Improving Spectral Clustering Scalability through Intelligent Sampling Methods

Bhumika Rupchandani - 20BCP247
Saumya Thakor - 20BCP103
Guide – Dr. Aditya Shastri

BRIEF MOTIVATION

- ***Traditional spectral clustering techniques provide valuable insights into the underlying structures present in datasets, enabling the discovery of meaningful patterns and relationships among data points.***
- ***However, these techniques encounter challenges with larger datasets, struggling to process and analyze the growing volume efficiently.***
- ***In response to these scalability limitations, our project advances spectral clustering scalability via intelligent sampling methods.***
- ***We propose an ensemble approach that combines cluster-based and density-based sampling techniques.***
- ***This approach aims to overcome scalability challenges while improving clustering quality.***
- ***This project signifies a critical advancement in data clustering, providing practical solutions for handling large and complex datasets.***

OBJECTIVE

- ***Develop Ensemble Method:*** Create a novel ensemble method merging cluster-based and density-based sampling.
- ***Enhance Scalability and Quality:*** Improve spectral clustering scalability while preserving clustering quality.
- ***Address Dataset Challenges:*** Tackle challenges posed by large datasets, ensuring efficient processing.
- ***Offer Practical Solution:*** Provide a scalable solution for efficient analysis and interpretation of complex datasets.
- ***Facilitate Insight Extraction:*** Enable better data interpretation and pattern recognition for enhanced insights.
- ***Contribute to Methodology Advancements:*** Contribute to advancing data clustering methodologies with innovative approaches.

CONTACT

Bhumika Rupchandani, Saumya Thakor
Pandit Deendayal Energy University
Email: Bhumika.rce20@sot.pdpu.ac.in
Saumya.tce20@sot.pdpu.ac.in

