DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

National Institute of Technology Patna

Improving Ensemble Clustering by Unified Diversification and Result Optimization

by

Bhola Das (2106038)
Navneet Kumar(2106045)
Rohini Kumari (2106008

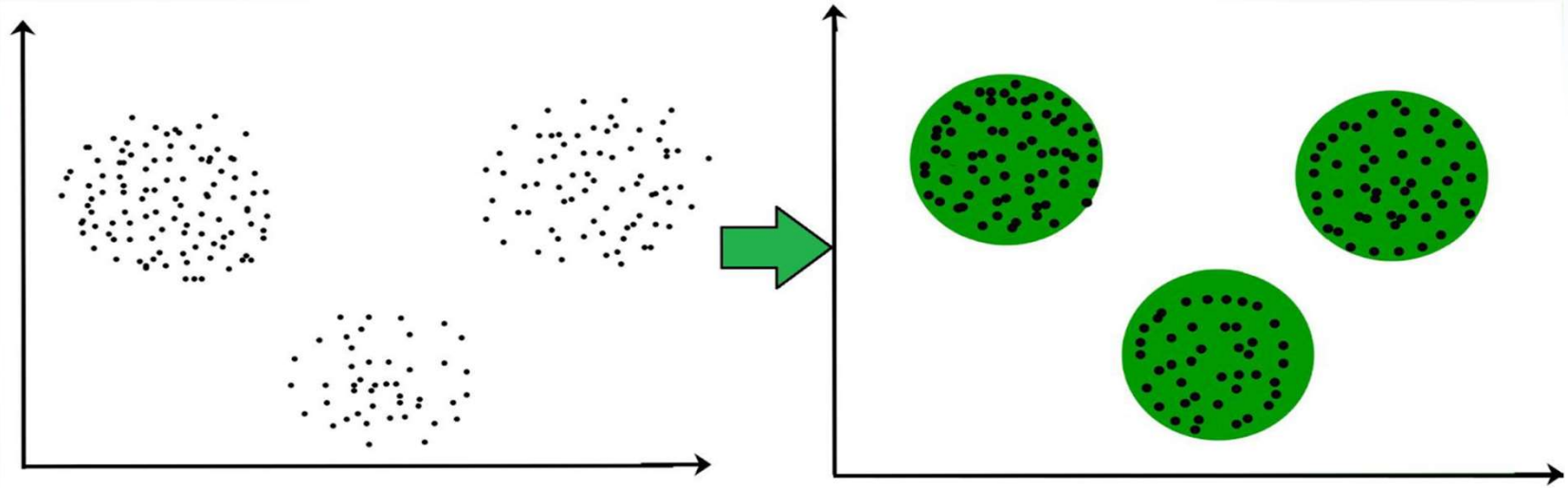
Under the supervision of

Dr. Krishna Kumar Sethi
Assistant Professor
Computer Science and Engineering
NIT Patna (Mahendru – 800005)

List of Contents:

Clustering:

Clustering is an unsupervised machine learning algorithm that organizes and classifies different objects, data points, or observations into groups or clusters based on similarities or patterns



Clustering Techniques: Basic vs. Advanced

1. Basic Clustering Techniques:

These algorithms are efficient for large datasets and relatively easy to implement.

Parition Based clustering: It partitions the data into K distinct clusters based on distance to the centroid of a cluster

Hierarchical clustering: Builds a hierarchy of clusters by merging or splitting data points based on similarity

Density-based clustering: Identifies clusters based on areas of high data point concentration.

Advanced Clustering Techniques:

These techniques address limitations of basic methods and handle complex data scenarios.

Variations of basic algorithms: Fuzzy C-Means (overlapping clusters), Hierarchical K-means (combines K-means with hierarchical clustering).

Ensemble clustering: Combines multiple clustering for a more robust solution

Model-based clustering: Assumes an underlying statistical model for the data in each cluster.

