Summer Report

Bipasha Garg
(2022111006)

July 2024

1 Abstract

This report details the progress and learnings from a topic focused on visualizing cluster data using the BEADS methodology and enhancing it with advanced algorithms and user-friendly features. Initially, the project involved studying Saujanya's thesis to understand clustering algorithms and lp-norm calculations, which laid the groundwork for BEADS visualization. Python was used to develop code for plotting clusters and calculating boundary shapes, followed by the application of the binning method for representing multidimensional data. To improve user interactivity, the project transitioned to incorporating JavaScript.

2 Project Timeline

2.0.1 Read Papers/Thesis

- Initial Phase: Began by reading Saujanya's thesis, focusing on the clustering algorithm and lp-norm calculations. This foundational understanding was crucial for comprehending the basis of BEADS visualization.
- Gained a thorough understanding of clustering algorithms and lp-norm calculations by studying Saujanya's thesis. This provided the theoretical background necessary for BEADS visualization.

2.0.2 Code in Python

- Wrote Python code to plot the BEADS cluster-wise.
- Plotted the boundaries in their respective shapes by calculating their lp-norm values.
- Utilized the algorithm provided in Saujanya's thesis to determine the optimal p value and the boundary coordinates.

2.0.3 Using binning method for multidimensional data

- Applied the binning method to plot multidimensional data after segmenting the radial chart into sectors.
- Leveraged insights and code snippets from the thesis to enhance this process.
- Learned how to segment radial charts into sectors for better data representation.

2.0.4 Shifting to javascript

- Transitioned from Python to a combination of Python and JavaScript to improve the user experience.
- This shift aimed to provide a more interactive and responsive visualization.

2.0.5 Implementing CURE algorithm

- Implemented the CURE (Clustering Using Representatives) algorithm.
- Focused on representing clusters using a few representative points, enhancing the efficiency and clarity of the visualization.

2.0.6 Datasets used

- Iris 2D
- Iris 4D
- Diabetes dataset
- Customer Dataset
- Mall Customer Dataset

3 Dataset Outputs

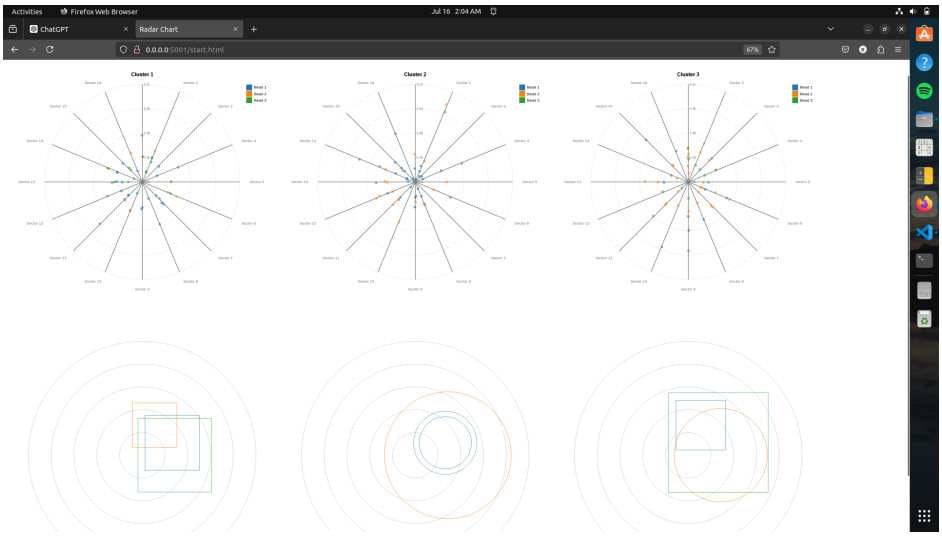

Figure 1: Iris 4D dataset before implementing CURE algorithm

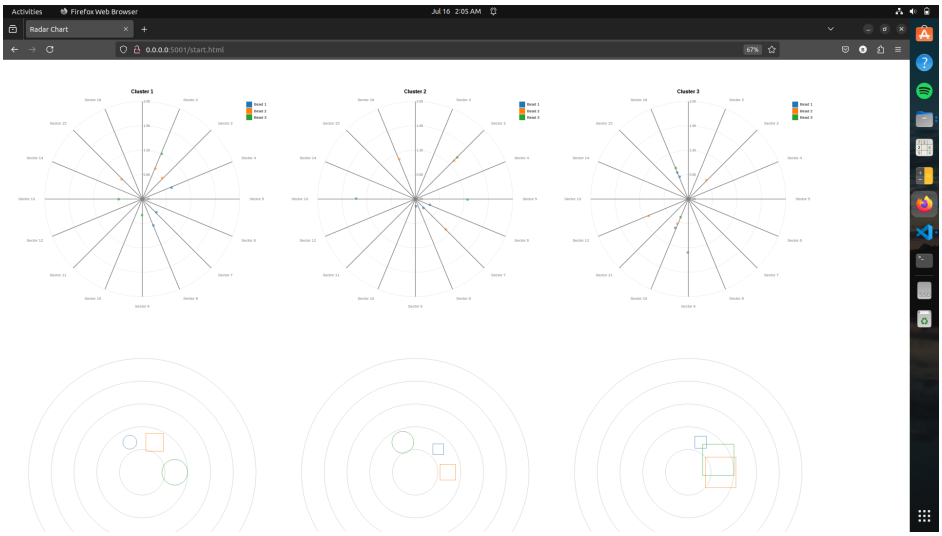

Figure 2: Iris 4D dataset after implementing CURE algorithm

4 Observation

- **Initial Visualization:** The initial visualization of the Iris 4D dataset using the BEADS methodology provided a clear representation of the cluster boundaries and their respective shapes based on lp-norm calculations. This allowed for an accurate depiction of the cluster distributions and their relationships.
- **Binning Method Efficiency:** Implementing the binning method for multidimensional data visualization significantly improved the clarity of the plots. Segmenting the radial chart into sectors helped in managing and representing complex data more effectively, making the visualization more comprehensible.
- **CURE Algorithm Implementation:** The application of the CURE algorithm enhanced the visualization by reducing the number of representative points per cluster. This not only made the plots less cluttered but also maintained the integrity of the cluster shapes, providing a balance between simplicity and accuracy.
- **Dataset Comparisons:** Comparing the dataset visualizations before and after implementing the CURE algorithm highlighted the improvements in efficiency and clarity. The post-CURE visualization depicted more streamlined clusters with fewer representative points.

5 Future Plans

- Implement a zooming feature in the figures to allow users to explore data in greater detail.
- Improve user experience by properly grouping plots and integrating pop-ups for additional information.
- Code improvement: Focus on refining and optimizing the code to enhance performance and maintainability.
- Develop more efficient and effective data filtering techniques to ensure relevant and high-quality data is visualized.
- Work on improving the speed of calculations to provide faster and more responsive data visualizations.
- Testing to be done on as many datasets possible especially larger datasets.