Improving Ensemble Clustering by Unified Diversification and Result Optimization

A Project Report

Submitted in Partial Fulfillment of the Requirements for
the Award of the Degree of

Bachelor of Technology

In

Engineering

Submitted by

Rohini Kumari
Roll No- 2106008

Navneet Kumar
Roll No- 2106045

Bhola Das
Roll No- 2106038

Under the supervision of

Dr. Krishna Kumar Sethi
Assistant Professor
Computer Science and Engineering NIT Patna

Department of Computer Science and Engineering

NATIONAL INSTITUTE OF TECHNOLOGY PATNA
PATNA - 800005

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

CERTIFICATE

This is to certify that **Rohini Kumari (** 2106008 **) , Navneet Kumar (** 2106045), **Bhola Das(** 2106038 **)** has carried out the project entitled " **"** as his 6th semester project under the supervision of Dr Krishna Kumar Sethi, Assistant Professor, Department of Computer Science and Engineering, National Institute of Technology Patna. This project report is a bonafide work done by him for the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

Dr. Krishna Kumar Sethi
(Project Supervisor)
Assistant Professor,
Department of CSE,
NIT Patna

Dr. Maheshwari Prasad Singh
(Head of Department)
Department of CSE,
NIT Patna .

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DECLARATION AND COPYRIGHT TRANSFER

We **Rohini Kumari (2106008) , Navneet Kumar (2106045), Bhola Das(2106038)** , registered candidates **for B.TECH Program** under the department of **Department of Computer Science and Engineering** of **National Institute of Technology Patna**, declare that this is my own original work and does not contain material for which the copyright belongs to a third party and that it has not been presented and will not be presented to any otherUniversity/ Institute for a similar or any other Degree award. We further confirm that for all third-party copyright material in my project report (including any electronic attachments), We have "blanked out" third party material from the copies of the thesis/ dissertation/book/articles etc. fully referenced the deleted materials and where possible, provided links (URL) to electronic sources of the material. We hereby transfer exclusive copyright for this project report to NIT Patna. The following rights are reserved by the author: a) The right to use,free of charge, all or part of this article in future work of their own, such as books and lectures, giving reference to the original place of publication and copyright holding b) The right to reproduce the article or thesis for their own purpose provided the copies are not offered for sale.

Signature of candidates:

Date:

ACKNOWLEDGEMENT

We would like to express my sincere gratitude to Dr Krishna Kumar Sethi ,Assistant Professor, Department of Computer Science and Engineering, National Institute of Technology Patna for providing me an opportunity to do my project work.

This Project has been a good experience for me in a way that it has given me an introduction to the corporate world and a chance to understand the real world outside the classroom.

I. **Rohini Kumari**(Roll No- 2106008)
II. **Navneet Kumar** (Roll No- 2106045)
III. **Bhola Das** (Roll No- 2106038)

ABSTRACT

Ensemble clustering has emerged as a powerful technique in data mining and machine learning, leveraging the collective intelligence of multiple clustering algorithms to enhance clustering accuracy and robustness. However, two critical challenges persist within ensemble clustering frameworks: the need for diversifying base clusters and the efficient storage of intermediate results.

In data mining and machine learning, clustering plays a pivotal role in uncovering hidden patterns and structures within datasets. Ensemble clustering further elevates this by integrating diverse clustering algorithms, leading to more comprehensive and accurate results. However, without effective diversification of base clusters, ensemble methods may suffer from redundancy or bias, limiting their effectiveness. Moreover, the storage of intermediate results poses a significant challenge, as it requires balancing information retention with resource efficiency.

Despite the advancements in ensemble clustering techniques, there remains a gap in addressing both the diversification of base clusters and the storage of intermediate results within a unified framework. Existing approaches often focus on one aspect while neglecting the other, leading to suboptimal performance or scalability issues.

Our research endeavors to bridge this gap by proposing a novel approach that tackles both challenges simultaneously. By developing a comprehensive ensemble clustering algorithm, we aim to diversify base clusters effectively while optimizing the storage of intermediate results with minimal information loss. Through empirical evaluations and comparative studies, we demonstrate the superiority of our method in terms of clustering accuracy, robustness, and scalability. Our findings not only advance the state-of-the- art in ensemble clustering but also provide practical insights for handling large-scale datasets efficiently in real-world applications.
.

TABLE OF CONTENTS

1.INTRODUCTION

Ensemble clustering is a machine learning technique that leverages the collective decision-making of multiple clustering algorithms to enhance the overall clustering performance. Instead of relying on a single clustering algorithm, ensemble clustering combines the outputs of several base clustering algorithms to produce a more robust and accurate clustering solution.