Ensemble Clustering

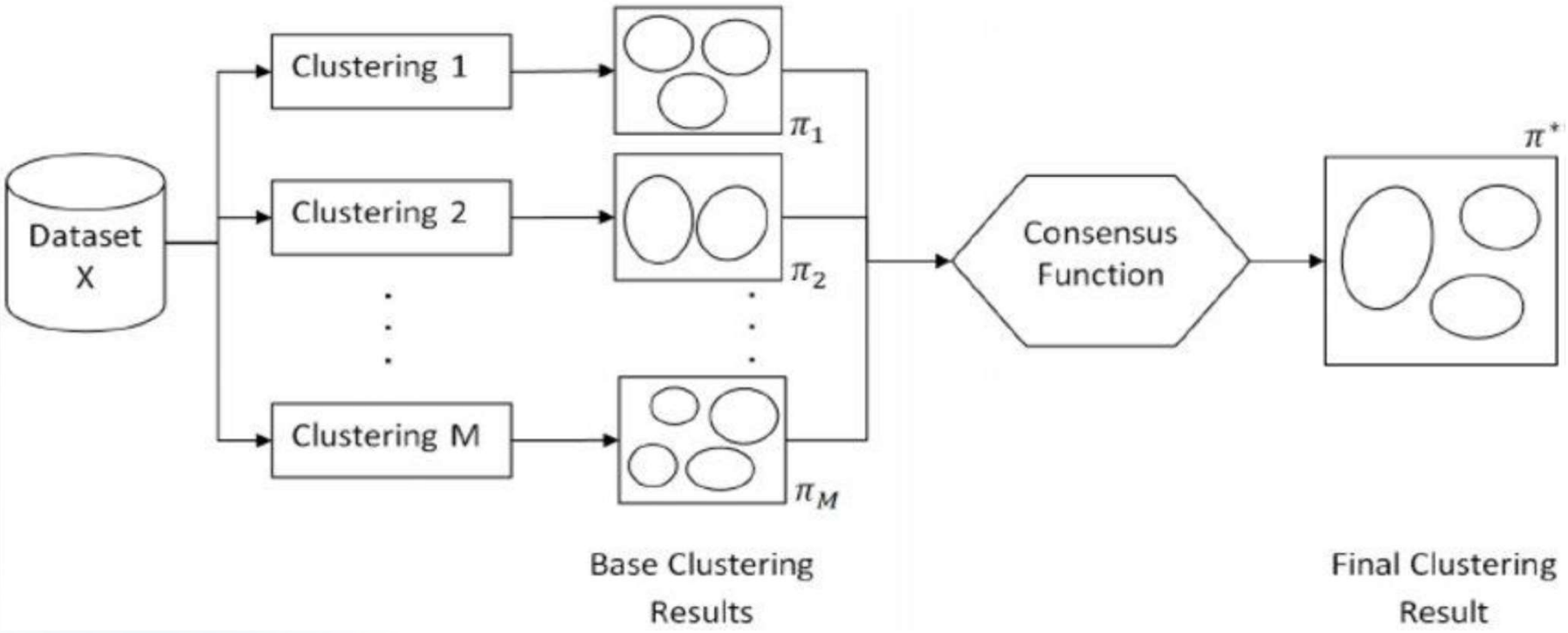
It is the process of combining results from multiple clustering algorithms to create a single , robust and potentially more accurate clustering solution.

Also known as consensus clustering, applies the ensemble approach to the unsupervised learning task of clustering

It aims to leverage the diversity and complementary strengths of different clustering methods

It consist of various approaches, including consensus-based methods, integration of diverse algorithms, and ensemble model selection strategies

Ensemble Clustering Model



Homogeneous vs. Heterogenous Ensemble clustering:

Feature	Homogeneous Ensemble	Heterogeneous Ensemble
Definition	Ensemble of multiple clustering algorithms of the same type.	Ensemble of multiple clustering algorithms of different types.
Approach	Combines multiple instances of the same clustering algorithm.	Combines diverse clustering algorithms.
Diversity	Achieved through variations in parameter settings or data subsets.	Achieved through different clustering algorithms with distinct characteristics.
Example	Ensemble of multiple K-means or hierarchical clustering instances.	Ensemble of K-means, DBSCAN, and hierarchical clustering algorithms.

