Data-Specific Feature Selection Method Identification for Most Reproducible Connectomic Feature Discovery Fingerprinting Brain States

Nicolas Georges, Islem Rekik(✉), and for the Alzheimers's Disease Neuroimaging Initiative

BASIRA lab, CVIP group, School of Science and Engineering, Computing, University of Dundee, Dundee, UK
irekik@dundee.ac.uk
http://www.basira-lab.com

Abstract. Machine learning methods present unprecedented opportunities to advance our understanding of the connectomics of brain disorders. With the proliferation of extremely high-dimensional connectomic data drawn from multiple neuroimaging sources (e.g., functional and structural MRIs), effective feature selection (FS) methods have become indispensable components for (i) disentangling brain states (e.g., early vs late mild cognitive impairment) and (ii) identifying connectional features that might serve as biomarkers for treatment. Strangely, despite the extensive work on identifying stable discriminative features using a particular FS method, the challenge of choosing the best one from a large pool of existing FS techniques for optimally achieving (i) and (ii) using a dataset of interest remains unexplored. In essence, the question that we aim to address in this work is: "Given a set of feature selection methods $\{FS_1, \ldots, FS_K\}$, and a dataset of interest, which FS method might produce the *most reproducible and 'trustworthy'* connectomic features that accurately differentiate between two brain states?" This paper is an attempt to address this question by evaluating the performance of a particular feature selection for a specific data type in fulfilling criteria (i) and (ii). To this aim, we propose to model the relationships between a set of FS methods using a multi-graph architecture, where each graph quantifies the feature reproducibility power between graph nodes at a fixed number of top ranked features. Next, we integrate the reproducibility graphs with a discrepancy graph which captures the difference in classification performance between FS methods. This allows to identify, for a dataset of interest, the 'central' node with the highest degree, which reveals the most reliable and reproducible FS method for the target brain state classification task along with the most discriminative features fingerprinting these brain states. We evaluated our method on multi-view brain connectomic data for late mild cognitive impairment vs Alzheimer's disease classification. Our experiments give insights into *reproducible* connectional features fingerprinting late dementia brain states.

G. Wu et al. (Eds.): CNI 2018, LNCS 11083, pp. 99–106, 2018.
https://doi.org/10.1007/978-3-030-00755-3_11

1 Introduction

Neurological and neuropsychiatric disorders, including Autism spectrum disorder (ASD), Alzheimer's disease (AD) or Mild Cognitive Impairment (MCI), are distinctive conditions that affect the morphology, cognition, and function of the brain. Understanding the connectomics of these brain disorders [1] can help improve diagnosis, prognosis, and patient treatment. To this aim, several works leveraged machine learning techniques [2–4] as well as graph analysis techniques [5] to discover distinctive brain features which reliably differentiate between normal subjects and disordered patients. These might serve as biomarkers, which can be targeted for developing efficient treatment. Due to the high dimensionality of connectomic data, many machine learning methods embed feature selection (FS) techniques to effectively reduce the dimensionality of data samples by selecting a subset of highly relevant features. Despite the great progress made over the last decade in devising robust and accurate FS methods [6], developing a new approach that would produce the best classification results and identify the most reliable feature for *all* data types seems to be an intractable problem. In fact, the ongoing proliferation of multi-source medical data, including structural and functional magnetic resonance imaging (MRI) data collected for the human brain connectome project [7], presents unprecedented challenges to devise feature selection methods that generate reproducible biomarkers across different data sources. This is because each data source has its unique characteristics and statistical distribution that might not match that of another data source. Hence, identifying the best feature selection method that unravels the inherent traits of a particular dataset remains a major challenge.

Despite the great potential that many FS methods hold for identifying connectomic biomarkers for neurological disorders (e.g., Tourette Syndrome, ASD) [8–10], training on small datasets comes with its limitations including an observable variability of most discriminative features. Being able to rely on a stable FS method that is 'optimal' for a specific dataset would constitute a radical change for detecting disordered brain changes through the connectome data. Our hypothesis is that the best performing FS method for a dataset of interest might not be optimal for a different dataset in terms of classification accuracy and feature reproducibility. To the best of our knowledge, existing FS assessment criteria have mainly focused on the stability criteria [11,12], which quantifies the sensitivity of feature selection methods to variations in the training set. However, this does not assess the suitability of the 'selected' FS method for a particular dataset. Basically, the question that we aim to address in this work is: Given a set of feature selection methods $\{FS_1, \ldots, FS_N\}$, and a particular dataset, which FS method might produce the *most reproducible and 'trustworthy'* connectomic features that accurately differentiate between two brain states (e.g., demented vs healthy)?