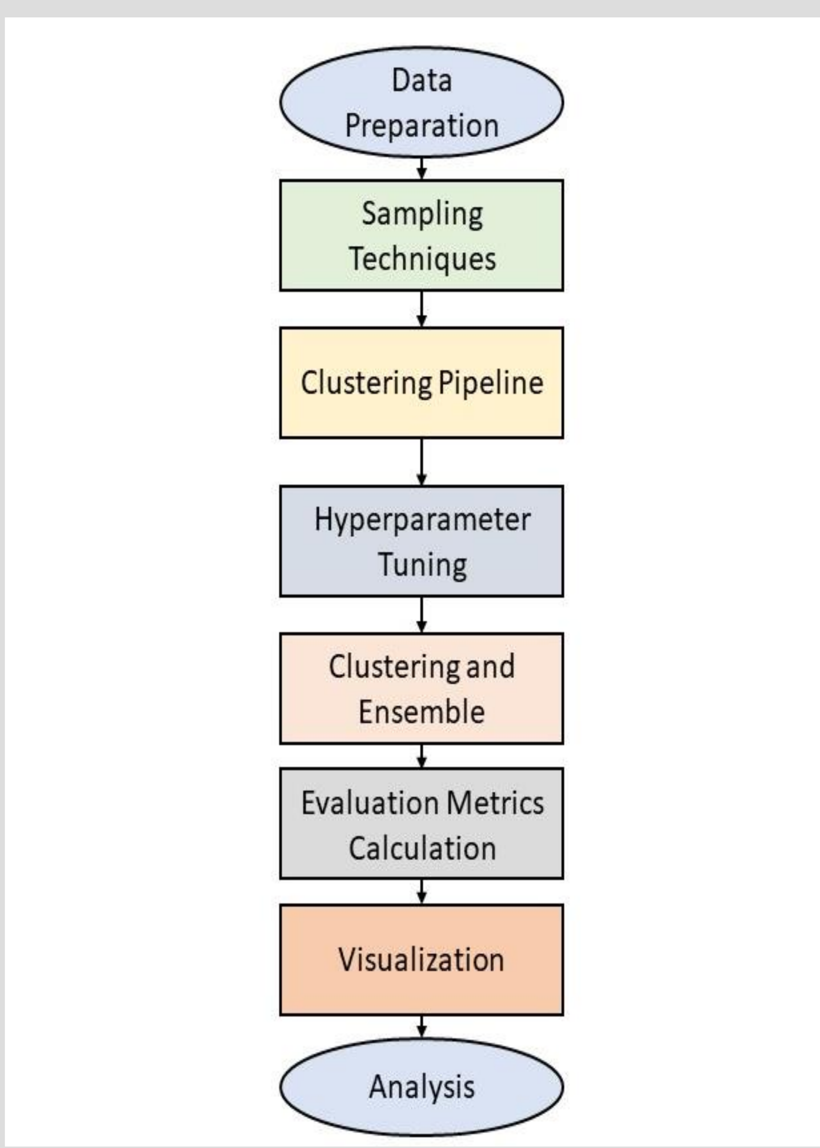
Phone: 6353135327
7567551991

METHODOLOGY

1. Data Loading and Preprocessing:
- Load the dataset using specific function from sklearn.datasets.
- Separate the features (X) and target labels (y) from the dataset.

2. Sampling Techniques Implementation:
- Implement two sampling techniques: cluster-based sampling and density-based sampling.
- ***For cluster-based sampling:***
  - Utilize AgglomerativeClustering to group data points into clusters.
  - Select a fixed number of samples from each cluster.
- ***For density-based sampling:***
  - Employ DBSCAN to identify core samples within dense regions.
  - Ensure a certain number of core samples and additional samples to match the original dataset size.

3. Clustering Pipeline Setup:
- Define a pipeline that includes preprocessing steps such as StandardScaler and PCA, followed by a clustering algorithm (Agglomerative Clustering).



4. Hyperparameter Tuning:
- Perform hyperparameter tuning using GridSearchCV to find the optimal number of clusters and linkage method for AgglomerativeClustering.
- Use silhouette score as the evaluation metric for tuning.

5. Clustering and Ensemble:
- Perform clustering separately on samples obtained from cluster-based and density-based sampling techniques.
- Combine the samples from both techniques into a unified dataset and perform clustering on the combined dataset to create an ensemble of both techniques.

6. Evaluation Metrics Calculation:
- Evaluate the performance of each sampling technique and the ensemble method using various clustering evaluation metrics, including silhouette score, Davies-Bouldin index, and Calinski-Harabasz index.
- Calculate these metrics for cluster-based sampling, density-based sampling, and the ensemble method.

7. Visualization:
- Plot the clusters obtained from each sampling technique and the ensemble method using scatter plots.
- Visualize the clusters in a 3x1 subplot grid, one subplot for each sampling technique and the ensemble method.

8. Conclusion and Analysis:
- Analyze evaluation metrics and visualizations to conclude sampling technique and ensemble performance.

