Accounting for Heterogeneity across Multiple Imaging Sites using Multi-Task Learning

Michelle T. Hromatka[1], Wei Liu[1], Jeffrey S. Anderson[2], Brandon A. Zielinski[3], Molly B. DuBray[4], and P. Thomas Fletcher[1]

[1] School of Computing
[2] Division of Neuroradiology
[3] Division of Child Neurology
[4] Interdepartmental Program in Neuroscience
University of Utah, Salt Lake City, UT 84112, USA

Abstract. Combining imaging data from multiple sites has the potential to increase statistical power in clinical studies. However, the increase in sample size is accompanied by an increase in variance due to inconsistencies across sites, e.g., different scanners, protocols, and demographics. In this paper, we present an approach for combining multi-site imaging data in classification tasks that takes this heterogeneity into account. The idea is to treat the classification problem as a multi-task learning problem, where each imaging site is treated as a "task". We employ a regularized support vector machine (SVM) that allows for differences in decision boundaries at individual sites, while at the same time leveraging the similarities in the decision boundaries across sites. We demonstrate the effectiveness of this approach in the classification of autism from multi-site functional magnetic resonance imaging (fMRI) from the Autism Brain Imaging Data Exchange (ABIDE). The proposed method achieves state-of-the-art accuracy and outperforms a comparable SVM classifier applied to pooled data as well as individual SVM classifiers applied per site.

1 Introduction

The last few years have seen an increase in multi-site imaging studies for diseases like Alzheimer's to spectrum disorders like Autism. The basis for these aggregate datasets is to accelerate knowledge of these diseases and disorders by providing larger sample sizes and making access to data readily available. Analysis using these datasets, however, is not that straightforward due to differences across site scanners, protocols, populations and diagnosis techniques(?). This variability introduces extra parameters that must be accounted for when using the aggregate data, but a standard way to deal with this variability has not been established.

In addition to the reasons cited above, many sites do not have a large enough sample size to use learning algorithms on site specific data. Meta-analysis, in which results across small studies are combined to extract patterns common in each, has previously been used to combine site-specific results, especially when the sample size is low. However, meta-analysis is not free of subjectivity of data variability, thus this method of combined analysis is also faulty [5].

Several groups have used the ABIDE dataset for classification with different, non-meta-analytic approaches. Nielsen, et al. combines the ABIDE dataset with a whole-brain approach, using a leave-one-out classifier to compute a classification score for each left-out subject based on age, gender and handedness. The correlations for each connection in turn were fit with a linear model, separating controls from ASDs, which was then adjusted by the difference between the subject's site mean for that connection and the overall mean. This approach yielded a maximum overall accuracy of 60.0% despite finding significant positive correlation between the classification score and several of the phenotypic behavioral measures [10]. A different study used histogram of gradients and applied this to several multi-site imaging studies which was able to achieve 61.7% accuracy on the ABIDE dataset and 62.6% on the ADHD-200 dataset [6]. While [10] accounted for some of the site differences, both studies approached the differences in population, imaging parameters and ******** as noise, instead of extra data that can be used when classifying an aggregate data set. This is the key principle in multi-task learning: differences between tasks can be accounted for while using a common mean to account for similarities between the different tasks. We decided to use this approach within an SVM classifier for the ABIDE dataset.

multi site imaging studies, aggregating data across sites with the goal that large amounts of data will *********** however This has led to an increase over the last several years in online data consortia like the ADHD-200, the 1000 Functional Connectomes project and the Autism Brain Imaging Data Exchange(ABIDE) [2] [1] [4]. These multi-site imaging studies have allowed researchers to take advantage of an increased sample size ************8 Many researchers fail to take these differences as a****8, but rather as increased noise when treating with the drawback of increased noise due to scanner, protocol and population differences.

2 Methods

The data is split evenly amongst individual sites into two sets; a third is used for training and the remaining two thirds for testing. The training set is used to remove nuisance factors from the data (e.g. age), SVM parameters through cross-validation and the features to be used in classification (see 2.2). These parameters are then used in a leave-one-out classifier on the testing data to determine the algorithm's accuracy, sensitivity and specificity. It is important to note that all parameter selection and feature selection is done exclusively on the *training* set, and the testing set is only used in the final classification to avoid inflation of results.

