THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

An Approach to Compare Clustering Results of Two Temporal fMRI Dataset

Presented By: Group 16

Ahmedur Rahman Shovon (ashovon)

Ashraful Islam (aislam)

Pratim Saha (psaha)

Shahariar Rabby (arabby)

April 20, 2022

Contents

Introduction

- Topological data analysis is an important topic in the field of data mining.
- This is especially useful to handle high-dimensional and noisy data.
- In this paper, we show an approach to compare different clustering results on an fMRI dataset of two temporal frequencies for a subject.
- fMRI (Functional Magnetic Resonance Imaging) is a non-invasive and non-detrimental process to quantify the neuronal activity of the brain during normal and diseased conditions
- We explore the similarity between two fMRI scans of the same people taken at two-time points.

Motivation

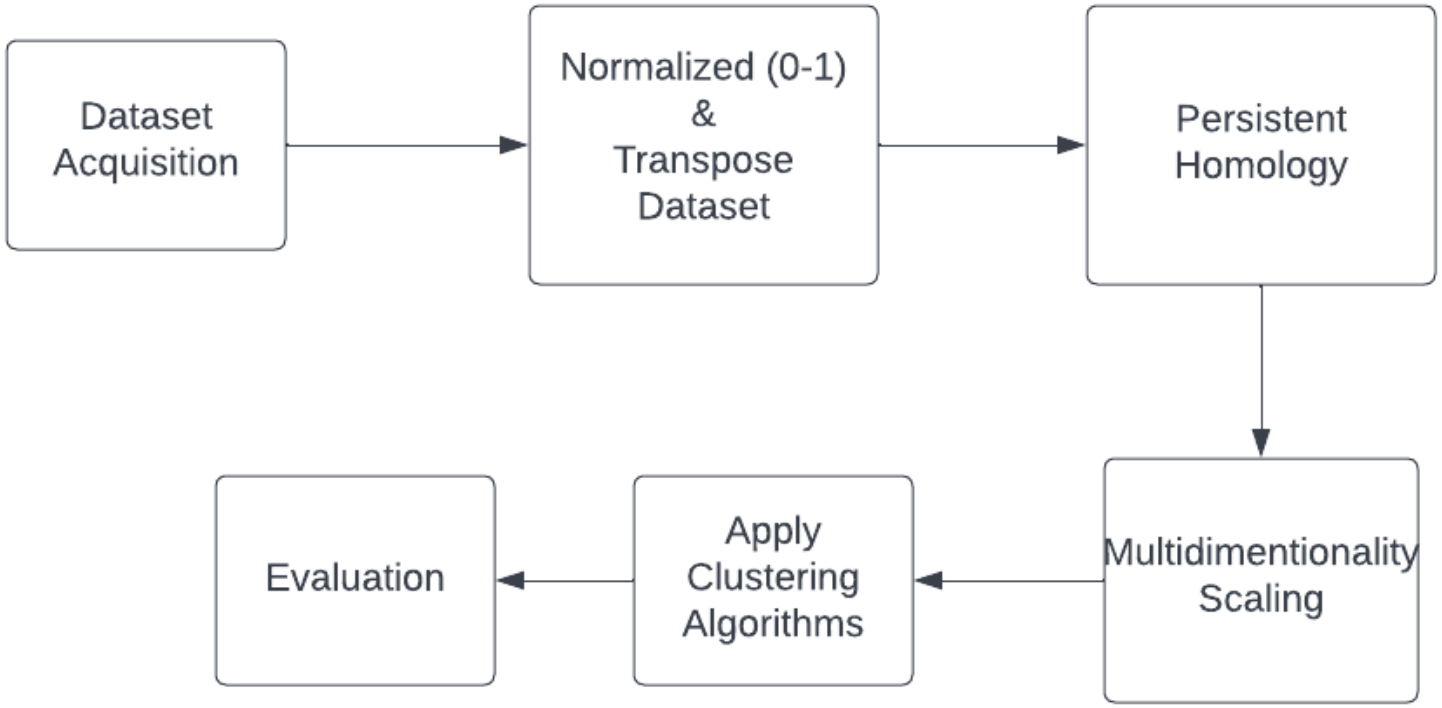
- Time-varying data analysis has increasingly become an integral part of fields such as Data Mining.
- Major objectives of understanding this type of data is to extrapolate meaningful information and correlation among the data points to forecast future outcome.
- fMRI is a time series data.
- We formulate the hypothesis that fMRI data for the same subjects at two time points should have similar natures.