The process of ensemble clustering typically involves three main steps: selection, combination, and fusion. In the selection step, a diverse set of base clustering algorithms is chosen to capture different aspects of the data and provide complementary perspectives. These algorithms can include density-based, partitioning, hierarchical, or any other clustering techniques.

In the combination step, the individual clustering results from each base algorithm are aggregated or combined to create a consensus clustering solution. This can be achieved through techniques such as clustering averaging, clustering consensus function.

The consensus clustering solution is refined or fused to generate the final ensemble clustering result. This may involve resolving conflicts between different base clustering results, refining cluster boundaries, or optimizing clustering quality metrics.

Ensemble clustering offers a powerful framework for improving clustering accuracy, robustness, and generalization by integrating diverse clustering perspectives and leveraging the collective intelligence of multiple algorithms

2.Problem Statement

The problem in ensemble clustering research stems from the tendency to overlook the dual challenges of effectively diversifying base clusters and efficiently managing intermediate results .While some approaches focus on integrating a diverse set of base clustering algorithms to enrich representation, they may neglect considerations regarding the storage of the intermediate results generated by these algorithms.

Conversely, others may emphasize efficient storage practices but overlook the importance of diversifying base clusters to capture the complexity of real-world datasets adequately.

This disjointed approach impedes the development of ensemble clustering algorithms that excel in both aspects simultaneously.

Literacy Survey:

1.From clustering to clustering ensemble selection [1]:-
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.

2.Combining multiple clustering using evidence accumulation [2]:-
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.

3.Clustering ensemble method [3] :-
The Evidence Accumulation (EAC) approach combines multiple clustering results to extract a consistent clustering, treating each partition as independent evidence, voting to create a similarity matrix . The method is evaluated using mutual information and bootstrapping.

4.K-Means-Based Consensus Clustering [4] :-
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.

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

Present Work-1

This project proposes a novel ensemble clustering framework implemented in Python. It leverages coassociation matrices to combine information from various base clusterings generated using different algorithms and parameters. The framework utilizes synthetic datasets and visualizes both the base clusterings and the final ensemble solution for evaluation.

1.Data Generation:

- Utilize Python libraries like scikit-learn to generate synthetic datasets in 2D using functions like make_blobs, make_moons, make_circles, and make_classification. These functions offer flexibility in creating datasets with varying cluster shapes and distributions.

2.Base Clustering:

- Implement different clustering algorithms using different number of K-Means.
- Run these algorithms multiple times with varying values for the number of clusters (k) to generate a diverse set of base cluster solutions.

3.Visualization:

- Employ the matplotlib library to visualize the generated datasets and the corresponding cluster assignments obtained from each base clustering run. This helps in visually inspecting the quality and diversity of the base clusterings.

4.Co-association Matrix Construction:

- For each data point, calculate the probability of it belonging to the same cluster across all base clustering solutions for a specific value of k.
- Based on these probabilities, construct a co-association matrix (CAM). This matrix captures the co-occurrence relationships between data points in the various base clusterings.

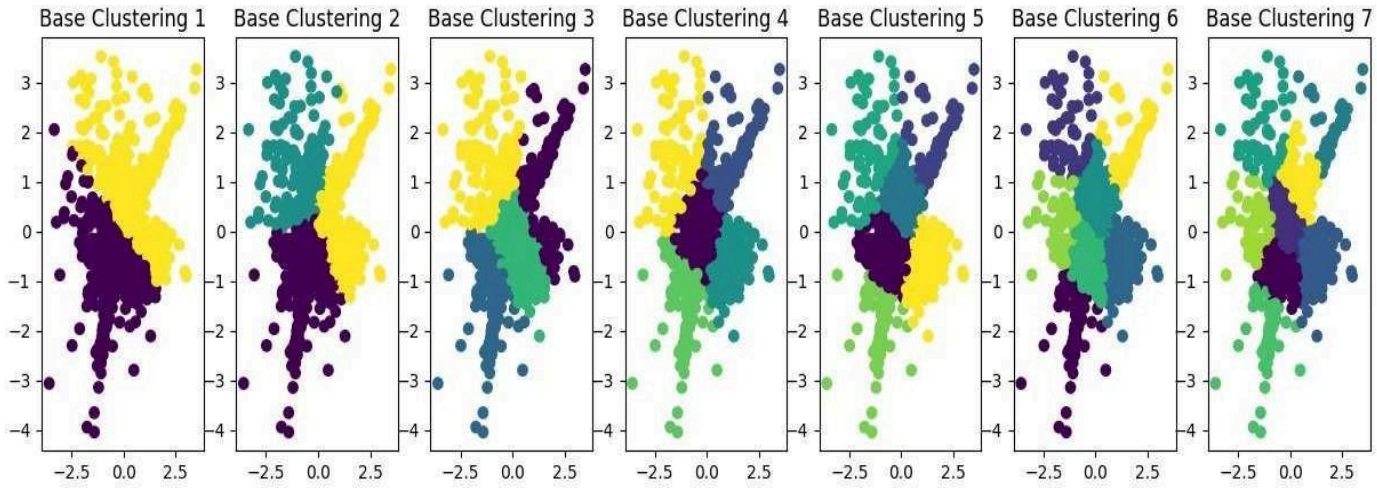
5.Ensemble Clustering:

