Network-Guided Group Feature Selection for Classification of Autism Spectrum Disorder

Veronika Cheplygina[1,3], David M.J. Tax[1], Marco Loog[1,2], and Aasa Feragen[2,3]

[1] Pattern Recognition Laboratory, Delft University of Technology, The Netherlands
[2] The Image Group, University of Copenhagen, Denmark
[3] Machine Learning and Computational Biology Group, Max Planck Institutes Tübingen, Germany

Abstract. We present an anatomically guided feature selection scheme for prediction of neurological disorders based on brain connectivity networks. Using anatomical information not only gives rise to an interpretable model, but also prevents overfitting, caused by high dimensionality, noise and correlated features. Our method selects meaningful and discriminative groups of connections between anatomical regions, which can be used as input for any supervised classifier, such as logistic regression or a support vector machine. We demonstrate the effectiveness of our method on a dataset of autism spectrum disorder, with an AUC of 0.76, outperforming baseline methods.

1 Introduction

The effect of neurological disorders on structural (and functional) brain connectivity can be studied through magnetic resonance imaging (MRI). Studies often focus on population differences between cases and controls for particular global variables, such as white matter volume [1] or global graph-theoretic properties of brain networks [2,3]. However, the *predictive power* of the selected features is often not tested [3,4], and weak statistical tests which can be inconclusive as to which variables of interest are predictive of the diagnosis [5], leading to poor generalization to unseen data. Classification models are therefore more interesting from a diagnostic perspective [6,7].

Furthermore, *interpretability*, in the sense that prediction should link back to concrete biological markers, is a desirable property. For example, global measures such as small-worldness of networks [3] or histograms of image gradient descriptors [8] may disregard local connectivity changes, and do not provide information about which brain pathways have been altered.

This calls for methods which (i) consider local information and (ii) are predictive. Measuring features on densely sampled regions of interest (ROIs) provides local information, but unfortunately the high dimensionality of the noisy, correlated features can easily lead to overfitting, i.e. predicting perfectly on the training data, but failing to generalize to previously unseen data. It is therefore necessary to reduce the dimensionality, either by clustering ROIs [9], selecting

G. Wu et al. (Eds.): MLMI 2014, LNCS 8679, pp. 190–197, 2014.

features with good predictive performance on the training data [6, 10], and/or using classifiers which penalize complex models [9, 11].

It is important to understand that these techniques do not necessarily lead to good generalization to test data – even feature selection methods can suffer from overtraining due to the large number of potential feature subsets that need to be evaluated using limited training data. Adding constraints to select *groups of features* (i.e. either all or none of the features in a group are selected) helps to reduce the size of this search space. Such an approach is taken in structured sparsity [9, 11], however there the feature selection is implicit and it is not straightforward to control how many feature groups are selected.

In this paper we leverage the advantages of clustering ROIs and group feature selection in order to create a robust predictive model for brain connectivity data. We study features which quantify **ROI-ROI connections**. Clustering ROIs therefore naturally results in groups of connections, which share their start and end clusters. We call the concatenation of all connections within such a feature group a *hyperedge*, see Fig. 1. We cluster the ROIs into data-driven or anatomical clusters, which in both cases may lead to clusters of different sizes. This leads to hyperedges of different dimensionalities. Our goal is to select a set of discriminative hyperedges, i.e. per hyperedge we aim to select all, or none of the features.

By assuming that adjacent ROI-ROI connections are likely to work together in disorders, we further propose to examine *connected* networks of hyperedges. At each step of the feature selection approach, we therefore add a hyperedge that (a) is connected to the already selected hyperedges via one or both of its clusters and (b) leads to the largest improvement in performance on the training set. The performance is evaluated by the nearest mean classifier, which is efficient and insensitive to overfitting.

We contribute an interpretable predictive model for brain connectivity graphs, which selects local discriminative brain connectivity patterns, implicated in neurological disorder. We combat the overfitting problem by grouping ROIs into anatomical or data-driven clusters, and thus grouping ROI-ROI connections into groups of features. We use the cluster assignments to guide the group feature selection process, which, for anatomical clusters, leads to interpretable brain networks. Our method outperforms competing approaches on a dataset of ASD [3, 12].

