Accounting for Heterogeneity Across Multiple Imaging Sites Using Multi-Task Learning

No Author Given

No Institute given

Abstract. Combining imaging data from multiple sites has the potential to increase statistical power in clinical studies. However, while pooling multiple data sources increases sample size, it also increases unwanted 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

Recent years have seen a movement towards combining neuroimaging data collected across multiple sites. Such multi-site data has the potential to accelerate scientific discovery by increasing sample sizes, providing broader ranges of participant demographics, and making data publicly available. Different approaches include large, coordinated multi-site neuroimaging studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) [14], as well as data sharing initiatives that combine multiple single-site studies, such as the Autism Brain Imaging Data Exchange (ABIDE) [5]. Analysis using these datasets, however, is not straightforward, due to differences across site scanners, protocols, populations, and diagnosis techniques. Treating a multi-site study as a single, homogeneous data set fails to account for this variability, which can be detrimental to the statistical power and counteract the gains made by increasing the sample size.

An alternative approach is to perform separate statistical analyses at individual sites and combine the results in a meta-analysis. This is often formulated as a random effects model [4], where each site is regarded as a random treatment effect. While this can be an effective way to combine statistical tests of low-dimensional summary measures, it is less applicable to learning problems on high-dimensional data such as images. For instance, applying independent classifiers at each site and then combining them post-hoc misses the opportunity that the classifiers could benefit from sharing information across sites during learning. Meta-analysis is also prone to publication bias, meaning only results from studies whose results were significant enough to publish are aggregated. By using shared raw data instead of widely published results, this bias can ideally be avoided [8].

In this paper, we propose an approach for combining multiple imaging studies in classification problems. The idea is to treat the problem as a multi-task learning problem, where classifiers for each site are estimated jointly, with a regularization that favors similarity across sites. This allows classifiers to share information during learning, but also provides the flexibility for them to differ at each site as needed. This is the key principle in multi-task learning: heterogeneity across similar tasks (or, in our case, sites) can be accounted for while using a common mean to account for similarities between the different tasks. We specifically employ a regularized support vector machine (SVM) introduced by Evgeniou and Pontil [7]. As a driving problem, we apply this method to the classification of autism from functional magnetic resonance imaging (fMRI) from the ABIDE database.

Several groups have reported classification results using the ABIDE data, in each case treating the pooled collection of images across all sites as a single homogeneous dataset during classification. Nielsen, et al. [15] 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 subjects with an Autism Spectrum Disorder (ASD), 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 [15]. 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 [9]. While [15] accounted for the site differences during feature generation and selection, both studies approached the differences across imaging sites as noise instead of extra data that can be leveraged when classifying an aggregate data set.

2 Methods

We formulate classification of multi-site imaging data as a multi-task learning problem, where each site, $s = 1, \dots, S$, is treated as a separate task. This results in a different classifier for each site, which allows for variability of decision boundaries across sites. At the same, a regularization term in the objective function favors similarities between sites, resulting in sharing of data across sites during training. We specifically use a regularized version of SVM, introduced by Evgeniou and Pontil [7], which we describe next. Following that, we describe a process for feature selection in the case of resting state fMRI, which is an important step to reduce the high dimensionality of the data.

2.1 Multi-Task Learning

Evgeniou and Pontil [7] introduce a method of multi-task learning based on kernel 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. Given N feature vectors $x_i \in \mathbb{R}^d$ with labels $y_i \in \{-1, 1\}$, the traditional minimization for a soft margin SVM [3] is

$$\arg\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i, \qquad \text{subject to: } y_i(w \cdot x_i - b) \geq 1 - \xi_i,\ \xi_i \geq 0, \quad (1)$$

where the weight vector w defines the hyperplane, $(w \cdot x + b)$, which is the boundary between groups, $C \in \mathbb{R}$ is a constant, and $\xi_i \in \mathbb{R}$ are slack variables.