In contrast to methods focusing on boosting the accuracy of FS methods [13] in classifying different brain states, our primary goal is not to maximize individual-level classification accuracy but to identify the best FS method that will produce reproducible brain features associated with a specific brain disorder

(i.e., potential biomarkers) for a particular dataset. To do so, given a set of FS methods, we first model the relationship between FS methods using a set of graphs, each graph quantifies the feature reproducibility power between neighboring nodes at a fixed number of top ranked features (i.e., a 'feature threshold' K). The weight of an edge connecting two FS nodes in the graph captures the overlap in top K ranked features. Next, we generate a discrepancy FS graph, where the strength of an edge connecting two FS nodes encodes the absolute difference in their classification accuracy. Ultimately, by merging all reproducibility and discrepancy graphs, we generate a holistic graph which allows the identify the central FS method with most reproducible features *in relation* to other FS methods in the graph. More importantly, the selected central FS method will be used to identify the most meaningful and reproducible connectomic features for a brain disorder of interest.

2 Multi-graph Based Identification of Data-Specific Feature Selection Method for Reproducible Discriminative Feature Discovery

In this section, we introduce the proposed pipeline to identify the FS method that produces 'the most agreed upon' features for distinguishing between two groups drawn from a connectomic data of interest. Fig. 1 displays the key steps of our framework.

Multi-view Connectomic Feature Extraction. Each brain is represented by a set of n_v networks $\{\mathbf{V}_i\}_{i=1}^{n_v}$, each encoding a particular view of the connectional brain construct. To train our classification model based on the identified FS method, we define a feature vector $\mathbf{v}_k$ for each brain network view k, whose elements belong to the off-diagonal upper triangular part of the corresponding connectivity matrix (Fig. 1).

FS-to-FS Multi-graph Construction. Given a particular data view, we aim to identify the best feature selection method that gives the most reproducible and reliable features allowing to tease apart two brain states. We hypothesize that the most reliable FS method is able to reproduce the majority discriminative features identified by other methods, thereby achieving the highest consensus with other FS methods. The most appealing characteristic of the approach is that it evaluates the importance of a given FS method while considering a set of FS methods at a given cut-off threshold K representing the number of top K ranked features selected to train the classifier (e.g., support vector machine –SVM). Given a set of N FS methods $\mathcal{F} = \{FS_1, \ldots, FS_N\}$, we construct an undirected fully-connected graph $G_K = (V_K, E_K)$; V_K is the set of nodes, each nesting an FS method in $\mathcal{F}$, while E_K represents weighted edges, which model pairwise overlap in top K features among FS methods. By varying the cut-off values K, we define a set of graphs $\mathcal{G}$ (or multi-graph) that model the overlap between FS methods at different levels. Next, for easily merging the generated multiple graphs, we represent each G_K as a similarity matrix $\mathbf{S}_K$ (Fig. 1), where each element $\mathbf{S}_K(i, j)$ denotes the overlap in top K ranked features between