2 Methods

Each subject's brain graph is represented by a symmetric $m \times m$ matrix which quantifies the brain connectivity between m ROIs in the brain, as illustrated in Fig. 1 (right). Each matrix is a collection of $M = (m^2 + m)/2$ individual connections (including the self-connections on the diagonal). That is, each subject is described by a vector $\mathbf{x}_i \in \mathbb{R}^M$ of M *features*, where each feature is an ROI-ROI fiber count. These m ROIs are associated with G clusters. The clusters organize these M features into $(G^2 + G)/2$ *feature groups* or *hyperedges*. Fig. 1 illustrates the relationship of the ROIs, clusters and the connectivity matrices.

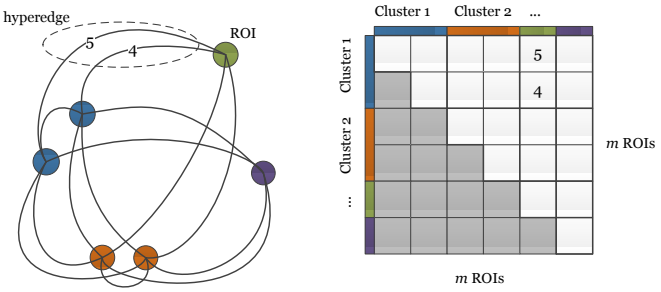

Fig. 1. Illustration of brain network (left) of $m = 6$ ROIs organized into $G = 4$ clusters (by color). Each connection is a feature; all features are summarized in a subject-specific data matrix (right). For each pair of ROI clusters, the set of connections between two clusters, or within a cluster, form a feature group called a *hyperedge*. Due to symmetry, this example has $M = 21$ unique features and 10 hyperedges (outlined in bold) to consider. There are 3 hyperedges with 1 feature, 4 with 2 features, 2 with 3 features, and 1 with 4 features.

For N subjects, this gives an $N \times M$ data matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\intercal$ and a vector $\mathbf{y} \in \{-1, 1\}^N$ of labels which describe the presence or absence of neurological disorder. Subsets of the features are denoted by X^s, where $s \in S = \{1, \ldots, M\}$. A hyperedge $H_{ij} \subset S$ is the set of indices of all the connections of clusters i and j.

Filter Approach to Feature Selection. We are interested in the feature subset $s^* \subset S$, which maximizes a goodness criterion c on the training data, $s^* = \arg\max_s c(X^s, \mathbf{y})$. Exhaustive evaluation of all choices for s is intractable for large M, while *individual feature selection* does not take feature correlations into account, rendering both approaches unsuitable for brain connectivity. A possible approach is therefore to perform *forward feature selection*: select the best feature, and iteratively grow the feature set with the feature j that leads to the largest improvement in $c(X^{s \cup j}, Y)$.

Network Group Feature Selection. Based on the assumption that discriminative information is contained in networks of anatomical regions, we propose to perform forward selection on *connected feature groups* rather than individual features. This further limits the flexibility of the feature selection methods and therefore helps to reduce overfitting. Furthermore, the selected feature set will be interpretable which is interesting from a diagnostic perspective.

We iteratively grow the feature set by adding the (i, j)-th hyperedge that leads to the largest increase in $c(X^{s \cup H_{ij}}, Y)$, and is adjacent to the already selected hyperedges. For example, in Fig. 1, {blue-green, green-purple} is allowed because the hyperedges are connected at the green cluster, but the feature set {blue-green, purple-orange} is not. A procedure overview is shown in Fig. 2.

Goodness Criterion. The goodness criterion c could be univariate, such as a t-test. However, we need a multivariate c because we want to evaluate groups of hyperedges. We define c as the average cross-validation performance of the nearest mean classifier (NMC) on the training set, $c(X, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} I(f(\mathbf{x}_i) == y_i)$ where $f(\mathbf{x}) = \arg\min_{l \in \{-1,1\}} \|\boldsymbol{\mu}_l - \mathbf{x}\|$ and $\boldsymbol{\mu}_l$ are the class means. This choice has several advantages: NMC is very inflexible and therefore relatively insensitive to overfitting [13], and its performances on different cross-validation folds can be computed very efficiently in matrix form.

The NMC errors are discretized into ranks (equal error = equal rank). To avoid discarding potentially good feature sets when feature sets s and s' have similar errors, we consider all feature sets with rank up to R, therefore reducing the greediness of our method. The added computational effort of this step is compensated by the fact that fewer feature sets need to be evaluated due to the network constraint.

