributions have the same mean. The probability density of the so-called *t-statistic*, equal to the difference between the sample means normalized by the standard error, under the null hypothesis is the well known Student distribution [17]. If the integral of the Student distribution over the values higher than the observed t-statistic is smaller than the desired significance level α, we reject the null hypothesis in favor of the alternative hypothesis that the means of the two distributions are different.

In order to perform hypothesis testing, we need to know the probability distribution of the selected statistic under the null hypothesis. Unfortunately, deriving a parametric distribution for a particular statistic requires making strong assumptions on the generative model of the data. Consequently, non-parametric techniques, such as permutation tests, can be of great value if the distribution of the data is unknown.

2.1 Permutation Tests

Suppose we have chosen an appropriate statistic $\mathcal{T}$ and the acceptable significance level α. Let $\{(\mathbf{x}_k, y_k)\}_{k=1}^{l}$ be the set of examples, and $\mathbb{Z}_l$ be a set of all permutations of indices $1, \ldots, l$. The permutation test procedure that consists of M iterations is defined as follows:

- Repeat M times (with index $m = 1$ to M):
  - sample a permutation $\mathbf{z}^m$ from a uniform distribution over $\mathbb{Z}_l$,
  - compute the statistic value $t^m = \mathcal{T}(\mathbf{x}_1, y_{z_1^m}, \ldots, \mathbf{x}_l, y_{z_l^m})$.
- Construct an empirical cumulative distribution

$$\hat{P}(T \leq t) = \frac{1}{M} \sum_{m=1}^{M} \Theta(t - t^m),$$

  where Θ is a step-function ($\Theta(x) = 1$, if $x \geq 0$; 0 otherwise).
- Compute the statistic value for the actual labels, $t_0 = \mathcal{T}(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_l, y_l)$ and its corresponding p-value p_0 under the empirical distribution $\hat{P}$.
- If $p_0 \leq \alpha$, reject the null hypothesis.

The procedure computes an empirical estimate of the cumulative distribution of the statistic $\mathcal{T}$ under the null hypothesis and uses it for hypothesis testing. Since the null hypothesis assumes that the two classes are indistinguishable with respect to the selected statistic, all the training data sets generated through permutations are equally likely to be observed under the null hypothesis, yielding the estimates of the statistic for the empirical distribution. An equivalent result is obtained if we choose to permute the data, rather than the labels, in the procedure above. Ideally, we would like to use the entire set of permutations $\mathbb{Z}_l$ to construct the empirical distribution $\hat{P}$, but it might be not feasible for computational reasons. Instead, we resort to sampling from $\mathbb{Z}_l$. It is therefore important to select the number of sampling iterations M to be large enough to guarantee accurate estimation. One solution is to monitor the rate of change in the estimated distribution and stop when the changes are below an acceptable threshold.

To better understand the difference between the parametric approach of the t-test and the permutation testing, observe that statistical significance does not provide an absolute measure of how well the observed differences between the sample groups will generalize, but it is rather contingent upon certain assumptions about the data distribution in each class $p(\mathbf{x}|y)$ being true. The t-test

assumes that the distribution of data in each class is Gaussian, while the permutation test assumes that the data distribution is adequately represented by the sample data. Neither estimates how well the sample data describe the general population, which is one of the fundamental questions in statistical learning theory and is outside the scope of this paper.

3 Statistical Significance in Classification

The permutation tests can be used to assess statistical significance of the classifier and its performance using the test error as a statistic that measures dissimilarity between two populations. Depending on the amount of the available data, the test error can be estimated on a large hold-out set or using cross-validation in every iteration of the permutation procedure. The null hypothesis assumes that the relationship between the data and the labels cannot be learned reliably by the family of classifiers used in the training step. The alternative hypothesis is that we can train a classifier with small expected error.

We use the permutations to estimate the empirical cumulative distribution of the classifier error under the null hypothesis. For any value of the estimated error e, the appropriate p-value is $\hat{P}(e)$ (i.e., the probability of observing classification error lower than e). We can reject the null hypothesis and declare that the classifier learned the (probabilistic) relationship between the data and the labels with a risk of being wrong with probability of at most $\hat{P}(e)$.

