Algorithm and Method Used

In this project, the MiniBatchKMeans clustering algorithm from scikit-learn was utilized. This approach is an optimized variant of the KMeans algorithm, designed specifically for large datasets, efficiently handling the complexity and size inherent in big data projects.

Suitability of MiniBatchKMeans

MiniBatchKMeans is particularly suitable due to its ability to:

- Efficiently cluster large datasets by processing mini-batches of data.
- Rapidly converge, reducing computational resources significantly.
- Adapt well to the known requirement of identifying clusters from datasets with n dimensions, as stipulated by the assignment guidelines.

Handling High-Dimensional Data

The method effectively manages high-dimensional data through:

- Standardization (Normalization): Implemented using StandardScaler to ensure all dimensions contribute equally and prevent features with larger scales from dominating the clustering process.
- Mini-batching: Reduces memory usage and computational complexity, facilitating the clustering of large-scale, multi-dimensional datasets.

Preprocessing and Hyperparameters

Key preprocessing steps and hyperparameters used include:

- Normalization: Applying StandardScaler to ensure dimensional homogeneity.
- Cluster Count: Explicitly set to 4n -1, fit the requirement.
- Initialization and Random Seed: Utilized n_init=100 and random_state=42 for reproducibility and stable cluster assignments.

Conclusion

The MiniBatchKMeans algorithm, combined with proper data normalization and thoughtful hyperparameter selection, demonstrates effectiveness and scalability, essential for analyzing and classifying extensive particle accelerator datasets. This methodology meets the outlined requirements, clearly clustering data into meaningful groups suitable for deeper scientific exploration.

GitHub Repository

https://github.com/r1407p/NTU-bigdata