Evaluation Procedure. We perform 10-fold cross-validation, the feature selection is performed only on the training set. The area under the receiver-operating characteristic (AUC) is used for evaluation; random performance is equal to 0.5 and perfect performance is equal to 1.

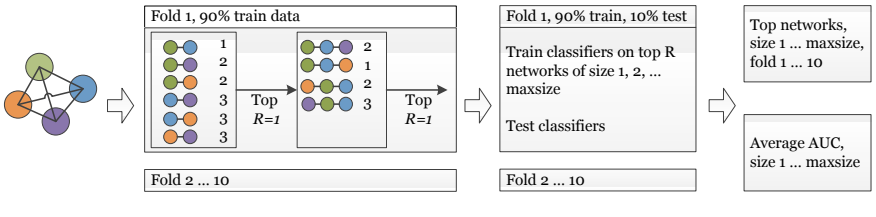

Fig. 2. Overview of procedure. The hyperedges are ranked, then the hyperedges with rank less than or equal to R (here $R = 1$) are selected for the next step. The selected hyperedges are extended to form pairs of hyperedges, which are again ranked. The procedure continues until the desired network size is reached.

3 Experiments and Results

We apply our algorithm to the structural connectivity matrices from the autism spectrum disorder dataset of [3, 12]. Each subject is represented by a 264×264 matrix encoding the number of paths connecting 264 ROIs from a functional atlas [14], where paths are found by deterministic tractography, using the fiber assignment by continuous tracking (FACT) algorithm in Diffusion Toolkit[1]. There are $N = 94$ subjects (51 ASD, 43 TD), matched for factors such as IQ and age. The ROIs are divided into clusters using (a) 88 anatomical regions, based on the coordinates of the 264 ROIs, and (b) 9 data-driven clusters obtained through Louvain modularity on the mean connectivity matrix.

[1] http://trackvis.org/dtk

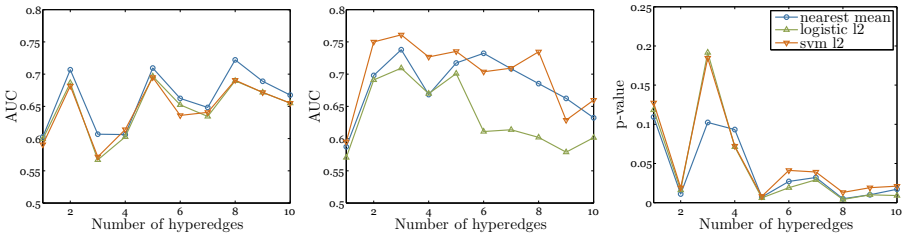

Fig. 3. AUC for an increasing number of hyperedges between data-driven clusters (left), anatomical clusters (middle) and p-values for the anatomical clusters (right). The best result achieved is **AUC of 0.76** for 3 anatomical hyperedges and an SVM.

We evaluate the network group forward selection on both (a) and (b). Fig. 3 shows how the performances change when more hyperedges (best hyperedge, best pair, ...) are added in the two models. As the anatomical clusters achieve superior, less noisy performances and improve interpretability, we use them in further experiments. To test significance of these results, we provide p-values based on a permutation test with 1000 repetitions in Fig. 3 (right).

Comparison. We first compare our method (`fwd-network` with 5 hyperedges and SVM) to filter feature selection with a t-test (`t-test`), minimum redundancy maximum relevance [15] (`mrmr`) from the FEAST toolbox [16] and SVM with recursive feature elimination [17] (`rfe`). We use 300 highest ranked features and an SVM (from PRTools [18]) as a classifier. We also compare to a sparse logistic (`sp-basic`) and group sparse logistic (`sp-group`), both from SLEP toolbox [19], and a tree-structured sparse logistic [9] (`sp-tree`), modified to suit our connectivity data. This method averages correlated groups of connections and enforces that within such correlated groups, all or none of the features are selected. Note that for the sparse methods, it is not possible to control the number of hyperedges explicitly, as in our method. Lastly, we compare to selecting 5 hyperedges without the network constraint (`fwd-group`). The results are shown in Table 1.

Table 1. AUC mean ± std, ×100, of selection of 300 best features (`t-test` to `rfe`), sparse methods (`sp-`) or forward selection (`fwd-`) with 5 hyperedges. 300 features and 5 hyperedges are chosen as default parameters, and NOT to correspond with best performances in Fig. 3.