Why Ensemble Clustering?

- It combines the strength of individual clustering model , enhance robustness and provide more reliable result
- By combining different algorithms, it can handle diverse data patterns and adapt to various scenarios.
- They are effective in identifying complex and non-convex clusters that may be missed by individual algorithms.
- It aims to find a consensus among individual clusterings. It helps identify common structures across different partitions, leading to more consistent and reliable results
- Consensus-based approaches reduce the risk of getting stuck in local optima

Phase Generation:

Phase 1: Generating Diverse Base Clustering:

- Preprocessing data for optimal clustering .Employing various clustering algorithms with diverse parameter settings.
- Running algorithms multiple times with different initializations or cluster numbers.
- Storing resulting base clustering with different assignments and cluster numbers

Phase 2: Consensus Function and Final Clustering:

- The consensus function takes all the base clustering as input and aims to find a single, "consensus" clustering that best reflects the agreement across the diverse base clustering
- **Final Clustering:** Based on the chosen consensus function, a final clustering is generated. This clustering represents a more robust and potentially more accurate grouping of the data points compared to any single base clustering

Categories of Consensus Function-

Voting-Based Approach:

- Treats each base clustering as a "vote" for data point grouping.
- Final clustering determined by majority vote across base clustering
- Effective when base clustering exhibit high agreement

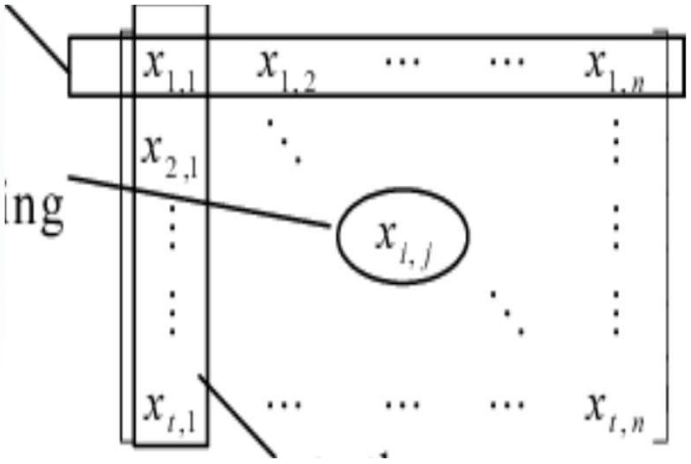
Hypergraph Partitioning Approach:

- Represents data points as vertices and co-occurrences in clusters as hyperedges.
- Aims to partition hypergraph into subgraphs with frequently co-clustered vertices
- Includes algorithms such as Hypergraph Spectral Clustering (HSPC) and Affinity Propagation(AP)

Co-Association Matrix:

- A co-association matrix acts as an intermediate result, storing information about the co-occurrence of data points in various clusters generated by the individual base clustering algorithms.

- The co-association matrix is a square matrix with dimensions N x N, where N is the number of data points in dataset. Each element cij in the matrix represents the co-association score between data points i and j. This score indicates how frequently these data points appear in the same cluster across the different base clustering



Literacy Survey:

1.From clustering to clustering ensemble selection:-

Ref[14]

Provides a comprehensive overview of data clustering, covering its historical development and basic techniques. It delves into clustering ensemble methods, exploring mechanisms for generating ensembles and consensus functions.

".Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence*, *104*, 104388"

2.Combining multiple clustering using evidence accumulation:

{Ref[4]}

The proposed Adaptive Clustering Ensemble algorithm combines multiple clustering models to enhance consistency and quality. It employs cluster similarity and a novel membership similarity. The method iteratively refines clusters, handling uncertain objects

".Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, *27*(6), 835-850".

