A two-step clustering-based pipeline for big dynamic functional network connectivity data

Mohammad S. E. Sendi Student Member, IEEE, Robyn L Miller, Member, IEEE, David H Salat, Vince D. Calhoun, Fellow

Abstract— Dynamic functional network connectivity (dFNC) estimated from resting-state functional magnetic imaging (rs-fMRI) studies the temporal properties of FNC among brain networks by putting them into distinct states using the clustering method. The computational cost of clustering dFCNs has become a significant practical barrier given the availability of enormous neuroimaging datasets. To this end, we developed a new dFNC pipeline to analyze large dFNC data without accessing huge processing capacity. We validated our proposed pipeline and compared it with the standard one using a publicly available dataset. We found that both standard and iSparse k-means generate similar dFNC states while our approach is 27 times faster than the traditional method in finding the optimum number of clusters and creating better clustering quality.

I. INTRODUCTION

Most previous studies assume that the functional network connectivity (FNC) estimated from resting-state fMRI is static over time and ignores brain dynamics [1]. Indeed, functional connectivity is highly dynamic, even during the resting state without external stimuli [2]. In recent years, dynamic functional network connectivity (dFNC) has been introduced to study the strength of connectivity among brain regions and the temporal properties of the FNC [3]–[5]. It has been shown that dFNC can improve classification between disordered and healthy conditions [6], [7] and provides more information about the pathology of neurological and neuropsychiatric disorders than its static counterpart [8]. The standard dFNC analytic pipeline includes four steps in which the first step estimates intrinsic components for the desired brain regions. The second step calculates the dFNC using a sliding window. The third step concatenates all dFCNs of all subjects and goes through an optimization process to find the clustering order based on the elbow criterion the fourth step estimates the final dFNC for the whole group and state vector for each individual calculates the dFNC features for statistical analysis.

The dFNC pipeline primarily employs k-means clustering due to its ease of deployment and ability to scale to massive datasets. [9]. Also, compared to other mainly used clustering approaches, k-means clustering is faster. [10]. However, working with a sizeable dFNC dataset is still slow and requires a lot of computing power.

We proposed iterative sparse k-means (iSparse k-means), an iterative clustering technique that scales to millions of high-dimensional observations, making it a valuable addition to the pipeline for large-scale dFNC analysis. We also examined the time it takes to find the optimal cluster number using iSparse kmeans compared and regular kmeans and found that our method is faster than the standard approach.

II. MATERIALS AND METHODS

Our proposed analytic pipeline includes rs-fMRI preprocessing, extracting independent components, calculating dFNC, estimating the optimum number of state and dFNC states using the proposed clustering method. The following subsection describes each step in more detail.

A. Preprocessing and independent components extraction

To preprocess the fMRI data, we applied the statistical parametric mapping (SPM12, https://www.fil.ion.ucl.ac.uk/spm/) in MATLAB2019. We removed the first five dummy scans were removed before preprocessing. Next, we applied the rigid body motion correction to account for participant’s head movement. Next, we used spatial normalization by echo-planar imaging (EPI) template into the standard Montreal Neurological Institute (MNI) space. Finally, we applied a Gaussian kernel to smooth the fMRI images using a full width at half maximum (FWHM) of 6mm.

To extract intrinsic connectivity networks (ICNs) for each subject, we used the Neuromark pipeline [11]. Using this pipeline, we extracted 53 ICNs for each subject and put them in seven domain networks, including subcortical network (SCN), auditory network (ADN), sensorimotor network (SMN), visual network (VSN), cognitive control network (CCN), the default-mode network (DMN), and cerebellar network (CBN). The details of the extracted ICNs are provided in [12].
**B. Dynamic functional network connectivity estimation**

We used a tapered sliding window to estimate the functional network connectivity based on Pearson correlation. Within each window, we estimated 1378 connectivity features. Next, we concatenated dFNC estimates of each window for each subject to form a matrix, called dFNC tensor hereafter, with the size of $T \times F$, where $T$ denotes the number of windows and $F$ denotes the number of connectivity features (Fig. 1).