For multi-task learning, the relationship between S tasks (or sites, in our case) must be described, which can be framed as a hierarchical model. This assumes that each task function comes from a class of probability distributions. The relationship is defined as

$$w_s = w_0 + v_s, \quad (2)$$

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

$$C\sum_{s=1}^{S}\sum_{i=1}^{m} \xi_{is} + \frac{\lambda_1}{S}\sum_{s=1}^{S} ||v_s||^2 + \lambda_2||w_0||^2, \quad (3)$$

where λ_1, λ_2 are positive regularization parameters, and C is still a constant. For high similarity between tasks, the v_s 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 (3) can be described using a feature mapping $\phi((x, s))$, which allows us to relate the dual of a multi-task learning problem to the dual of (1) as

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

where

$$\phi((x, s)) = \Big(\frac{x}{\sqrt{\mu}}, \underbrace{0, ..., 0}_{s-1}, x, \underbrace{0, ..., 0}_{S-s} \Big), \quad \text{for } \mu = \frac{S\lambda_2}{\lambda_1}. \quad (5)$$

As can be seen in (4), this is the same dual problem as for a single-task SVM, with the data transformed by $\phi((x,s))$ into the multi-task feature space. This can be implemented as a kernel method, without explicitly computing the higher-dimensional feature vectors $\phi((x,s))$, since only inner products between features are needed in the SVM optimization.

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. For a feature space $x \in \mathbb{R}^d$, the multi-task kernel $\phi((x,t))$ would yield a feature space of $d(t+1)$ dimensions. Feature selection can and should be employed to remove redundancy and increase relevancy of the data while reducing computation time [10].

One approach is to use the values of the weight vector w to choose the most relevant features. Recall that the decision boundary in an SVM is defined by $w \cdot x$, meaning that the highest magnitudes in the weight vector denote the features that best define the decision boundary between groups. This is the basis for the SVM recursive feature elimination(SVM-RFE) method demonstrated in [11], [5], [6]: an iterative process where a user-specified number of features corresponding to the lowest w values are removed from the feature set after each iteration. We modify SVM-RFE to better suit the f-MRI pairwise correlation data by ranking the features not by the associated w value of each pairwise correlation, but by the $l2$-norm of an entire region's pairwise correlation w vector values. This means for each region, r, we extract the w values for all pairwise correlation datapoints involving region r and find the $l2$-norm of these w values. This is done for all regions in the dataset which are then ranked and a user-specifed number of regions corresponding to the lowest associated $l2$-norms are removed from the dataset after each iteration. This approach effectively reduces the data dimension for feature selection purposes while still leveraging the correlation data.

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 [5]. 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 [12] [13] [5]. Further inclusion details can be found in [5].

Preprocessing All data was preprocessed using the Functional Connectomes-1000 preprocessing scripts which includes skull stripping, motion correcting, registration, segmentation and spatial smoothing [2]. Twelve subjects were removed because of failure during the preprocessing. Two OHSU subjects were missing the resting fMRI file and 10 UCLA subjects were missing the anatomical scan file which is required in the preprocessing pipeline. 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 [16]. 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 all other 263 regions, resulting in a 34,716 dimensional feature space for each subject. After feature selection, this number was reduced to 74 regions of the original 264, yielding a final feature space of 2701 features.

3.2 Results

We split the ABIDE data into two sets; one half of the data is used for training and the other half is used exclusively to test our approach. This split is performed at the site level, randomly removing half of the ASD subjects and half of the TD subjects per site, and then aggregating these subjects into a testing set to preserve the ratios between sites and groups across the two data sets.

We present results for three cases: one where each site is classified individually, a second where all site data is pooled and considered a single site and the third using the multi-task learning approach described in 2.1. For all of the results presented within, a linear kernel is used in the SVM, with the error term parameter, C, determined by cross validation on the training set. Multi-task learning requires an additional parameter, μ, which is also determined through cross validation on the training set. Additionally, the training set is used to determine nuisance factor regression on subject age, and feature selection as described in 2.2. It is important to note that all parameter tuning, feature selection, and nuisance factor regression was performed exclusively on the training set. Involving the testing set in any of these procedures can bias the classifier and inflate the results.