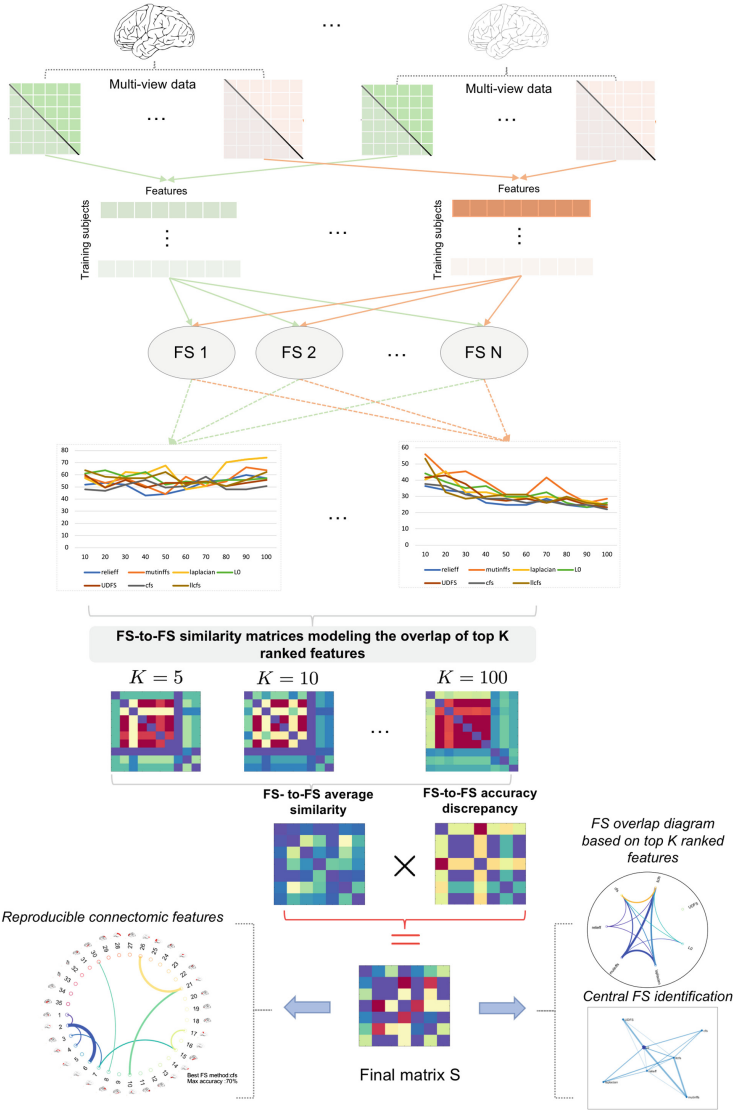

Fig. 1. *Proposed data-specific feature selection method identification pipeline.* For each subject, we define connectomic feature vectors, each derived from a particular brain view. We note that the performance of different FS methods varies with data types. Given a particular data view, we define multiple graphs, each represented as a similarity matrix modeling the consensus in top K ranked features among other selection methods. Next, we define a accuracy discrepancy matrix measuring the pairwise absolute difference in average accuracy between FS methods. By merging consensus similarity defined at multiple thresholds K with the accuracy discrepancy matrix, we generate a final matrix $\mathbf{S}$. The best FS method for the dataset of interest is defined as the node with the highest centrality in $\mathbf{S}$, thereby allowing to identify the most reproducible features distinguishing between two brain states (e.g., healthy vs disordered states).

FS methods i and j. We generate an average similarity matrix $\bar{\mathbf{S}}$ by merging all similarity matrices across all thresholds, thereby capturing the *average FS method consensus* with other methods.

FS-to-FS Accuracy Discrepancy Matrix Construction. Since classification accuracy influences the credibility of the produced distinctive features, we propose to model the relationship between FS methods in terms of discrepancy in average classification accuracy. Hence, we define an average accuracy discrepancy matrix $\bar{\mathbf{A}}$, where the cost $\bar{\mathbf{A}}(i,j)$ of an edge connecting two nodes i and j is defined as $\bar{\mathbf{A}}(i,j) = |\bar{a}_i - \bar{a}_j|$, where $\bar{a}_i$ represents the average accuracy of FS method i at different cut-off thresholds. Next, we merge both $\bar{A}$ and $\bar{S}$ to output the final FS similarity matrix $\mathbf{S}$ (Fig. 1).

FS Method Identification. We assign a score c_i for each FS_i in $\mathbf{S}$, that quantifies the consensus in top selected feature set as well as classification performance among other methods. In particular, inspired from graph analysis theory, we define c_i as the centrality measure, indicating the number of times that FS method is visited on whatever path of a given length. The final FS method is selected as the one with the highest centrality in $\mathbf{S}$. Once, we identify the most reliable FS method, we train an SVM classifier using the top K selected features by FS to reveal the most discriminative ones.

3 Results and Discussion

Evaluation Dataset. To perform the classification, we used leave-one-out cross validation on 77 subjects (41 AD and 36 late MCI) from ADNI data[1], each with structural T1-w MR image [14]. We reconstructed both right and left cortical hemispheres for each subject from T1-w MRI using FreeSurfer software [15]. Then we parcellated each cortical hemisphere into 35 cortical regions using Desikan-Killiany Atlas [15]. We generated two brain network datasets derived from the maximum principal curvature brain view and the mean cortical thickness brain view, respectively. For each cortical attribute, we produced a morphological brain network, where the strength of a connection linking ROI i to ROI j is defined as the absolute difference between the averaged attribute values in both ROIs [2,3].

FS Methods and Training. We used the Feature Selection Library [16] provided by Matlab to select 7 FS methods: relieff [17], mutinffs [18], laplacian [19], L0 [20], UDFS [21], llcfs [22], and cfs [23]. We adopted a leave-one-out cross-validation (CV) strategy to train each FS in combination with an SVM classifier. For FS methods that required parameter tuning, we used nested CV. For each FS method, we evaluated the performance of SVM classifier across different number of top K selected features varying from 10 to 100, with a step size of 10 features.

Findings and Future Improvements. Fig. 2 confirms our hypothesis that the best FS method for one data type might not be the best for another data type.

[1] http://adni.loni.usc.edu/.