Dataset Description

- Dataset was collected from Auburn University MRI research center.
- Two fMRI scan of 316 subjects were used in our experiment.
  - Dynamic_FC_2500:
    - Number of slices: 86
    - Dimension of each slice: 114*114
  - Dynamic_FC_1400:
    - Number of slices: 336
    - Dimension of each slice: 114*114
- Dataset was provided in the matrix format.

Methodology

Workflow Diagram



Normalization of Dataset

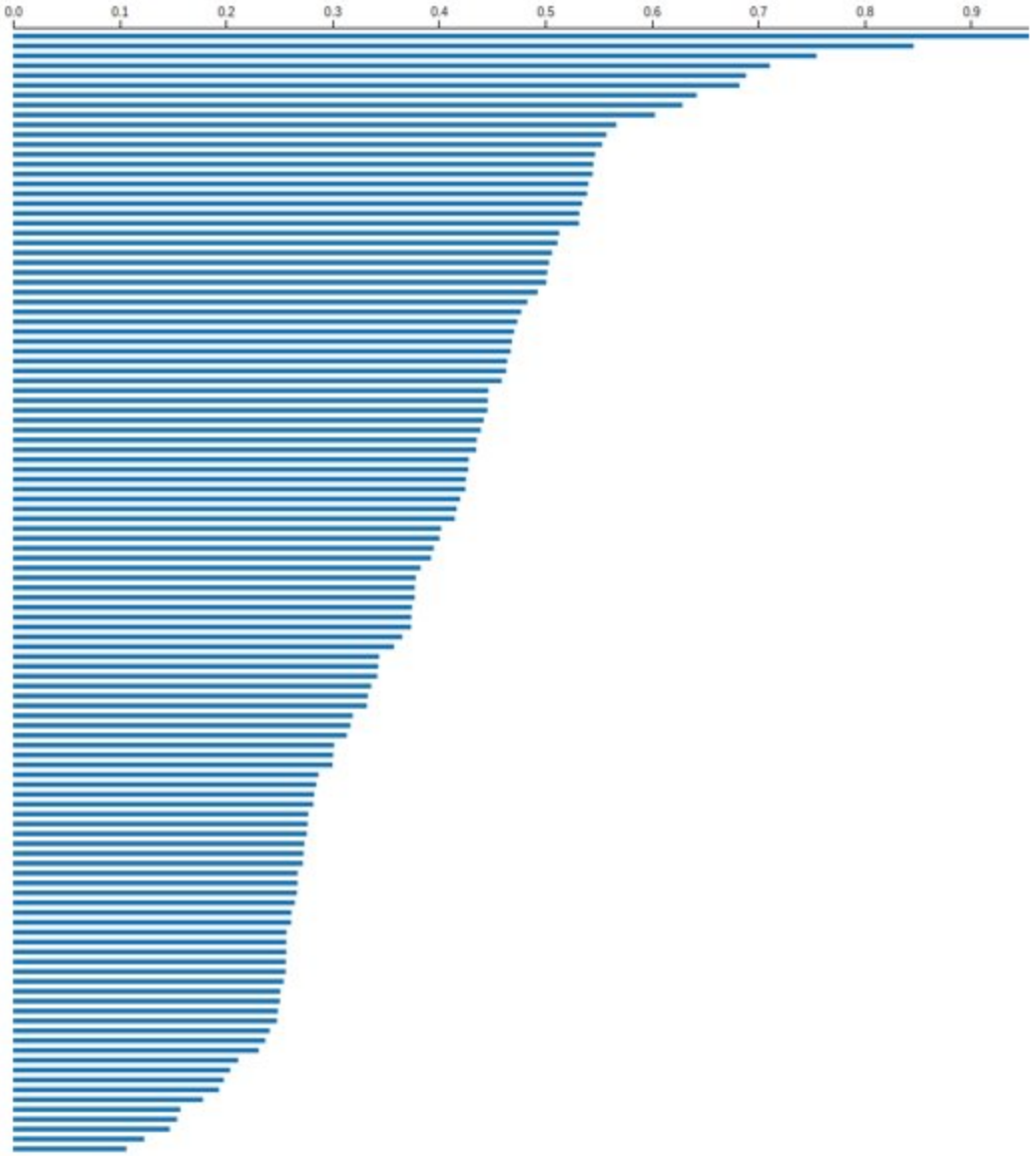
- We normalized the input matrix data within a range of 0 to 1.
- The following formula was used to normalize the data:

$$[x]' = \sqrt{1 - C([x]^T)}$$

- $[x]' = normlized\ matrix$
- $C = correlation\ coefficient$
- $[x]^T = transposed\ input\ matrix\ [x]$

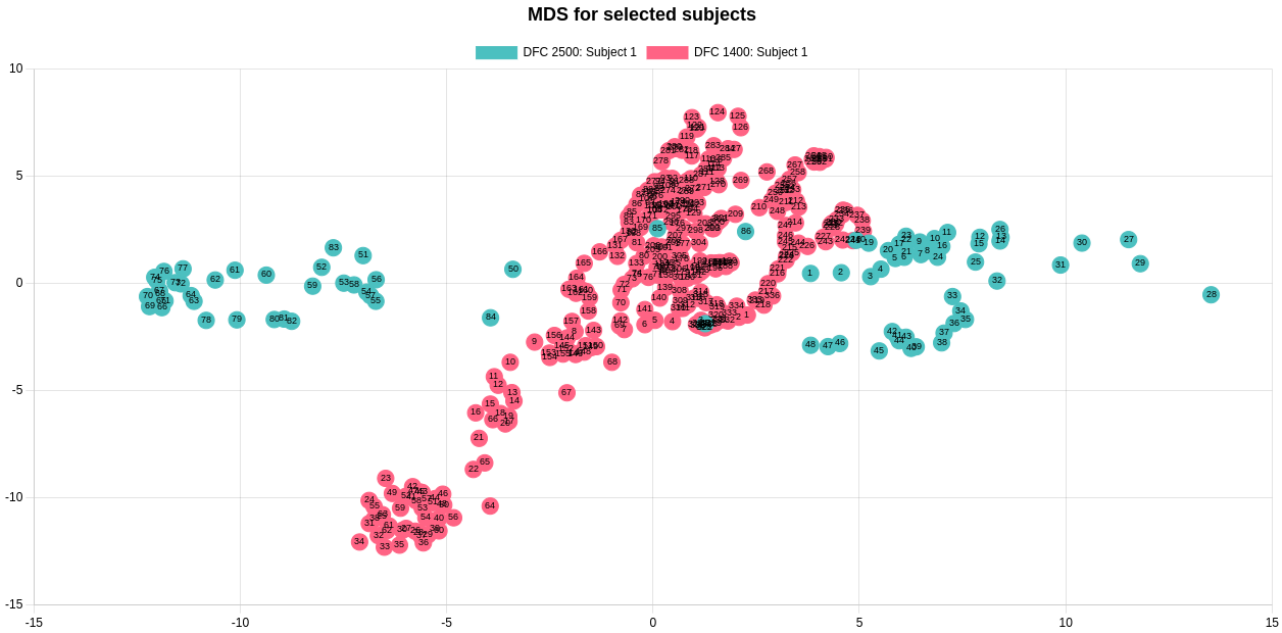
Persistent Homology

- Computed 0-dimensional persistent homology (PH) for time frames of each subjects
- Generated 0-dimensional barcodes from calculated PH value with maximum value of 1
- Generated 316 * 2 JSON files, 1 for each subject for 1-Wasserstein distance matrix of the timeframe barcodes
- Matrix size 86 * 86 and 336 * 336