To underscore the point made in the previous section, the test uses only the available data examples to evaluate the complexity of the classification problem, and is therefore valid only to the extent that the available data set represents the true distribution $p(\mathbf{x}, y)$. Unlike standard convergence bounds, such as bounds based upon VC-dimension, the empirical probability distribution of the classification error under the null hypothesis says nothing about how well the estimated error rate will generalize. Thus permutation tests provide a weaker guarantee than the convergence bounds, but they can still be useful in testing if the observed classification results are likely to be obtained by random chance.

Note that the estimated empirical distribution also depends on the classifier family used in the training step. It effectively estimates the expressive power of the classifier family with respect to the training data set. The variance of the empirical distribution $\hat{P}$ constructed by the permutation test is in inverse relationship with the difficulty of the classification problem. Consequently, the same accuracy value is more likely to be significant for more complex problems because a classifier trained on the permuted data set is unlikely to perform well on the test set by chance.

3.1 Related Work

As we mentioned in the introduction, permutation tests have been used before in statistical studies of neuroanatomy and function [1,15,16]. In such studies, the anatomical labels or the functional signal intensity in each voxel were tested

for significance and the permutation tests were used to replace the Student distribution with the non-parametric distribution that was constructed for each pixel separately from the images. Consequently, this approach addresses the concern that the value distribution at each voxel was not necessarily Gaussian under the null hypothesis. In addition, non-parametric tests were derived for identifying statistically significant spatially contiguous clusters of voxels that exceeded a particular threshold. Testing for clusters of voxels implicitly accounts for dependencies among the values in neighboring voxels and therefore represents an important advance towards global significance assessment.

Our approach extends the non-parametric testing of statistical significance to the entire ensemble of the features extracted from the images by utilizing the classification framework to measure the predictive power of the feature set with respect to the given labels.

4 Example Applications

We illustrate the use of permutation testing in application to two different examples: a study of changes in the cortical thickness due to Alzheimer's disease and a comparison of brain activation, as measured by fMRI, in response to different visual stimuli.

In all experiments reported in this section, we used linear Support Vector Machines [19] to train a classifier[3], and jacknifing (i.e., sampling without replacement) for cross-validation. The number of cross-validation iterations was 1,000, and the number of permutation iterations was 10,000. All data sets, both training and testing, were perfectly balanced, containing equal number of examples from each class.

4.1 Cortical Thickness Study

In this study, we compare the thickness of the cortex in 50 patients diagnosed with dementia of the Alzheimer type (DAT) and 50 normal controls of matched age. The gray/white matter interface and the pial surface were automatically segmented from each MRI scan [3,7,8], followed by a registration step that brought the surfaces into correspondence by mapping them onto a unit sphere while minimizing distortions and then non-rigidly aligning the cortical folding patterns [5, 6]. The cortical thickness was densely sampled at the corresponding locations for all subjects. And while visualization of the detected differences and understanding the physiological implications of the detected differences is of great interest and is the topic of our current research, we limit our report here to the statistical guarantees that are the focus of this paper.

We first study the classification error with the goal of identifying a sufficient training set size for reliable detection of differences between the two groups. We

[3] A linear classifier function $f(\mathbf{x}) = \langle \mathbf{x}.\mathbf{w} \rangle + b$, where $\mathbf{w}$ is the normal to the separating hyperplane, also called a projection vector, and b is the offset.

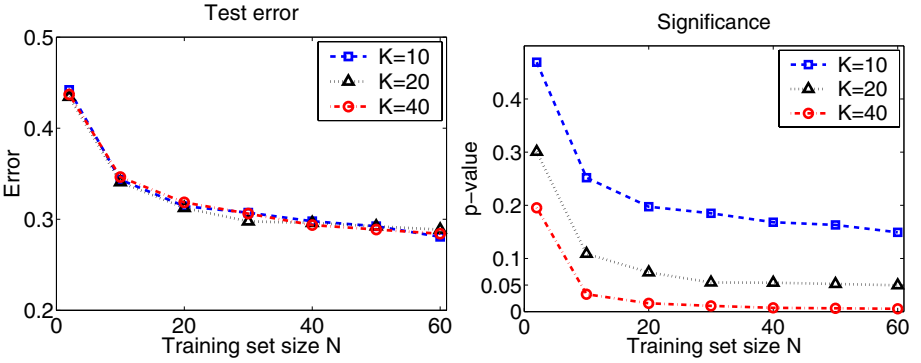

Fig. 1. Cross-validation error (left) and statistical significance (right) for the cortical thickness study for different training set size N and test set size K.