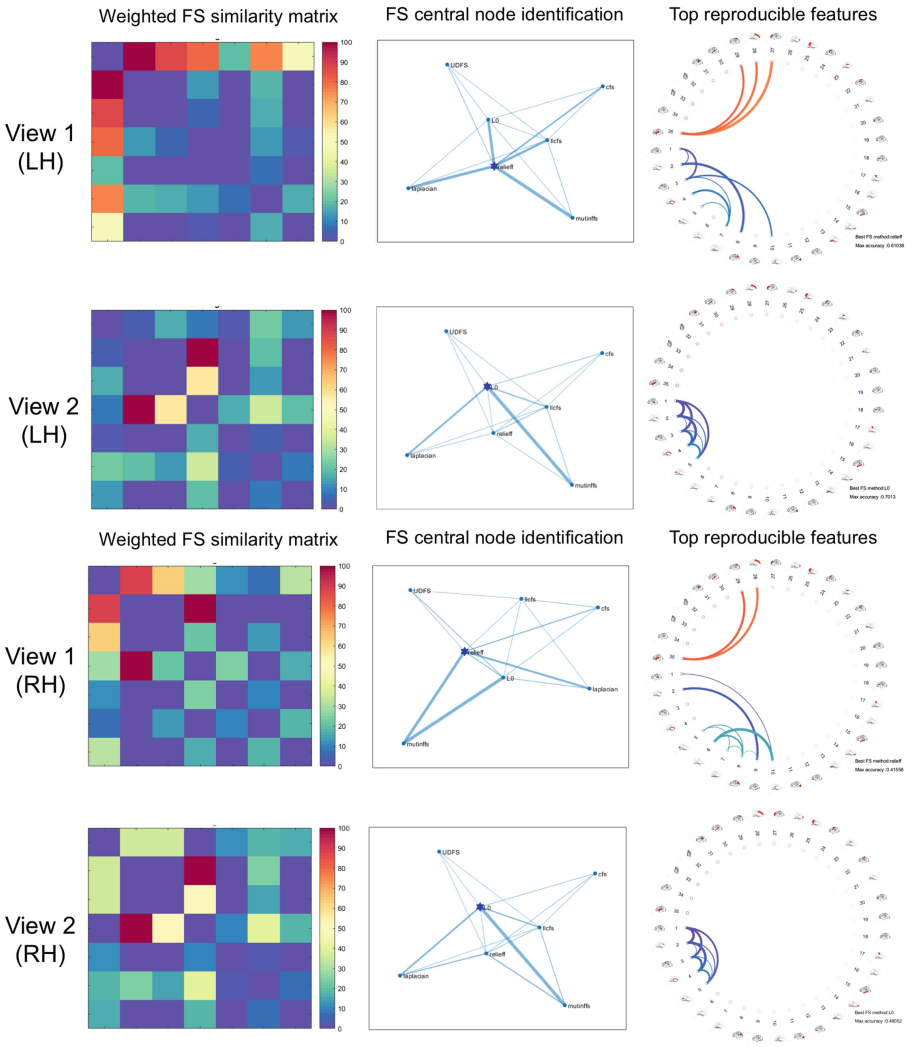

Fig. 2. *Top 10 reproducible discriminative features identification using the best identified feature selection (FS) method for each network brain view data.* Selected FS methods ($\star$), corresponding classification accuracy, and top reproducible features varied across data types and right and left hemispheres (RH and LH).

For instance, relieff was identified for view 1 LH connectomic data with a classification accuracy of 61.03%, while L0 was identified for view 2 LH connectomic data with a classification accuracy of 70.3%. Overall, the identified discriminative features distinguishing between LMCI and AD brain states varied across views and cortical hemispheres. However, we note that nodes 1, 2 and 5 corresponding with the bank of the superior temporal sulcus, caudal anterior-cingulate cortex,

and cuneus cortex were frequently selected. These regions were reported in other studies on AD [24].

There are several future directions to explore to further improve our seminal work. First, instead of pre-defining a similarity matrix modeling the relationship between FS methods in terms of top ranked feature consensus, we can instead learn these associations in a more generic way. Second, we will integrate the feature stability criteria for FS method identification. Third, we will evaluate our method on multiple connectomic datasets, including functional and structural connectomes. Fourth, ideally, the FS method giving the best classification accuracy would identify the most discriminative and reproducible features. We aim to further improve our framework to identify the data-specific FS method that satisfies both criteria.