Multidimensional Scaling

- Applied classical metric Multidimensional scaling (MDS) with precomputed distance (1-Wasserstein)
- It represents a low-dimensional view of the data in which the distances respect well the distances in the original high-dimensional space
- Generated 316 * 2 JSON files, 1 for each subject for MDS matrix
- Matrix size 86 * 2 and 336 * 2



Algorithms Used

We've used 8 clustering algorithms:

- KMEANS
- KMEANS++
- Affinity Propagation
- Birch
- Mean Shift
- Spectral Clustering
- DBSCAN
- OPTICS

Clustering Parameters

- Silhouette Score:

  It is a metric used to calculate the goodness of a clustering techniques.

- For iterative(Kmeans, Kmean++) and graph-based (Spectral) clustering we generated Silhouette Score for the number of clusters ranging from 2 to 15.

- The number with the highest score was picked as the number of cluster.

- For DBSCAN, OPTICS we set the $\in = 1.5$ and $min\_point = 5$ and generate clusters for each fMRI scan.

- For Mean Shift we set the $bandwidth = 2$, BIRCH $n_{cluster} = None, threshold = 1.5$ and For Affinity Propagation we set the $iteration = 200$, with $damping = 0.5$.

Affinity Propagation

- Affinity Propagation creates clusters by sending messages between pairs of samples until convergence.
- A dataset is then described using a small number of exemplars, which are identified as those most representative of other samples.
- The messages sent between pairs represent the suitability for one sample to be the exemplar of the other, which is updated in response to the values from other pairs.
- This updating happens iteratively until convergence, at which point the final exemplars are chosen, and hence the final clustering is given.

Mean Shift

- Mean Shift clustering aims to discover blobs in a smooth density of samples.
- It is a centroid-based algorithm
- It works by updating candidates for centroids to be the mean of the points within a given region.
- Each point try to find its group by moving towards the weighted mean of its local area in each step.
- The destination of each point will be the centroid of the data cluster that the point belongs to.
- All the data points with the same destination point can be labeled with the same cluster.

Spectral Clustering

- Graph based clustering algorithm
- Spectral Clustering performs a low-dimension embedding of the affinity matrix between samples,
- Find the Laplacian matrix of the input matrix by subtracting the adjacency matrix from input matrix.
- First non-zero eigenvalue is called spectral gap which gives us the notion about the density of the graph.
- First large gap of the eigenvalues determines the number of clusters
- Eigenvectors correspond to the eigenvalues determine the actual cluster label.

BIRCH (Balanced Iterative Reducing & Clustering using Hierarchies)

- The Birch builds a tree called the Clustering Feature Tree (CFT) for the given data.
- The data is essentially lossy compressed to a set of Clustering Feature nodes (CF Nodes).
- The CF Nodes have several subclusters called Clustering Feature subclusters (CF Subclusters)
- These CF Subclusters located in the non-terminal CF Nodes can have CF Nodes as children.

Evaluation

- We compare the number of clusters for fMRI of a subject.
- We find the percentage of population with same number of clusters in both fMRI data.
- For mismatched clusters, we find the difference in cluster number between 2 scans and calculated the mean mismatch distance.
- Percentage match and mean mismatch distance was used to evaluate the clustering algorithms.
- K-means++ provides the best performance with 56.33% matches and MMD of 1.28.