`t-test`	`mrmr`	`rfe`	`sp-basic`	`sp-group`	`sp-tree`	`fwd-network`	`fwd-group`
56.6	52.4	48.9	51.3	47.4	50.1	**73.5**	67.8
± 25.7	± 22.2	± 17.4	± 21.6	± 20.3	± 23.3	± **16.6**	± 16.2

Interpretation. Table 2 shows the subnetworks frequently selected by our method. In each fold, we record which networks have ranks 1 to 5. Not considered networks are ranked with a 10. By averaging ranks over the folds, we

Table 2. Average ranks (1 = best, 10 = worst) for networks with 1-3 hyperedges. R Putamen (RP), R Superior Parietal Lobule (RSPL), R Parahippocampal Gyrus posterior (RPGp), R Thalamus (RT), L Insular Cortex (LIC), L Planum Temporale (LPT), L Precentral Gyrus (LPG), R Superior Temporal Gyrus posterior (RSTGp), L Inferior Frontal Gyrus pars opercularis (LIGFpo), L Caudate (LC)

Size 1	Rank	Size 2	Rank	Size 3	Rank
RP-RSPL	1.8	LPG-RP-RSPL	2.9	LPG-RP-RSPL-RSTGp	6.2
RPGp-RT	4.5	RP-LIC-LPT	6.7	LC-LPG-RP-RSPL	6.4
LIC-LPT	5.6	RP-RSPL-RSTGp	7.0	LIFGpo-LPG-RP-RSPL	7.6

quantify how frequently a network is selected: if a network is the best in all folds, its average rank is 1, if a network is never selected, its average rank is 10.

4 Discussion and Conclusions

Our method selects groups of connectivity features based on prior anatomical knowledge and outperforms competing approaches which either do not consider groups of features and/or do not select features prior to training a classifier.

Comparison of Classifiers. As expected, methods which select features individually (`t-test`) perform very poorly because feature interactions are not taken into account. Forward selection methods which add or remove one feature per iteration (`mrmr` and `rfe`) perform even worse, because overfitting becomes a problem due to the amount of possible feature subsets that are evaluated. Selecting groups of features in a forward fashion (`fwd-`) therefore yields superior results, and network selection (`fwd-network`) outperforms selecting groups without such network constraints (`fwd-group`).

The sparse classifiers (`sp-`) perform very poorly. We suspect this is mainly due to the high dimensionality and correlations of the feature space, which lead to overfitting. While structured sparsity aims at selecting few features or feature groups, this cannot be controlled explicitly because a relaxation of the desired l_0 norm is used. Therefore, solutions that are less sparse than desirable could still be chosen. Our method explicitly controls the sparsity by choosing the network size, and therefore removes some potentially harmful solutions. To this end, it would also be interesting to investigate methods from computational biology which explicitly optimize the l_0 norm, such as [20].

In general, we suspect the problems are also caused by the difficulty of the data, because other methods, such as averaging of features for each ROI [7] (AUC ≈ 0.5 on our dataset) or structured sparsity [9], perform well on related problems, but not on this ASD dataset. Perhaps the numbers of fiber tracts do not contain enough discriminative information for this study, because of the differences in the tractography procedure.

Anatomical vs. Data-Driven Clusters. In our method, anatomical ROI clusters outperform data-driven clusters. This may be caused by the dimensionality

of the hyperedges, which is higher for the data-driven clusters because the features are divided into less groups, resulting in more features per group. Perhaps more importantly, prior knowledge acts as an intrinsic regularization. We assume that the discriminative information is contained in connected subnetworks of anatomical regions, which reduces the solution search space. The fact that we obtain superior results indicates that the corresponding search space reduction removes solutions (which are not connected subnetworks) that would fit the training data perfectly but not generalize to test data.

Interpretability. Our method selects networks of anatomical regions, and therefore allows interpretation of results from a neurological perspective. The networks selected for this dataset often contain the right putamen. A study of DTI measures in ASD [21], finds significant differences in white matter tracts passing through the putamen (primarily left) to the frontal cortex, which is in part consistent with our results. Other studies of white matter pathways in ASD (reviewed in [22, 23]), find differences in the connections between temporal and occipital lobes and between the cingulate cortex and medial temporal structures. Although we do not find these specific connections, differences in the data acquisition and methodology could also be leading to inconsistencies.