gradually increase the training set size and estimate the test error and statistical significance, reported in Figure 1. Every point in the plots is characterized by the corresponding training set size N and test set size K. For each such pair, we ran 1,000 iterations of cross-validation, selecting randomly $N + K$ examples from the original data set ($N/2 + K/2$ from each group), used N examples to train a classifier and the remaining K examples to test it. The graph on the left shows the average test error over 1,000 cross-validation iterations. By examining the plots, we conclude that at approximately $N = 40$, the performance saturates at 71% ($e = 0.29$). In addition, we ran 10,000 iterations of permutations. In each iteration, we randomly selected $N + K$ examples, arbitrarily relabeled them, while maintaining an even distribution of labels, and performed 1,000 iterations of the cross-validation procedure (training on N examples and testing of the remaining K examples) on this newly labeled data set. The graph on the right shows p-values estimated for various training and test set sizes.

It is not surprising that increasing the number of training examples improves the robustness of classification as exhibited by both the accuracy and the significance estimates. Increasing the number of independent examples on which we test the classifier in each iteration does not significantly affect the estimated classification error, but substantially improves the statistical significance of the same error value, as can be seen in Figure 1. This is to be expected: as we increase the test set size, a classifier trained on a random labeling of the training data is less likely to maintain the same level of testing accuracy.

Figure 2 illustrates this point for a particular training set size of $N = 40$. It shows the empirical distribution $\hat{P}(e)$ curves for the test set sizes $K = 10, 20, 40$. The right graph zooms on the small framed area on the left graph. The filled circles represent classification performance on the true labels and the corresponding p-values. We note again that the three circles represent virtually same accuracy, but substantially different p-values. For this training set size, if we set the significance threshold at $\alpha = 0.05$, testing on $K = 20$ examples is sufficient

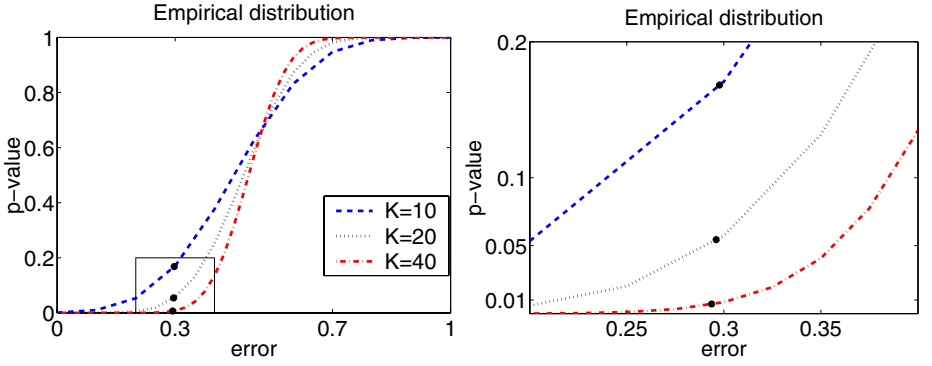

Fig. 2. Empirical distribution $\hat{P}(e)$ of the classifier error estimated using permutation tests in a cross-validation procedure (left); the small framed area of the graph is shown at higher resolution on the right. The size of the training set in all experiments is $N = 40$. Filled circles indicate the classifier performance on the true labels for the corresponding training and test set sizes ($N = 40$, $K = 10$: $e = 0.30$, $\hat{P}(e) = 0.17$; $N = 40$, $K = 20$: $e = 0.29$, $\hat{P}(e) = 0.05$; $N = 40$, $K = 40$: $e = 0.29$, $\hat{P}(e) = 0.007$).

to establish significance. Testing on $K = 40$ independent leads to $p < 0.007$, achieving significance under a much more strict threshold of $\alpha = 0.01$.

To gain an additional insight into these results, let's consider the classical situation of two-sided t-testing, with the data generated by two one-dimensional Gaussian densities. Our goal is to test the null hypothesis that postulates that the two distributions have identical means. There are two factors affecting statistical significance: the distance between the means and the amount of data we have to support it. We can achieve low p-values in a situation where the two classes are very far apart and we have a few data points from each group, or when they are much closer, but we have substantially more data. P-value by itself does not indicate which of these two situations is true. Using discriminative analysis allows us to estimate how far apart the two classes are[4], in addition to establishing statistical significance of the detected differences.