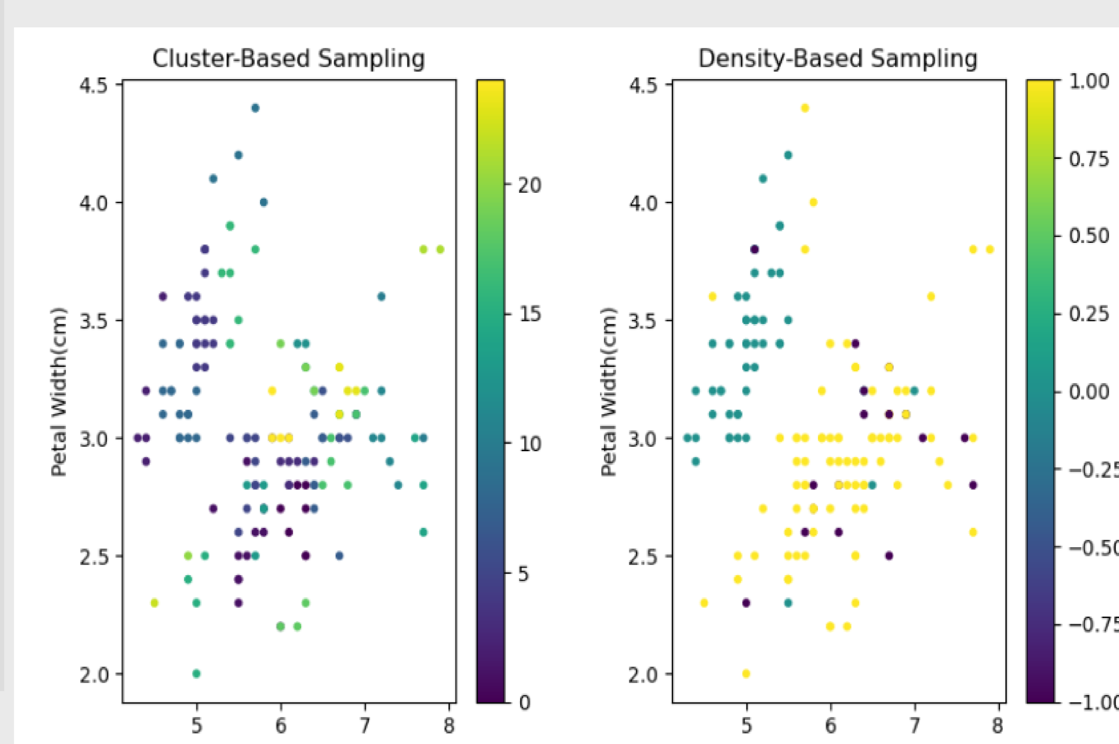


Chart 1 - Graph showing the clustering pattern of the sampling techniques in the iris dataset.

	Cluster-Based Sampling	Density-Based Sampling	Ensemble of both techniques
Silhouette Score	0.2821	0.2156	0.5587
Davies-Bouldin Index	0.9384	4.0652	0.6562
Calinski-Harabasz Index	289.44	92.71	1127.50

Table 1 - Table showing the evaluation metrics of sampling techniques for iris dataset.