2.1 Multi-Task Learning

Evgeniou and Pontil (2004) introduces a method of multi-task learning based on kernel based methods typically used for single task learning. This method relies on minimizing regularization functions, such as that for SVM, to capture both overall similarity between tasks and individual task differences. The traditional minimization for a soft margin SVM is:

$$\frac{1}{2}w^2 + CF(\sum_{i=1}^{t}\xi_i) \tag{1}$$

where C is a constant and $F(\mu)$ is a "monotonic convex function with $F(0) = 0$ [3]. In the case of SVMs, the weight vector w is used to define the hyperplane, $(w \cdot x + b)$, which is the boundary between groups.

For multi-task learning, the relationship between T tasks must be described, which Evgeniou and Pontil approach using the hierarchical Bayes method. This assumes that each task function comes from a class of probability distributions ******more****. The relationship is defined as:

$$w_t = w_0 + v_t, \tag{2}$$

where w_0 is the mean of the data and each task t has its own weight vector, v_t. Multi-task learning allows for simultaneous learning of the mean of all tasks, w_0, and each task weight vector, v_t, so the minimization function then becomes:

$$C\sum_{t=1}^{T}\sum_{i=1}^{m}\xi_{it} + \frac{\lambda_1}{T}\sum_{t=1}^{T}||v_t||^2 + \lambda_2||w_0||^2, \tag{3}$$

where λ_1, λ_2 are "positive regularization parameters" and C is still a constant. For high similarity between tasks, the v_t will be small in relation to w_0; this relationship is described by the hyperparameters λ_1, λ_2 that must be chosen by the user.

The dual of equation 3 can be found by defining a set of functions $f_t(x) = w_t \cdot x$ which can be simplified to $F(x,t) = f_t(x)$. This can be described by a kernel function $\phi((x,t))$ which allows us to relate the dual of a multi-task learning problem to the dual of Equation 1.

$$\max_{\alpha_{it}}\left\{\sum_{i=1}^{m}\sum_{t=1}^{T}\alpha_{it} - \sum_{i=1}^{m}\sum_{s=1}^{T}\sum_{j=1}^{m}\sum_{t=1}^{T}\alpha_{is}y_{is}\alpha_{jt}y_{jt}\phi((x,t))\right\} \tag{4}$$

where

$$\phi((x,t)) = \big(\frac{x}{\sqrt{\mu}}, \underbrace{0,...,0}_{t-1}, x, \underbrace{0,...,0}_{T-t}\big), \quad \textbf{for } \mu = \frac{T\lambda_2}{\lambda_1}. \tag{5}$$

As you can see in Equation 4, this is the same dual problem as for a single task-SVM, with the data transformed by $\phi((x,t))$ into the multi-task kernel space.

2.2 Feature Selection

Data extraction in imaging studies typically leads to very high dimensional data spaces. For f-MRI, a typical choice for data is the pairwise correlation between n

predefined regions of the brain. This yields a dataspace of $\frac{n(n+1)}{2}$ dimensionality, which, even for a relatively small number of regions, can be computationally expensive. The multi-task learning above further increases dimensionality with the number of tasks; the method described would yield a feature space of $(t+1)d$ dimensions, where d is the dimension of the feature space and t the number of tasks. Thus feature selection can be employed to remove redundancy and increase relevancy of the data while reducing computation time [7].

A ****** approach is to use a simple hypothesis test to determine which features would be most useful in classification. Nuisance factors, such as age or ******, should be accounted for prior to the hypothesis test to ideally isolate differences attributed to the disease or disorder being studied. A linear model is then fit to the training set one feature at a time based on group. If the coefficient for the group has a $p < .001$, the feature is kept and used in the SVM classification. This test is done on each feature in turn, ultimately reducing the data into a set that has significant differences between groups.

3 Evaluation

3.1 Data

The Autism Brain Imaging Data Exchange(ABIDE) database is an online consortium of resting-state functional-MRI data from 17 international sites, resulting in brain imaging data for 539 individuals with ASD and 573 typically developing(TD) controls [4]. All ASD subjects were diagnosed by either the Autism Diagnosis Observation Schedule-General(ADOS-G) or the Autism Diagnostic Interview-Revised tests and removed from the study if other co-morbid disorders were present [8] [9] [4]. Further inclusion details can be found at (put url? citation to... website? or the abide paper?)