4.2 Categorical Differences in fMRI Activation Patterns

This experiment compares the patterns of fMRI activations in response to different visual stimuli in a single subject. We present the results of comparing activations in response to face images to those induced by house images, as these categories are believed to have special representation in the cortex. The comparison used 15 example activations for each category (for details on data acquisition, see [18]). The fMRI scans were aligned to the structural MRI of the same subject using rigid registration. For each scan, the cortical surface was

[4] In this particular study, the classification accuracy saturates at 71%, indicating there is a substantial overlap between the two classes.

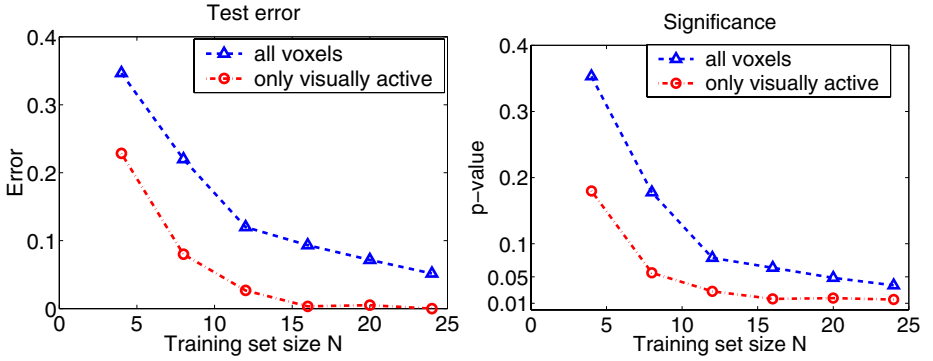

Fig. 3. Cross-validation error (left) and statistical significance (right) for two different feature sets in the fMRI study.

extracted using the same techniques as in the study of cortical thickness, and the average activation values were sampled along the cortical surface. A surface-based representation models the connectivity of the cortex and is therefore well suited for studies of fMRI activation of cortical regions.

First, we used all voxels in the fMRI scan to perform the comparison between the two categories. We then repeated the experiments with only the "visually active" region of the cortex. The mask for the visually active voxels was obtained using a separate visual task. The goal of using the mask was to test if removing irrelevant voxels from consideration improves the classification performance.

When the training and the test set sizes were varied in this experiment, we observed similar trends in the test error and significance behavior to those discussed in the previous section. This study contains substantially fewer examples than the previous one, forcing us to work with smaller test sets. Here, we report the results for the test set size $K = 6$, as it was the smallest test set size for which we we could demonstrate statistical significance and it allows testing over a wider range of training set sizes. Figure 3 reports the accuracy and the p-value estimates for both feature sets. As expected, reducing the feature set to locations relevant to the visual pathways increases the classification accuracy and statistical significance of the detected differences. For example, for training size $N = 20$, using the entire cortex yields 93% accuracy ($e = 0.07, p < 0.05$), while using only visually active voxels improves the accuracy to 100% ($e = 0, p < 0.02$).

In contrast to the cortical thickness study, we achieve statistical significance with substantially fewer training and testing examples in this experiment. The two classes in this study are so far apart in the feature space that the learning problem is much easier, requiring fewer examples to achieve robust detection.

One way to visualize the detected differences is to display the projection vector $\mathbf{w}$. Each original feature x_i is assigned weight w_i by the training algorithm. The magnitude of the weight is indicative of the predictive power of the corresponding feature. We can visualize $\mathbf{w}$ by displaying it in the same way we