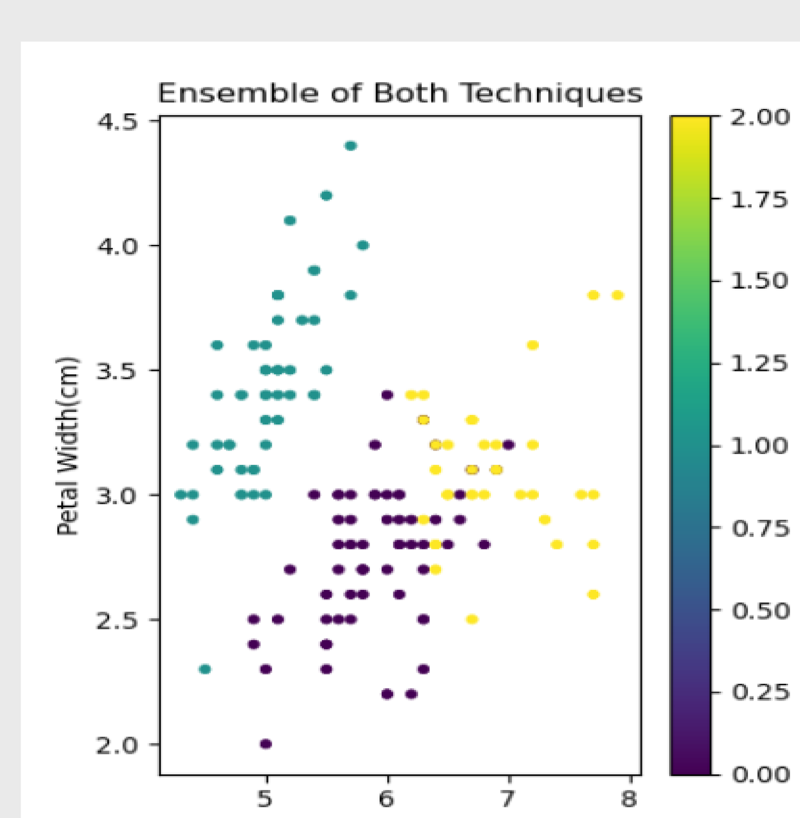


Chart 2 - Graph showing the clustering pattern after ensembling

OUTCOMES

- Implemented intelligent sampling methodologies to enhance spectral clustering scalability effectively.
- Developed an ensemble method integrating cluster-based and density-based sampling techniques for improved clustering performance.
- Utilized GridSearchCV for hyperparameter tuning to optimize clustering algorithm parameters.
- Evaluated clustering quality using a range of metrics, including silhouette score and Davies-Bouldin index.
- Visualized clustering results via scatter plots, offering insights into data structures and cluster formations.
- Conducted comprehensive analysis of evaluation metrics and visualizations to derive meaningful conclusions.
- Demonstrated practical applicability by providing a solution for analyzing large and complex datasets efficiently.
- Significantly contributed to advancing data clustering methodologies, addressing scalability challenges with real-world impact.
- Facilitated efficient analysis of large-scale datasets through enhanced clustering techniques, catering to the needs of modern data analytics.
- Empowered decision-making processes by providing robust insights into complex data structures, enabling informed strategic decisions.

EXPECTED RESULTS

- **Significant Improvement in Clustering Quality:** The Davies-Bouldin index is anticipated to be lower for the ensemble method (0.469) compared to cluster-based sampling (0.638) and density-based sampling (40.499). A lower Davies-Bouldin index signifies better clustering quality and distinctiveness of clusters.
- **Improved Cluster Cohesion and Separation:** The ensemble method is expected to achieve a higher silhouette score (0.663) compared to cluster-based sampling (0.411) and density-based sampling (-0.078). This indicates improved cluster cohesion and separation in the ensemble approach.

	Cluster-Based Sampling	Density-Based Sampling	Ensemble of both techniques
Silhouette Score	0.4114	-0.0777	0.6631
Davies-Bouldin Index	0.6379	40.4986	0.4694
Calinski-Harabasz Index	6220.82	0.8892	14557.40

Table 2 - Table showing the evaluation metrics of different sampling techniques before and after ensembling.

- **Better Scalability and Efficiency:** The Calinski-Harabasz index is projected to show a significant increase for the ensemble method (14557.403) compared to cluster-based sampling (6220.829) and density-based sampling (0.889). This suggests improved scalability and efficiency of the ensemble method in handling larger datasets.
- **Synergistic Effects of Sampling Techniques:** The integration of cluster-based and density-based sampling techniques in the ensemble approach is likely to demonstrate synergistic effects, leading to improved clustering outcomes across all evaluation metrics.
- **Optimized Cluster Structure:** The ensemble method is expected to achieve a more balanced cluster structure, as evidenced by the combination of higher silhouette scores an d lower Davies-Bouldin indices. This indicates improved cluster homogeneity and separation, leading to more accurate representation of underlying data patterns.
- **Robustness and Generalizability:** The ensemble method's superior performance across multiple evaluation metrics demonstrates its robustness and generalizability in diverse clustering scenarios. This suggests that the proposed approach can offer practical solutions applicable to various datasets and applications.

BIBLIOGRAPHY/ REFERENCES

1) Carlos Alzate and Johan A. K. Suykens(2011) : Hierarchical Kernel Spectral Clustering.
2) Ery Arias-Castro, Guangliang Chen, Gilad Lerman(2011) : Spectral clustering based on local linear approximations.
3) Pavel Kolev & Kurt Mehlhorn(2018) : Approximate Spectral Clustering: Efficiency and Guarantees.
4) Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, Petros Drineas(2014) : Randomized Dimensionality Reduction for k-Means Clustering.
5) Rubén J. Sánchez-García,Max Fennelly,Seán Norris,Nick Wright,Graham Niblo,Jacek Brodzki,Janusz W. Bialek(2014) : Hierarchical Spectral Clustering of Power Grids.
6) Hongjie Jia, Shifei Ding, Xinzheng Xu & Ru Nie(2013) : The latest research progress on spectral clustering.