3.Clustering ensemble method:

Ref[6]

The Evidence Accumulation (EAC) approach combines multiple clustering results to extract a consistent clustering, treats each partition as independent evidence, voting to create a similarity matrix . The method is evaluated using mutual information and bootstrapping

"Alqurashi, T., & Wang, W. (2019). Clustering ensemble method. *International Journal of Machine Learning and Cybernetics*, *10*, 1227-1246.".

4.K-Means-Based Consensus Clustering:-

{Ref[4,5]}

Propose a novel method centered around K-means based consensus clustering (KCC). It relies on a specifically designed utility function to measure how well a consensus clustering aligns with the original individual clustering.

".Wu, J., Liu, H., Xiong, H., Cao, J., & Chen, J. (2014). K-means-based consensus clustering: A unified view. *IEEE transactions on knowledge and data engineering*, *27*(1), 155-169"

5.Ensemble Clustering via Co-Association Matrix Self-Enhancement:

Ref[4]

Proposing a unified framework that simultaneously tackles the challenges of diversifying base clusters and optimizing the storage of intermediate result, it aims to enhance the performance, scalability, and robustness of ensemble clustering algorithms

"Jia, Y., Tao, S., Wang, R., & Wang, Y. (2023). Ensemble clustering via co-association matrix self-enhancement. *IEEE Transactions on Neural Networks and Learning Systems*"

Problem statement -

The problem in ensemble clustering overlook the dual challenges of effectively diversifying base clusters and efficiently managing intermediate results .

1. Methods prioritize integrating diverse clustering algorithms to enhance representation, but overlook the storage of intermediate results.
2. Focus on efficient storage but neglect the crucial task of diversifying base clusters to capture dataset complexity.

This disjointed approach hinders the development of ensemble clustering algorithms that excel in both representation and storage optimization simultaneously.

Implementation:

Develop a strategy to generate a set of base clustering for diversifying base cluster

We introduce a novel strategy aimed at diversifying the ensemble of base clusters in the initial phase of ensemble clustering. This involves utilizing a variety of clustering algorithms to generate a diverse set of base clusters,enhancing the representation and robustness of the ensemble..

Introducing a Probabilistic Co-Association Matrix Model for Ensemble Clustering

The co-association matrix captures the probability of data points A and B being assigned to the same cluster across different clustering solutions with varying numbers of clusters (k). The matrix accumulates these probabilities for all possible k values based on the initial cluster assignments

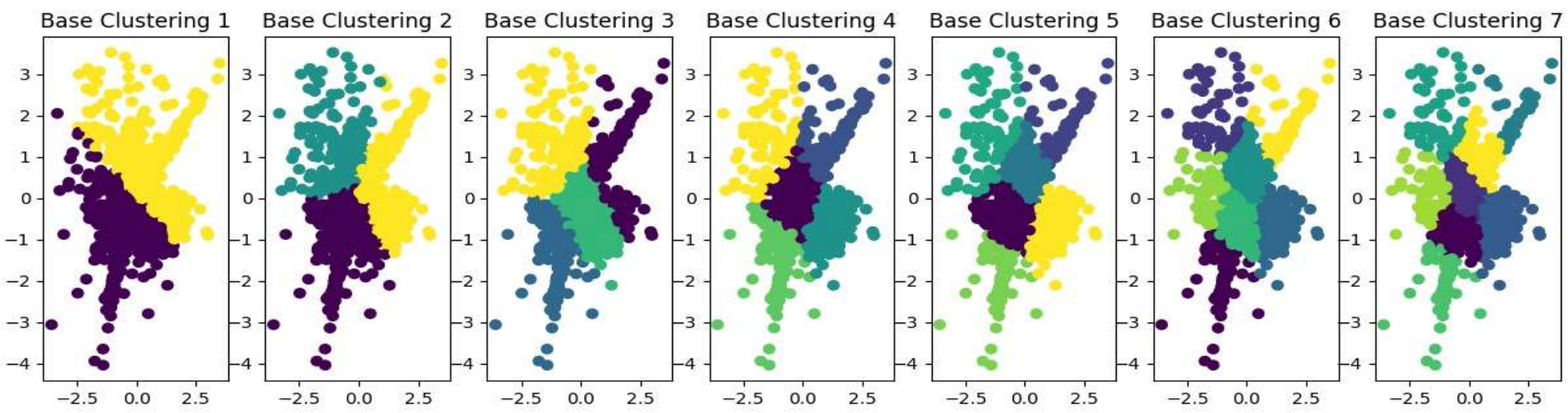
Calculating Co-association Matrix using co-occurrence


```
def calculate_co_association_probability(base_cluster_solutions):

  n_data =  len(base_cluster_solutions[0])
  n_clusters =  len(base_cluster_solutions)

  # Create the co-association matrix
  co_assoc_matrix = np.zeros((n_data, n_data))

  co_assoc_matrix = np.zeros((n_data, n_data))
  for i in range(n_data):
      for j in range(i, n_data):
        for k in range(n_clusters):
          if base_cluster_solutions[k][i] == base_cluster_solutions[k][j]:
            co_assoc_matrix[i][j] += (pro_value[k])
            co_assoc_matrix[j][i] += (pro_value[k])
          else:
            co_assoc_matrix[i][j] += 1
            co_assoc_matrix[j][i] += 1

  return co_assoc_matrix
```