**C. Estimating dFNC state using iSparse kmeans**

Fig. 1 shows the proposed iSparse kmeans clustering method for estimating dFNC states and the cluster order. This method includes four steps. In the first step, we sub-sample subjects dFNC tensors. Next, in the second step, we run standard kmeans clustering on the subsampled data with different k values from 2 to L and group them into $r \frac{L(L+1)}{2} - 1$ states in each iteration. By repeating this process $r$ times, we would have $r \frac{L(L+1)}{2} - 1$ states. Next, we use elbow criteria to find the optimum number of clusters using all $r \frac{L(L+1)}{2} - 1$ states. After finding the optimum number of clusters, called $K_{opt}$ here, we use conventional k-means clustering to group $r \frac{L(L+1)}{2} - 1$ states into $K_{opt}$ states, called final states. Finally, using the final $K_{opt}$ states, we assign the dFNC of each subject to one of the estimated states.

**D. Clustering quality assessment**

To assess the clustering quality for each dFNC data, we calculated the distance between the dFNC tensor and its associated cluster centroid. Next, we calculated the distance between each dFNC sample with the second closest cluster centroids. Then, we calculated the ratio of the latter to the former one for each dFNC instance, called distance ratio here. Finally, we averaged all distance ratios out for each participant.

$$R_p = \frac{1}{\tau} \sum_{i=1}^{\tau} \frac{d_{i,c}}{d_{i,sc}} \quad (1)$$

$d_{i,c}$ is the distance between each sample to the cluster centroid of the state the sample belongs, and $d_{i,sc}$ is the distance between each sample and the second closest cluster centroid. Also, $R_p$ is the averaged distance ratio for each participant. It is worth mentioning that a higher ratio means better quality in clustering. We also used Euclidian distance to calculate the distance between each sample and the cluster centroids.
To validate the proposed method, we used the rs-fMRI and demographic information collected from the 833 young, healthy adults (average age: 28.65; range: 22-37 years; female/male: 443/390) from the Human Connectome Project (HCP) [13], which is available on the HCP website (https://www.humanconnectome.org). This project was approved by the Washington University - University of Minnesota Consortium of the Human Connectome Project (WU-Minn HCP). The rs-fMRI data were collected on a Siemens Skyra 3T with a 32-channel RF receiver head coil. High resolution T2*-weighted functional images were acquired using a gradient-echo EPI sequence with TE = 33.1 ms, TR= 0.72 s, flip angle = 52°, slice thickness = 2 mm, 72 s slices and 2 mm isotropic voxel, a field of view: 208×180 mm (RO×PE), and duration: 14:33 (min: sec). We estimated dFNC feature through the method explained in Section 2.A and 2.B. The size of the estimated dFNC data is 848,827×1378 (8542 MB), in which the dFNC size of each participant is 10.25 MB.

III. RESULTS

A. Similarity between Standard kmeans and iSparse kmeans in estimating dFNC states

The first question we were interested to know was whether standard kmeans and iSparse kmeans would produce similar dFNC states or not. To test this, we clustered the dFNC with standard kmeans and iSparse kmeans with different L values (as shown in Fig.1). In the iSparse kmeans, we used 3% (~256.26 MB) of the entire dataset in each iteration. We found that the optimal number of clusters is 2 through both the standard and proposed kmeans clustering approaches based on the elbow criteria. Additionally, we assessed the similarity of dFNC states estimated by iSparse kmeans (with different L) with those estimated by conventional kmeans. We used the correlation across the matched states as a similarity metric. Fig 2A shows the similarity between matched states with varying values of L. We found that the similarity between the matched states generated by both approaches is more than 99%, with any value L of more than five. The estimated states with conventional kmeans and iSparse kmeans (L=6) are shown in Fig.2B. Finding the minimum L value significantly reduces the computational cost.

B. iSparse kmeans finds the optimum cluster number faster than the conventional kmeans

We compared the speed of our method in finding the optimum number of clusters to the standard kmeans clustering method, which uses the entire dataset, after determining the minimal reliable value of L. In iSparse kmeans, we used different percentages of the data, as shown in Fig. 3. We found that when we use a smaller percentage of data in each iteration, iSparse kmeans is faster. However, the similarity between the matched states calculated with standard kmeans and iSparse is still greater than 98 percent. Additionally, our proposed method is 27 times faster in finding the cluster order than the traditional method when we use only 0.12% of data (one subject or 10.25 MB) in each iteration.
C. iSparse kmeans has better cluster quality than the standard kmeans

We used a two-sample t-test to compare the distance ratio of standard kmeans vs. iSparse one for k=2. We found iSparse kmeans shows improved cluster quality than the standard approach with a higher distance ratio (p<0.0001, N=833).