Figure OPTICS clustering for Subject 1

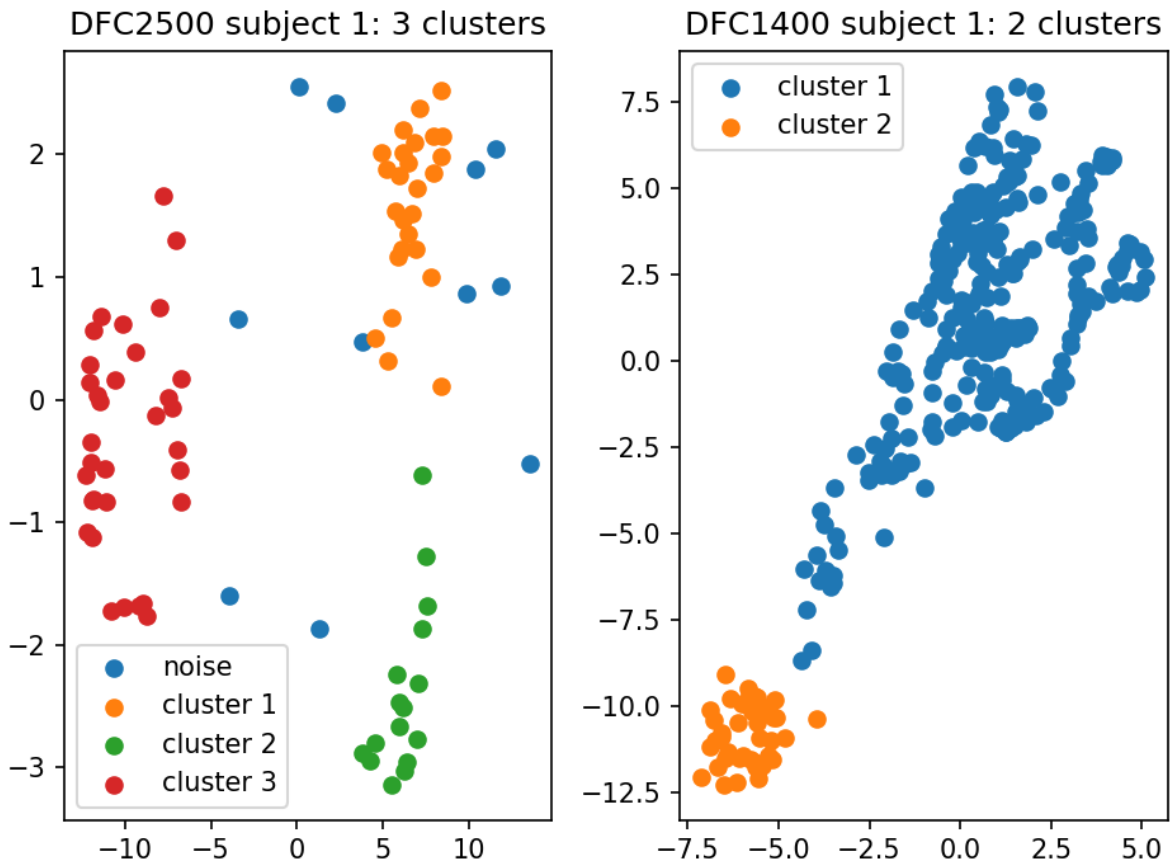


Fig. Clusters generated by OPTICS

Figure KMeans clustering for Subject 1

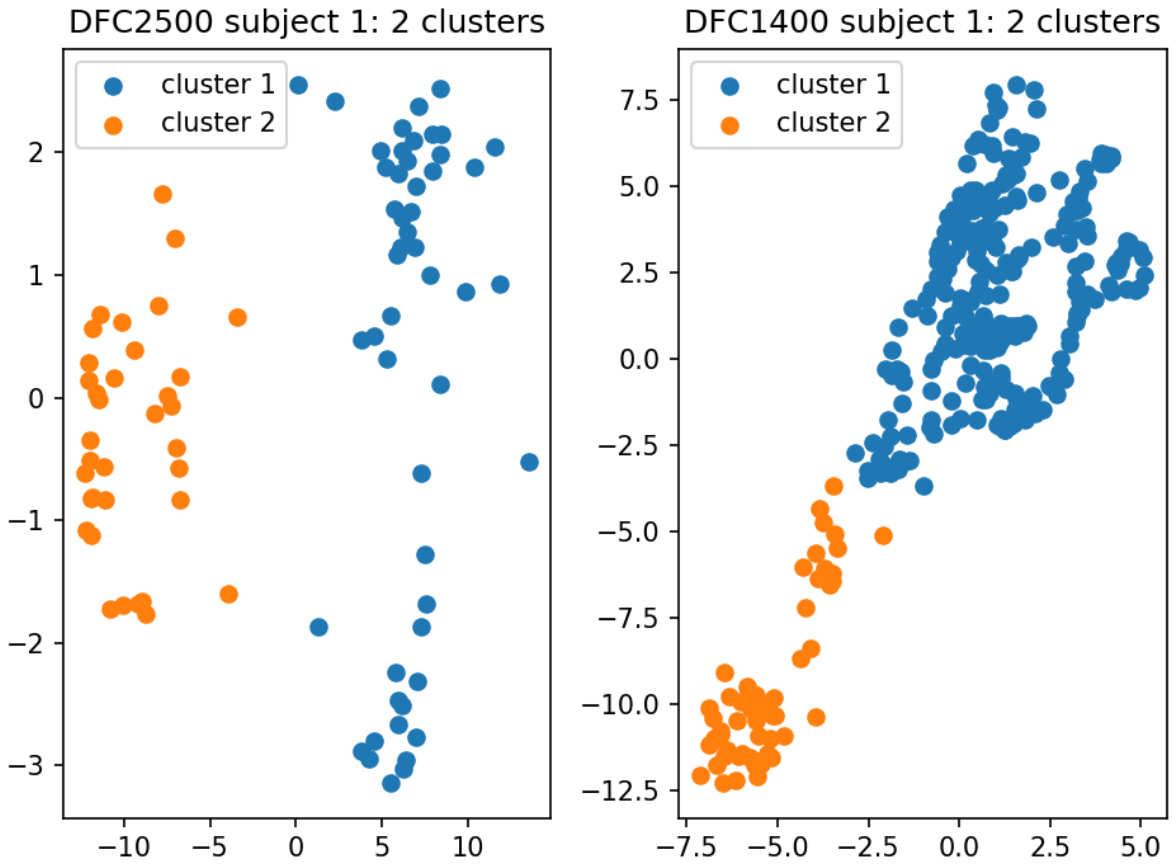


Fig. Clusters generated by KMeans clustering algorithm

Figure DBSCAN clustering for Subject 1

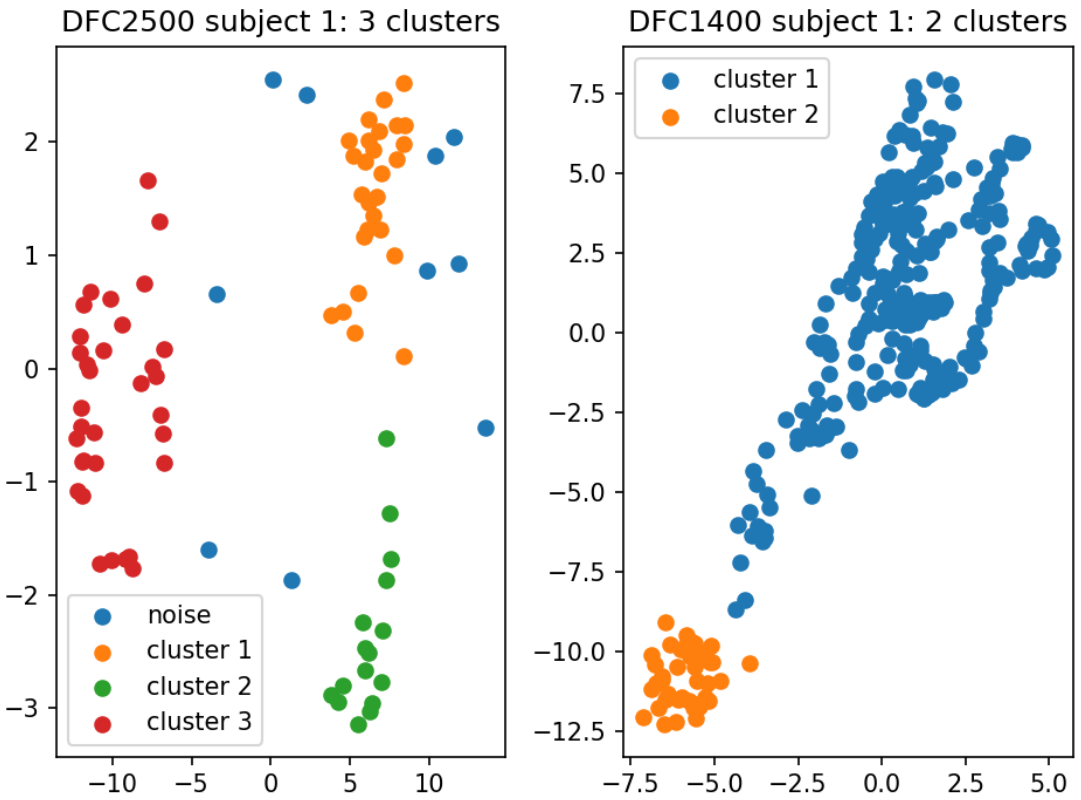


Fig. Clusters generated by DBSCAN clustering algorithm

Figure Spectral clustering for Subject 1

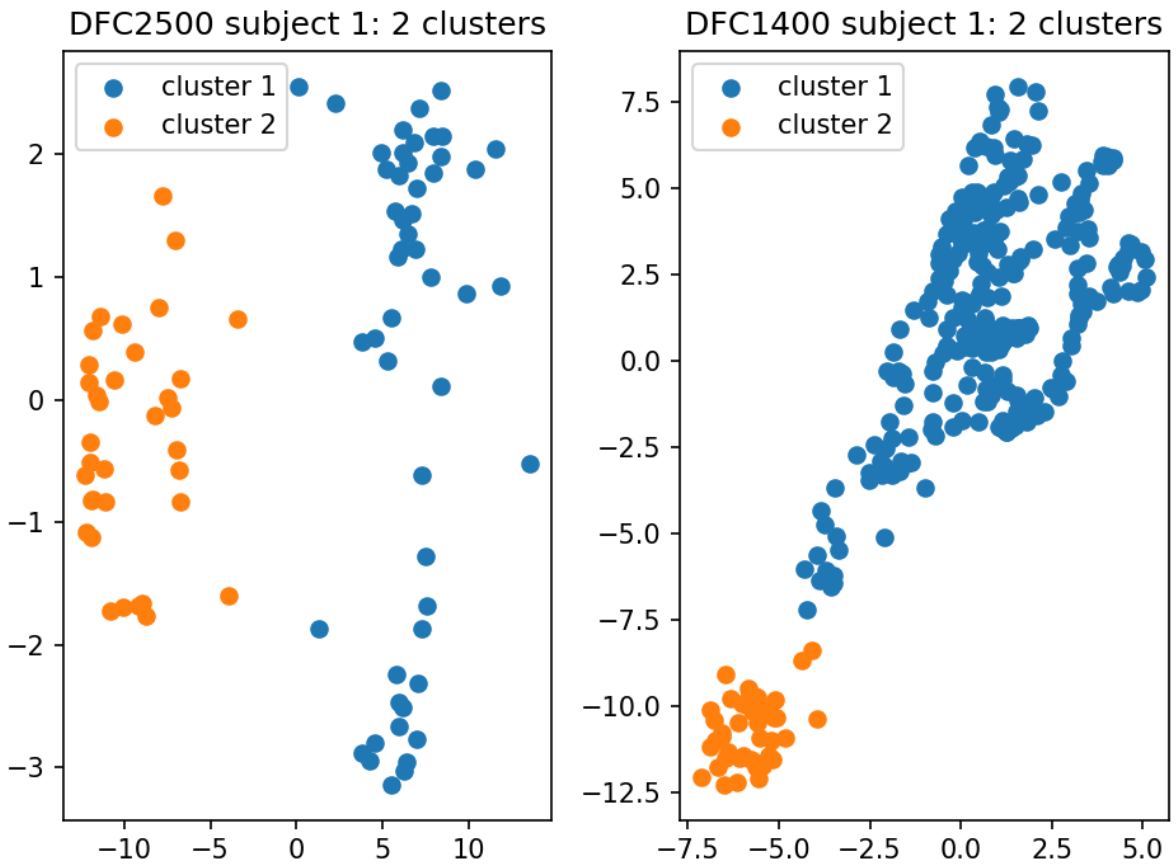


Fig. Clusters generated by Spectral Clustering

Figure Clustering summary


```
Clustering Method: KMeans
Best cluster selection using Silhouette Score in 2-15 range
Total subjects: 316
Total matches: 172
Total match percentage: 54.43
Mean mismatch distance: 1.50

Clustering Method: dbscan
Parameters: eps=1.5, min_samples=5
Total subjects: 316
Total matches: 137