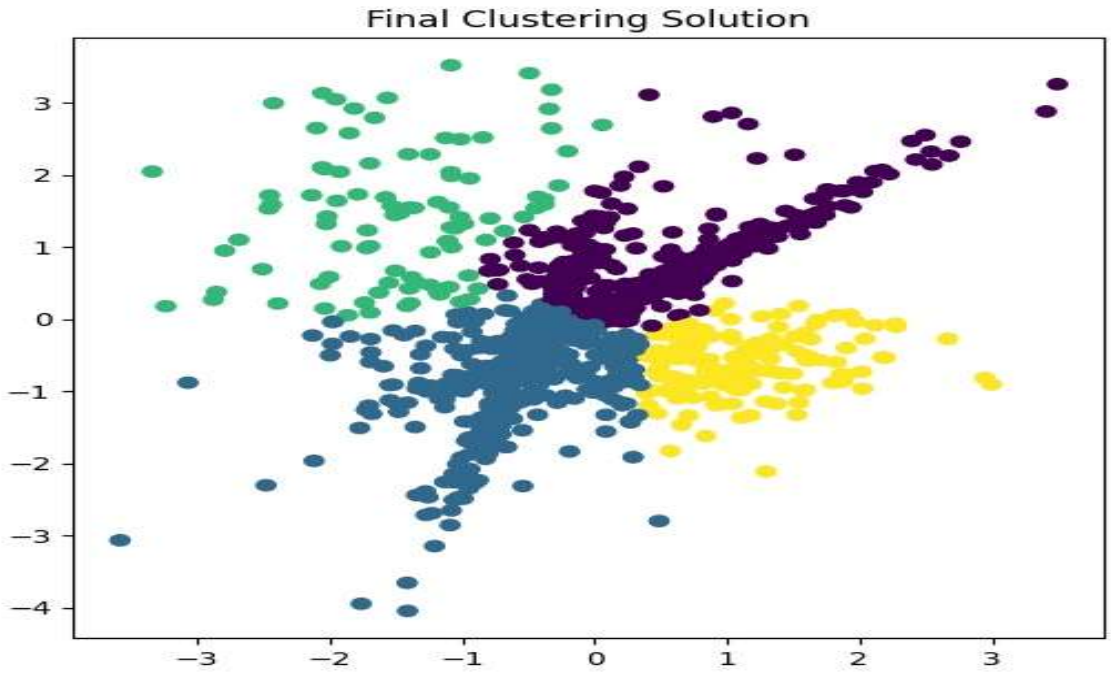
Calculating Co-association Matrix using Probability

```
def calculate_co_association_with_co_occur(base_cluster_solutions):

  n_data =  len(base_cluster_solutions[0])
  n_clusters =  len(base_cluster_solutions)

  # Create the co-association matrix
  co_assoc_matrix = np.zeros((n_data, n_data))

  co_assoc_matrix = np.zeros((n_data, n_data))
  for i in range(n_data):
      for j in range(i, n_data):
          val = sum(base_cluster_solutions[k][i] == base_cluster_solutions[k][j] for k in range(n_clusters))
          co_assoc_matrix[i][j] = co_assoc_matrix[j][i] = val

  return co_assoc_matrix
```