The multi-task learning approach achieved significantly better overall accuracy than the single task approach, which in turn improved upon the individual site classifiers. Using MTL, we classified the ABIDE data with an overall accuracy of 64.9%, with 67% sensitivity and 63% specificity. Pooling all data resulted in 62.9% overall accuracy, 67% sensitivity and 59% specificity. The summed overall accuracy of each individual site was 58.1% with sensitivity of 64% and specificity of 53%.

Several sites benefited greatly from the increase in data size, most notably Caltech, CMU, MaxMun, and Olin. All but one of the sites in the bottom 50%

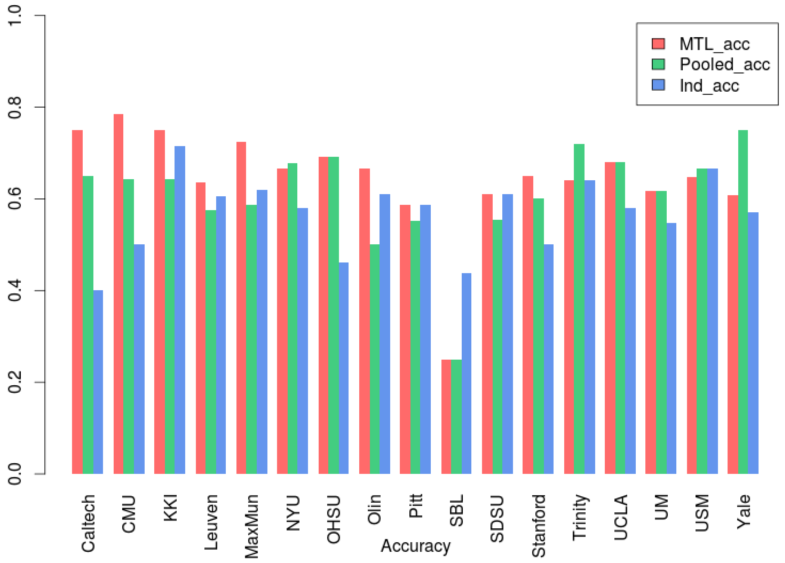

Fig. 1. The overall accuracies by site, as found by each of the three experiments.

for sample size either equaled or improved upon the pooled result using MTL. For sites that were relatively stable to begin with because of large individual site sample size (e.g., NYU, UCLA, UM, USM) the MTL approach found little to no improvement to that site's overall accuracy. This is an expected result of multi-task learning, as the larger sites have sufficient data to design a competent classifier. By requiring these sites to be near the combined site mean, a decrease in accuracy is likely as the decision boundary is deviating from what was already a stable result.

4 Discussion and Conclusion

Figure 2 displays the regions identified via the method described in 2.2. The regions show a strong overlap with ROIs previously identified as network hubs thought to be abnormal in autism, namely the default mode network (regions LCGpf, RCGad, RPC, etc.) and socioemotional salience network (regions RIC, L/RPcG, **). Specifically, these regions include atypical network structure and function, identified in structural covariance MRI [17] and also functional connectivity studies (reviewed in [1]), may contribute to many of the behaviors associated with the disorder.

We presented a novel way to classify multi-site data, specifically neuroimaging fMRI data, in a way that leverages similarity across sites while accounting