Preprocessing All data was preprocessed using the Functional Connectomes-1000 preprocessing scripts [1]. This includes:

1. MRI Deoblique, reorient, skull strip
2. f-MRI Reorient, motion correct, skull strip, smooth
3. registration
4. Segmentation - csf, white matter
5. extracting global signal, from csf and wm
6. extract time series, Z-transform correlations
7. spatial smoothing, register to atlas
8. some sort of regression

Twelve subjects were removed because of failure during the preprocessing. Two Oregon subjects were missing the resting fMRI file and 10 UCLA subjects were missing the anatomical scan file which is required in step 1 of the preprocessing pipeline above. This resulted in 1100 subjects for analysis, 530 ASD and 570 TD controls.

Data Extraction From each subject's postprocessed image, the time series for each of 264 regions is extracted based on Power's regions of interest [11]. These 264 regions are spread out among the cerebral cortex, subcortical structures and cerebellum, where each region is a sphere of 5mm in radius and regions are separated by a minimum distance of 10mm so as to avoid detection of a shared signal. The Fisher transformed Pearson correlation coefficient is then found between each region and the other 263 regions, resulting in a 34,716 dimensional feature space for each subject. After feature selection, this number is reduced to 312 features per subject. ***add in featsel, multiplication/interdependence of features??

3.2 Results

4 Discussion and Conclusion

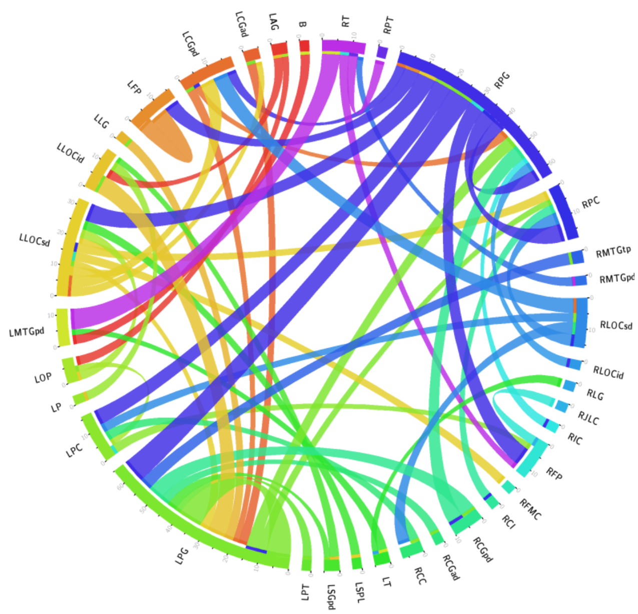

Fig. 1. The pairwise connections, represented as ribbons connecting each region, selected by the hypothesis test described in 2.2. The width of each ribbon is determined by the weights from the w_0 vector in the SVM.

References

1. Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan

Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

2. ADHD-200 Consortium et al. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6, 2012.
3. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
4. A Di Martino, CG Yan, Q Li, E Denio, FX Castellanos, K Alaerts, JS Anderson, M Assaf, SY Bookheimer, M Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 2013.
5. Hans J Eysenck. Meta-analysis and its problems. *BMJ: British Medical Journal*, 309(6957):789, 1994.
6. Sina Ghiassian, Russell Greiner, Ping Jin, and Matthew RG Brown. Learning to classify psychiatric disorders based on fmr images: Autism vs healthy and adhd vs healthy.
7. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
8. Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook Jr, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3):205–223, 2000.
9. Catherine Lord, Michael Rutter, and Ann Le Couteur. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, 24(5):659–685, 1994.
10. Jared A Nielsen, Brandon A Zielinski, P Thomas Fletcher, Andrew L Alexander, Nicholas Lange, Erin D Bigler, Janet E Lainhart, and Jeffrey S Anderson. Multisite functional connectivity mri classification of autism: Abide results. *Frontiers in human neuroscience*, 7, 2013.
11. Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.