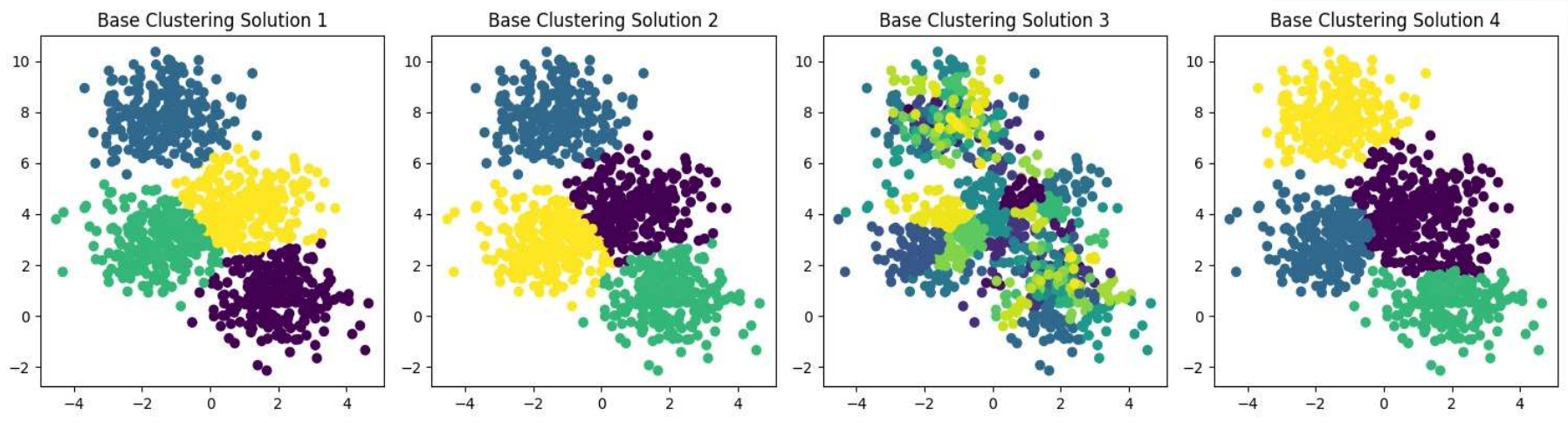
Base clustering solution using different value of k