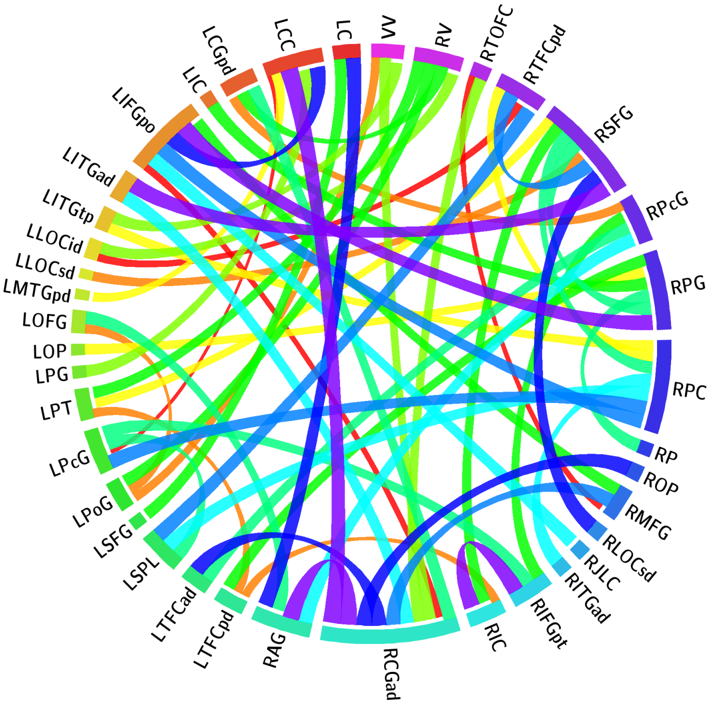

Fig. 2. The most discriminative pairwise connections as selected by the method described in 2.2. The width of each ribbon is determined by the weights from the w_0 vector (i.e., utility) in the SVM.

for individual site differences. Additionally, we introduced a feature selection approach that is better equipped to handle fMRI pairwise correlation data. The utility of these ideas was demonstrated by achieving state-of-the-art classification accuracy on the ABIDE dataset.

References

1. J S Anderson. Cortical underconnectivity hypothesis in autism: Evidence from functional connectivity mri. In V B Patel, V R Preedy, and C R Martin, editors, *Comprehensive Guide to Autism*, pages 1457–1471. Springer New York, 2014.
2. B B Biswal, M Mennes, X-N Zuo, S Gohel, C Kelly, S M Smith, C F Beckmann, J S Adelstein, R L Buckner, S Colcombe, et al. Toward discovery science of human brain function. *PNAS*, 107(10):4734–4739, 2010.
3. C Cortes and V Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

4. R DerSimonian and N Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.
5. 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.
6. C Ecker, V Rocha-Rego, P Johnston, J Mourao-Miranda, A Marquand, E M Daly, M J Brammer, C Murphy, and D G Murphy. Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. *Neuroimage*, 49(1):44–56, 2010.
7. T Evgeniou and M Pontil. Regularized multi–task learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.
8. H J Eysenck. Meta-analysis and its problems. *BMJ: British Medical Journal*, 309(6957):789, 1994.
9. S Ghiassian, R Greiner, P Jin, and M RG Brown. Learning to classify psychiatric disorders based on fmr images: Autism vs healthy and adhd vs healthy.
10. I Guyon and A Elisseeff. An introduction to variable and feature selection. *JMLR*, 3:1157–1182, 2003.
11. I Guyon, J Weston, S Barnhill, and V Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.
12. C Lord, S Risi, L Lambrecht, E H Cook Jr, B L Leventhal, P C DiLavore, A Pickles, and M Rutter. The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism. *J. of autism and developmental disorders*, 30(3):205–223, 2000.
13. C Lord, M Rutter, and A 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.
14. S G Mueller, M W Weiner, L J Thal, R C Petersen, C R Jack, W Jagust, J Q Trojanowski, A W Toga, and L Beckett. Ways toward an early diagnosis in alzheimers disease: The alzheimers disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.
15. J A Nielsen, B A Zielinski, P T Fletcher, A L Alexander, N Lange, Er D Bigler, J E Lainhart, and J S Anderson. Multisite functional connectivity mri classification of autism: Abide results. *Frontiers in human neuroscience*, 7, 2013.
16. J D Power, A L Cohen, S M Nelson, G S Wig, K A Barnes, J A Church, A C Vogel, T O Laumann, F M Miezin, B L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.
17. B A Zielinski, J S Anderson, A L Froehlich, M BD Prigge, J A Nielsen, J R Cooperrider, A N Cariello, P T Fletcher, A L Alexander, N Lange, et al. scmri reveals large-scale brain network abnormalities in autism. *PloS one*, 7(11):e49172, 2012.