Further Investigation. Age is important in the development of the brain during ASD [4, 22], increasing the variability in brain connectivity inside each class. To this end, we analyzed the correlation of performance, and the similarity of ages in the training and test set. Moderate correlations suggest that it might be advantageous to train age-dependent classifiers. Our initial efforts to do so did not outperform the proposed method, probably because of the even further reduced sample sizes. A remaining question is how to incorporate age in the classification procedure, without splitting the data into smaller subsets.

Conclusions. We propose a network-guided group feature selection method for structural brain connectivity data. The approach reduces overfitting by incorporating prior anatomical knowledge about ROI-ROI connections, and outperforms both methods where group structure is not considered, and data-driven methods. Our method provides interpretable output in the form of connected subnetworks between anatomical regions of the brain, which are discriminative for patients and controls. On a dataset of ASD, we obtain an AUC of 0.76 and select subnetworks which point in the direction of brain areas to be investigated in ASD. Future improvements could include incorporating the subjects' ages into classification.

Acknowledgements. We thank prof. dr. Karsten Borgwardt and dr. Cédric Koolschijn for their valuable advice concerning the paper. Aasa Feragen is supported by The Danish Council for Independent Research — Technology and Production.

References

1. Stigler, K.A., et al.: Structural and functional magnetic resonance imaging of autism spectrum disorders. Brain Research 1380, 146–161 (2011)
2. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10(3), 186–198 (2009)
3. Rudie, J., Brown, J., et al.: Altered functional and structural brain network organization in autism. NeuroImage: Clinical (2012)
4. Ghanbari, Y., Smith, A.R., Schultz, R.T., Verma, R.: Connectivity subnetwork learning for pathology and developmental variations. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 90–97. Springer, Heidelberg (2013)
5. Rubinov, M., Bullmore, E.: Fledgling pathoconnectomics of psychiatric disorders. Trends in Cognitive Sciences 17(12), 641–647 (2013)
6. Ecker, C., et al.: Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. Neuroimage 49(1), 44–56 (2010)
7. Ingalhalikar, M., Kanterakis, S., Gur, R., Roberts, T.P.L., Verma, R.: DTI based diagnostic prediction of a disease via pattern classification. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 558–565. Springer, Heidelberg (2010)
8. Ghiassian, S., et al.: Learning to Classify Psychiatric Disorders based on fMR Images: Autism vs Healthy and ADHD vs Healthy. In: MLINI (2013)
9. Jenatton, R., et al.: Multiscale mining of fMRI data with hierarchical structured sparsity. SIAM J. on Imaging Sciences 5(3), 835–856 (2012)
10. Orrù, G., et al.: Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosc. Biobeh. Rev. 36(4), 1140–1152 (2012)
11. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Structured sparsity through convex optimization. Statistical Science 27(4), 450–468 (2012)
12. Brown, J.A., et al.: The UCLA multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis. Frontiers in Neuroinformatics 6 (2012)
13. Skurichina, M., Duin, R.P.W.: Stabilizing classifiers for very small sample sizes. In: International Conference on Pattern Recognition, vol. 2, pp. 891–896. IEEE (1996)
14. Power, J.D., et al.: Functional network organization of the human brain. Neuron 72(4), 665–678 (2011)
15. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE TPAMI 27(8), 1226–1238 (2005)
16. Brown, G., et al.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. JMLR 13, 27–66 (2012)
17. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46(1-3), 389–422 (2002)
18. Duin, R.P.W., et al.: PRTools, a MATLAB toolbox for pattern recognition (2010), `http://www.prtools.org`
19. Liu, J., Ji, S., Ye, J.: SLEP: Sparse Learning with Efficient Projections (2009)
20. Azencott, C.A., et al.: Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics 29(13), i171–i179 (2013)
21. Langen, M., et al.: Fronto-striatal circuitry and inhibitory control in autism: findings from diffusion tensor imaging tractography. Cortex 48(2), 183–193 (2012)
22. Vissers, M.E., et al.: Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links. Neurosci. Biobehav. Rev. 36(1), 604–625 (2012)
23. Travers, B.G., et al.: Diffusion tensor imaging in autism spectrum disorder: a review. Autism Research 5(5), 289–313 (2012)