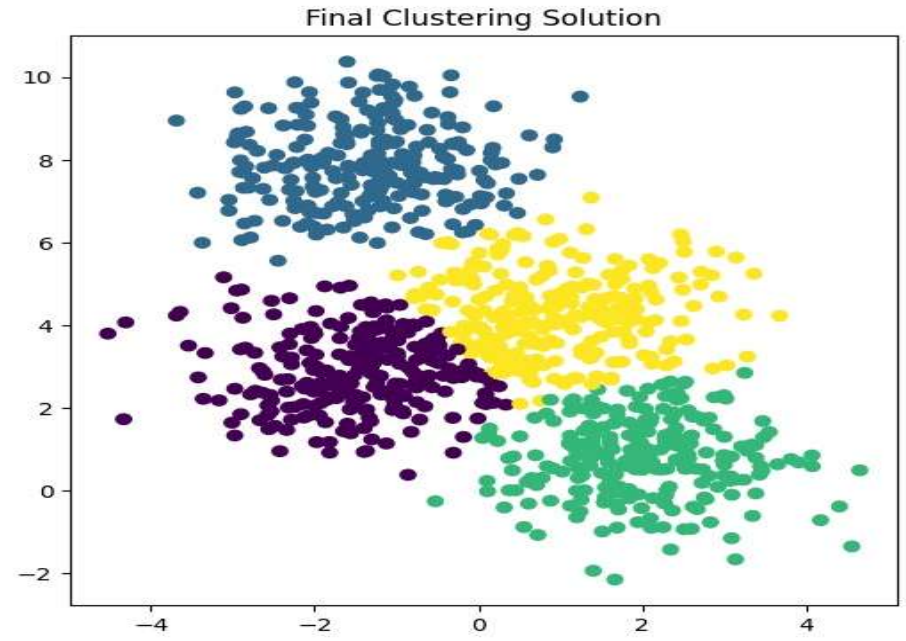
Final clustering solution using co-occurrence co-association matrix



Different base clusters using different Clustering algorithms



Final clustering solution using probabilistic coassosiation matrix



Evaluation Metric:-

Silhouette Score: It quantifies the cohesion and separation of clusters, where a higher score indicates well-separated clusters with instances tightly packed within clusters .

Calinski_harabasz_score : It measures the ratio of between-cluster dispersion to within-cluster dispersion, indicating the compactness and separability of clusters.

Davies-Bouldin Score: It assesses the average similarity between each cluster's center and its closest cluster center, providing insights into the clustering's effectiveness

Result of Evaluation :-

S.No	Evaluation Metric	Probability Based Co-association Matrix (P)	Co-occurrence Matrix (C)
01	Silhouette Score	0.3916000724961737	0.38811479506223295
02	**calinski_harabasz_score**	457.03394351840274	449.7002524337332
03	Davies-Bouldin Score	0.7903649988985719	0.8076498069262531

Conclusion:

This project achieved two main goals:

1. **Diversified Cluster Analysis**: We explored the data using a variety of clustering algorithms, including density-based, centroid-based, hierarchical-based, and distribution-based methods. To identify the optimal configuration for our data, we experimented with different numbers of clusters for each algorithm.

2. **Enhanced Co-association Matrix**: We improved the co-association matrix by incorporating the probability of a data point being assigned to the same cluster. This refined representation captures the likelihood of co-occurrence between data points more effectively

Future Scope:

Attention-Based Co-Association Matrix for Improved Clustering:
This work proposes an attention mechanism inspired by neural networks. It assigns weights to data point pairs based on their relative importance, capturing intricate connections. This refined co-association matrix is then fed into a clustering algorithm, potentially leading to more accurate data grouping

Self-Organizing Maps for Enhancing Co-Association Matrix
proposes using Self-Organizing Maps (SOMs) to enhance co-association matrices in clustering. SOMs learn data structure, and distances between neurons in the trained SOM become the refined co-association matrix, potentially improving cluster accuracy

References:

1. Jia, Y., Tao, S., Wang, R., & Wang, Y. (2023). Ensemble clustering via co-association matrix self-enhancement. *IEEE Transactions on Neural Networks and Learning Systems*

2.Wu, J., Liu, H., Xiong, H., Cao, J., & Chen, J. (2014). K-means-based consensus clustering: A unified view. *IEEE transactions on knowledge and data engineering*, *27*(1), 155-169.

3.Wu, Z., Cai, M., Xu, F., & Li, Q. (2024). PCS-granularity weighted ensemble clustering via Co-association matrix. *Applied Intelligence*, 1-18

4.Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. *Engineering Applications of Artificial Intelligence*, *104*, 104388

5.Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, *27*(6), 835-850.

6.Alqurashi, T., & Wang, W. (2019). Clustering ensemble method. *International Journal of Machine Learning and Cybernetics*, *10*, 1227-1246.