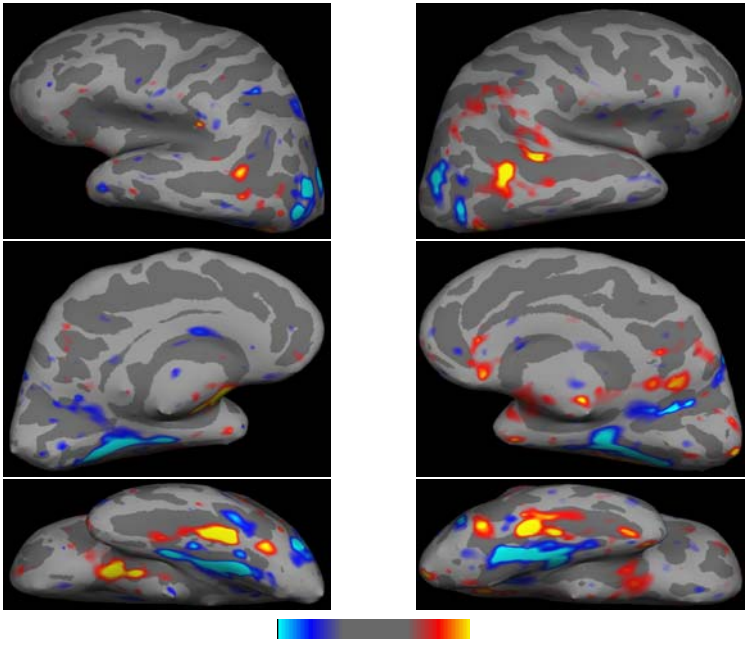

Fig. 4. Weight maps for the face class (positive class) in comparison with the house class (negative class). The six views show (top-to-bottom) lateral, medial and inferior views of the left and the right hemispheres. The color is used to indicate the weight of each voxel, from light blue (negative) to yellow (positive).

visualize the original feature vectors $\mathbf{x}_k$. Figure 4 and Figure 5 show the two feature sets in the fMRI study by paining the weights w_i onto the inflated cortical surface. The grayscale pattern shows cortical folding, while the color is used to indicate the magnitude and the sign of the weights. The weights have been thresholded for visualization purposes, leaving only the areas of high weighting painted. If used on "visual only" set of features, the method produces a map that is very similar to the corresponding subset of the one for the entire cortex, indicating robustness of the estimation. One of our current goals is to develop an automated feature selection mechanism capable of detecting relevant locations and reducing the feature set to include just those areas.

5 Conclusions

In this paper, we adapt the permutation tests for estimation of statistical significance of a classifier function. We demonstrate the tests on several examples of neuroimaging studies comparing two or more sets of anatomical or functional scans. The test is useful in experiments for which the standard convergence bounds fail to produce meaningful guarantees due to the high dimensionality of the input space and the extremely small sample size. We hope other researchers

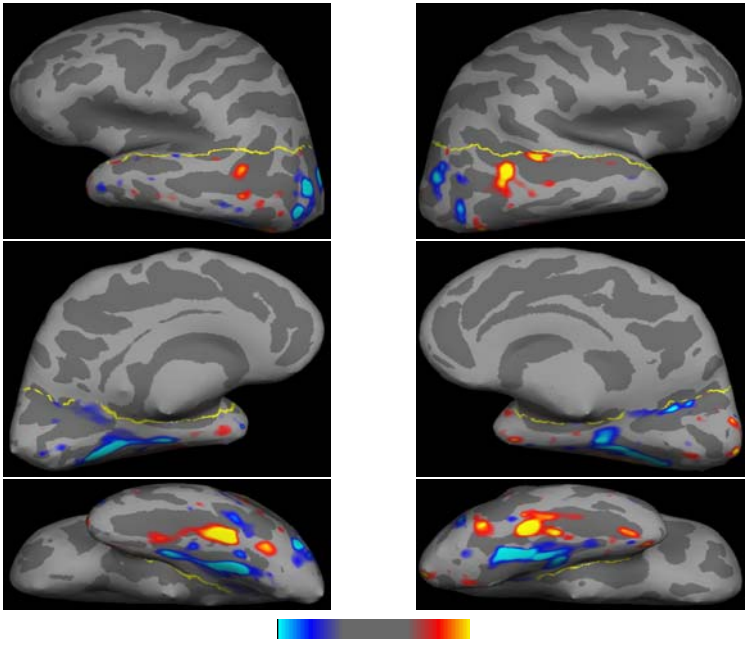

Fig. 5. Weight maps for the face class (positive class) in comparison with the house class (negative class) using just the visually active area of the cortex (posterior). The boundary of the "visually active" mask is shown in yellow. The six views show (top-to-bottom) lateral, medial and inferior views of the left and the right hemispheres. The color is used to indicate the weight of each voxel, from light blue (negative) to yellow (positive).

in the community will find the technique useful in assessing statistical significance of observed results when the data are high dimensional and are not necessarily generated by a normal distribution.