Total match percentage: 43.35
Mean mismatch distance: 0.81
```

```
Clustering Method: spectral
Best cluster selection using Silhouette Score in 2-15 range
Total subjects: 316
Total matches: 161
Total match percentage: 50.95
Mean mismatch distance: 1.58
```

Fig. Clustering result summary using KMeans, DBSCAN, and Spectral

Evaluation

	Method	Algorithm Type	Subjects	Matches	Match (%)	Mean Mismatch Distance
1	K-means	Iterative	316	172	54.43	1.50
2	**K-mean++**	**Iterative**	**316**	**178**	**56.33**	**1.28**
3	Affinity propagation	Hierarchical Clustering	316	1	0.32	6.95
4	Spectral Clustering	Graph based	316	161	50.95	1.58
5	DBSCAN	Density Based	316	137	43.35	0.81
6	OPTICS	Density Based	316	137	43.35	0.81
7	BIRCH	Hierarchical Clustering	316	5	1.58	14.56
8	Mean Shift	Centroid based	316	37	11.08	1.86

Conclusion and Future Work

- Time-varying fMRI data is becoming increasingly important in data analysis.
- Analyze the structural changes of time-varying fMRI data at different time points using unsupervised machine learning techniques.
- Applied different clustering algorithms to all fMRI scans to find their clustered nature.
- Found the difference between the number of clusters over two-time points to log the quality of data changes over time.
- Performance of the clustering result was evaluated by percentage similarity of matches and MMD.
- KMeans++ achieves a maximum of 56.33% matches and 1.28 MMD which outperformed other clustering algorithms we have adopted in this work.
- In future we are planning to apply deep learning-based algorithm.
  - Auto-encoder to analyze quality of clusters.

References

1. Mustafa Hajij, Bei Wang, Carlos Scheidegger, and Paul Rosen. Visual detection of structural changes in time-varying graphs using persistent homology. In 2018 ieee pacific visualization symposium (pacificvis), pages 125–134. IEEE, 2018.
2. J. Han, J. Pei, and M. Kamber. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.
3. C.C. Aggarwal. Data Mining: The Textbook. Springer International Publishing, 2015.
4. RNA Henson. Analysis of fmri time series: Linear time-invariant models, event-related fmri and optimal experimental design. Elsevier, 2003.
5. Ulderico Fugacci, Sara Scaramuccia, Federico Iuricich, and Leila De Floriani. Persistent homology: a step-by-step introduction for newcomers. In STAG, pages 1–10, 2016.
6. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vander- plas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duch-esnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.


THE UNIVERSITY OF
ALABAMA AT BIRMINGHAM
THANK YOU