4 Conclusion

In this work, we investigated a novel problem arising from the need to discover the most reproducible and reliable clinical biomarkers that distinguish between two groups (e.g., healthy and disorders brains) by identifying the best feature selection method suited for the dataset of interest. We first proposed the concept of FS similarity multi-graph to model the relationships between different FS methods in terms of overlap top ranked features at multiple thresholds. By further integrating an accuracy discrepancy graph with the similarity multigraph to enforce a consistency between high classification performance and feature reproducibility when identifying the best FS method for the target input data. By exploring the topological properties of the merged graph, we mark the central FS node with the highest the centrality score as the most reliable one. Our preliminary findings showed that the performance of a particular FS method to train a typical classifier varies with the data type. Besides classification accuracy, it is also possible to integrate feature stability as a measure to identify the best FS method. Another line of our ongoing work is to study the reproducibility of the identified features by the 'best' FS methods across *multi-source* medical datasets.

References

1. Fornito, A., Zalesky, A., Breakspear, M.: The connectomics of brain disorders. Nat. Rev. Neurosci. **16**, 159 (2015)
2. Mahjoub, I., Mahjoub, M.A., Rekik, I.: Brain multiplexes reveal morphological connectional biomarkers fingerprinting late brain dementia states. Sci. Rep. **8**(1), 4103 (2018)
3. Lisowska, A., Rekik, I.: Joint pairing and structured mapping of convolutional brain morphological multiplexes for early dementia diagnosis. Brain Connectivity (2018). https://doi.org/10.1089/brain.2018.0578
4. Zhao, F., Zhang, H., Rekik, I., An, Z., Shen, D.: Diagnosis of autism spectrum disorders using multi-level high-order functional networks derived from resting-state functional MRI. Front. Hum. Neurosci. **12**, 184 (2018). https://doi.org/10.3389/fnhum

5. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. **10**, 186 (2009)
6. Liu, H., Motoda, H.: Computational methods of feature selection (2007)
7. Van Essen, D.C., Glasser, M.F.: The human connectome project: progress and prospects. Cerebrum Dana Forum Brain Sci. **2016** (2016)
8. Lisowska, A., Rekik, I.: Pairing-based ensemble classifier learning using convolutional brain multiplexes and multi-view brain networks for early dementia diagnosis. In: Wu, G., Laurienti, P., Bonilha, L., Munsell, B.C. (eds.) CNI 2017. LNCS, vol. 10511, pp. 42–50. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67159-8_6
9. Soussia, M., Rekik, I.: High-order connectomic manifold learning for autistic brain state identification. In: Wu, G., Laurienti, P., Bonilha, L., Munsell, B.C. (eds.) CNI 2017. LNCS, vol. 10511, pp. 51–59. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67159-8_7
10. Wen, H., et al.: Combining disrupted and discriminative topological properties of functional connectivity networks as neuroimaging biomarkers for accurate diagnosis of early tourette syndrome children. Mol. Neurobiol. **55**, 3251–3269 (2018)
11. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. **12**, 95–116 (2007)
12. Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S.: Measuring stability of feature selection in biomedical datasets. In: AMIA Annual Symposium Proceedings, vol. 2009, p. 406. American Medical Informatics Association (2009)
13. Chen, Y.W., Lin, C.J.: Combining SVMs with various feature selection strategies. In: Feature extraction, pp. 315–324. Springer, Heidelberg (2006). https://doi.org/10.1007/978-3-540-35488-8_13
14. Mueller, S.G.: The alzheimer's disease neuroimaging initiative. Neuroimaging Clin. North Am. **10**, 869–877 (2005)
15. Fischl, B.: Automatically parcellating the human cerebral cortex. Cereb. Cortex **14**, 11–22 (2004)
16. Roffo, G.: Feature selection library (MATLAB toolbox). arXiv preprint arXiv:1607.01327 (2016)
17. Kononenko, I., Šimec, E., Robnik-Šikonja, M.: Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl. Intell. **7**, 39–55 (1997)
18. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. IEEE Trans. Neural Netw. **20**, 189–201 (2009)
19. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in Neural Information Processing Systems, pp. 507–514 (2006)
20. Han, J., Sun, Z., Hao, H.: l0-norm based structural sparse least square regression for feature selection. Pattern Recogn. **48**, 3927–3940 (2015)
21. Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X.: l2, 1-norm regularized discriminative feature selection for unsupervised learning. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, p. 1589 (2011)
22. Zeng, H., Cheung, Y.m.: Feature selection and kernel learning for local learning-based clustering. IEEE Trans. Patt. Anal. Mach. Intell. **33**, 1532–1547 (2011)
23. Hall, M.A.: Correlation-based feature selection for machine learning (1999)
24. Wee, C.Y., Yap, P.T., Shen, D., Initiative, A.D.N.: Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum. Brain Mapp. **34**, 3411–3425 (2013)