IV. DISCUSSION

In recent years, clustering for the big data has focused on finding the most informative features and then running a kmeans on the reduced set. For example, a recent study used principal component analysis and projected the dimension of the data set from p to m (p>m). This is followed by application of kmeans clustering to the projected dataset [14]. Since this method needs to load the entire dataset, it is computationally demanding for large samples. Also, in this method, the kmeans is applied to the project space. Then, we do not estimate the cluster centroid in the original space. Even though we might transfer the cluster centroid to the original space, this estimate is inaccurate and yields lower cluster quality than the standard kmeans approach.

Each iteration of our method loads a portion of the data. As a result, both the required memory and the computational time are reduced. Our suggested technique is like mini batch kmeans in this regard, as it partially loads the data and does not require expensive processing resources. But as [15] shows, the cluster quality for mini-batch kmeans is less than the standard approach, especially when the number of clusters increases. Unlike the mini-batch kmeans approach, iSparse kmeans shortsens the clustering process (Fig.3) while also improving the quality of the results.

V. CONCLUSION

Previous dFNC analytics pipelines use standard kmeans clustering, which does not work for big dFNC data. In this study, we developed a dFNC pipeline that used our proposed iSparse kmeans clustering that reduced the evaluation time for finding the cluster order while we only loaded a portion of the dataset in several iterations. As a result, we reduce the needed computer power for our new method compare to what the traditional method requires for a large dataset. We showed that our method produces similar brain states as the standard method while running 27 times faster than the conventional kmeans for estimating the cluster order.

VI. REFERENCES

[1] A. A. Ioannides, “Dynamic functional connectivity,” Current Opinion in Neurobiology, vol. 17, no. 2, pp. 161–170, 2007.
[2] M. S. E. Sendi et al., “Alzheimer’s Disease Projection From Normal to Mild Dementia Reflected in Functional Network Connectivity: A Longitudinal Study,” Frontiers in Neural Circuits, vol. 14, no. January, 2021.
[3] M. S. E. Sendi, E. Zendehrouh, C. Ellis, J. Turner, and V. Calhoun, “Dynamic Functional Connectivity of Default Mode Network in Schizophrenia and Links to Symptom Severity,” Biological Psychiatry, vol. 89, no. 9, pp. S262–S263, 2021.
[4] M. S. E. Sendi et al., “Abnormal Dynamic Functional Network Connectivity Estimated from Default Mode Network Predicts Symptom Severity in Major Depressive Disorder,” Brain Connectivity, vol. 11, no. 10, pp. 838–850, 2021.
[5] M. S. E. Sendi et al., “Aberrant Dynamic Functional Connectivity of Default Mode Network in Schizophrenia and Links to Symptom Severity,” Front. Neural Circuits, vol. 15, 2021.
[6] B. Rashid et al., “Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity,” in IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 251–254.
[7] D. K. Saha, E. Damara, B. Rashid, S. M. Plis, and V. D. Calhoun, “A Classification-Based Approach to Estimate the Number of Resting Functional Magnetic Resonance Imaging Dynamic Functional Connectivity States,” Brain Connectivity, vol. 11, no. 2, pp. 132–145, 2021.
[8] S. S. Xiong and K. Krishnamurthy, “A Comparison of Static and Dynamic Functional Connectivity for Identifying Subjects and Biological Sex Using Intrinsic Individual Brain Connectivity,” Scientific Reports, vol. 9, no. 1, pp. 1–11, 2019.
[9] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and repeats?,” Pattern Recognition, vol. 93, pp. 95–112, 2019.
[10] L. McNees and J. Healy, “Accelerated Hierarchical Density Based Clustering,” in IEEE International Conference on Data Mining Workshops, ICDMW, 2017, no. May 2017, pp. 33–42.
[11] Y. Du et al., “NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders,” NeuroImage: Clinical, vol. 28, no. August, p. 102375, 2020.
[12] M. S. E. Sendi et al., “Alzheimer’s Disease Projection From Normal to Mild Dementia Reflected in Functional Network Connectivity: A Longitudinal Study,” Frontiers in Neural Circuits, vol. 14, no. January, pp. 1–15, 2021.
[13] M. F. Glasser et al., “The Human Connectome Project’s neuroimaging approach,” Nature Neuroscience, vol. 19, no. 9, pp. 1175–1187, 2016.
[14] D. Feldman, M. Schmidt, and C. Sohler, “Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering,” in 24th Annual ACM-SIAM Symp. Discrete Algorithms, 2013, pp. 1434–1453.
[15] J. Béjar Alonso, “K-means vs Mini Batch K-means: a comparison,” 2013.