Acknowledgements. This research was supported in part by NSF IIS 9610249 grant, Athinoula A. Martinos Center for Biomedical Imaging collaborative research grant and NIH R01 RR16594-01A1 grant. The Human Brain Project/Neuroinformatics research is funded jointly by the NINDS, the NIMH and the NCI (R01-NS39581). Further support was provided by the NCRR (P41-RR14075 and R01-RR13609). The authors would like to acknowledge Dr. M. Spiridon and Dr. N. Kanwisher for providing the fMRI data, Dr. R. Buckner for providing the cortical thickness data and Dr. D. Greve for help with registration and feature extraction in the experiments discussed in this paper. Dr. Kanwisher would like to acknowledge EY 13455 and MH 59150 grants. Dr. Buckner would like to acknowledge the assistance of the Washington University ADRC, James S McDonnell Foundation, the Alzheimer's Association, and NIA grants AG05682 and AG03991.

References

1. E.T. Bullmore, *et al.* Global, Voxel, and Cluster Tests, by Theory and Permutation, for a Difference Between Two Groups of Structural MR Images of the Brain. *In IEEE Transactions on Medical Imaging*, 18(1):32–42, 1999.
2. J. G. Csernansky, *et al.* Hippocampal Morphometry in Schizophrenia by High Dimensional Brain Mapping. *In Proceedings of National Academy of Science*, 95(19):11406–11411, 1998.
3. A. M. Dale, *et al.* Cortical Surface-Based Analysis I: Segmentation and Surface Reconstruction. *NeuroImage*, 9:179–194, 1999.
4. B. Efron. The Jacknife, The Bootstrap, and Other Resampling Plans. *SIAM*, Philadelphia, PA. 1982.
5. B. Fischl, *et al.* Cortical Surface-Based Analysis II: Inflation, Flattening, a Surface-Based Coordinate System. *NeuroImage*, 9:195–207, 1999.
6. B. Fischl, *et al.* High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8:272–84, 1999.
7. B. Fischl, *et al.* Measuring the thickness of the human cerebral cortex from magnetic resonance images. *In PNAS*, 26:11050–5, 2000.
8. B. Fischl, A. Liu, A.M. Dale. Automated Manifold Surgery: Constructing Geometrically Accurate and Topologically Correct Models of the Human Cerebral Cortex. *In IEEE Transactions on Medical Imaging*, 20(1):70–80, 2001.
9. G. Gerig, *et al.* Shape versus Size: Improved Understanding of the Morphology of Brain Structures. *In Proc. MICCAI'2001*, LNCS 2208, 24–32, 2001.
10. P. Golland, *et al.* Discriminative Analysis for Image-Based Studies. *In Proc. MICCAI'2002*, LNCS 2488:508–515, 2002.
11. P. Good. Permutation Tests: A Practical guide to Resampling Methods for Testing Hypothesis. *Springer-Verlag*, 1994.
12. I. Guyon, *et al.* What Size Test Set Gives Good Error Rate Estimates? *In IEEE Trans. Pattern Analysis and Machine Intelligence.* 20(1): 52–64, 1998.
13. J. V. Haxby, *et al.* Distributed and Overlapping Representations of Faces and Objects In Ventral Temporal Cortex. *Science*, 293:2425–2430, 2001.
14. J. Martin, A. Pentland, and R. Kikinis. Shape Analysis of Brain Structures Using Physical and Experimental Models. *In Proceedings of CVPR'94*, 752–755, 1994.
15. T.E. Nichols and A.P. Holmes. Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 15:1–25, 2001.
16. P. M. Thomson, *et al.* Dynamics of Gray Matter Loss in Alzheimer's Disease. *Journal of Neuroscience*, 23(3), 2003.
17. L. Sachs. Applied Statistics: A Handbook of Techniques. Springer Verlag. 19984.
18. M. Spiridon and N. Kanwisher. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron*, 35(6):1157–1165, 2002.
19. V. N. Vapnik. Statistical Learning Theory. *John Wiley & Sons*, 1998.