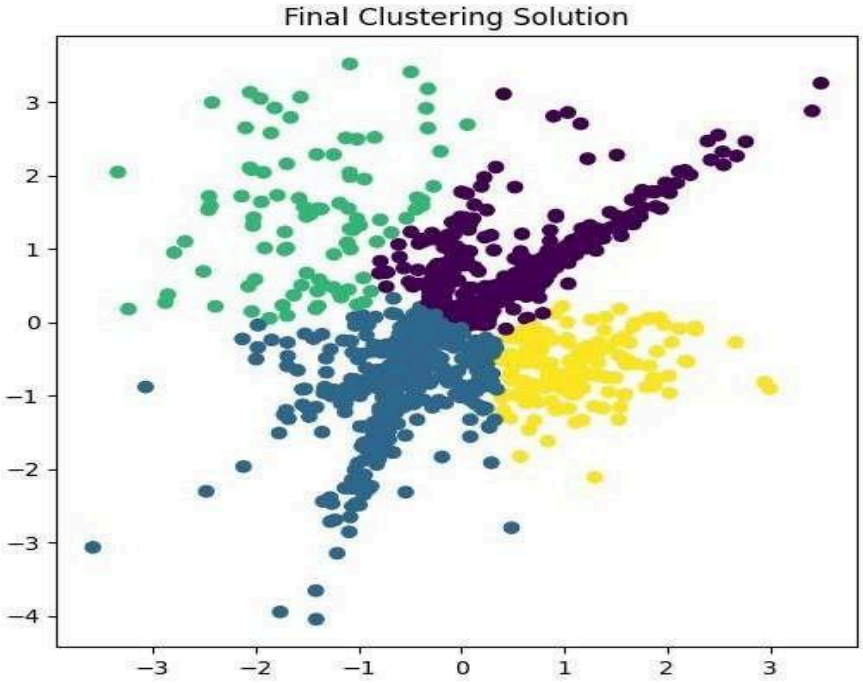
- Apply agglomerative hierarchical clustering to the co-association matrix. This leverages the information from all base clusterings to produce a final ensemble clustering solution.

6.Evaluation:

- Employ appropriate metrics like silhouette score or Calinski-Harabasz index to assess the quality and cohesiveness of the final clustering solution. This helps determine the effectiveness of the ensemble clustering approach

Result of Base Cluster using different values of K:-





Present Work -2

This project proposes a novel ensemble clustering framework. The framework leverages co-association matrices to combine information from diverse base clusterings and achieve a more robust final clustering solution.

1.Data Generation:

- We utilize the scikit-learn library to generate various synthetic datasets in 2D for our experiments. Functions like make_blobs, make_moons, make_circles, and make_classification offer flexibility in creating datasets with different cluster shapes and distributions.

2.Base Clustering:

- We implement different clustering algorithms to generate a diverse set of base cluster solutions. This exploration aims to capture a broader range of potential cluster structures within the data. Here are the specific algorithms employed:

a. K-Means: A popular partitioning clustering algorithm that groups data points into a predefined number of clusters based on their proximity.
b. Hierarchical Clustering: A bottom-up or top-down approach that iteratively merges or splits clusters based on a distance measure.
c. Affinity Propagation (AP): A message-passing algorithm that identifies exemplars (representative data points) for each cluster without requiring a predefined number of clusters.
d. Spectral Clustering: Utilizes techniques from spectral graph theory to identify clusters by analyzing the similarities between data points.

3.Visualization:

- We leverage the matplotlib library to visualize the generated synthetic datasets.

4.Co-association Matrix Construction:

- For each data point, we calculate the probability of it belonging to the same cluster across all base clustering solutions for a specific value of k. This probability reflects the co-occurrence of data points within clusters across the various base clusterings. Based on these probabilities, we construct a co-association matrix (CAM).

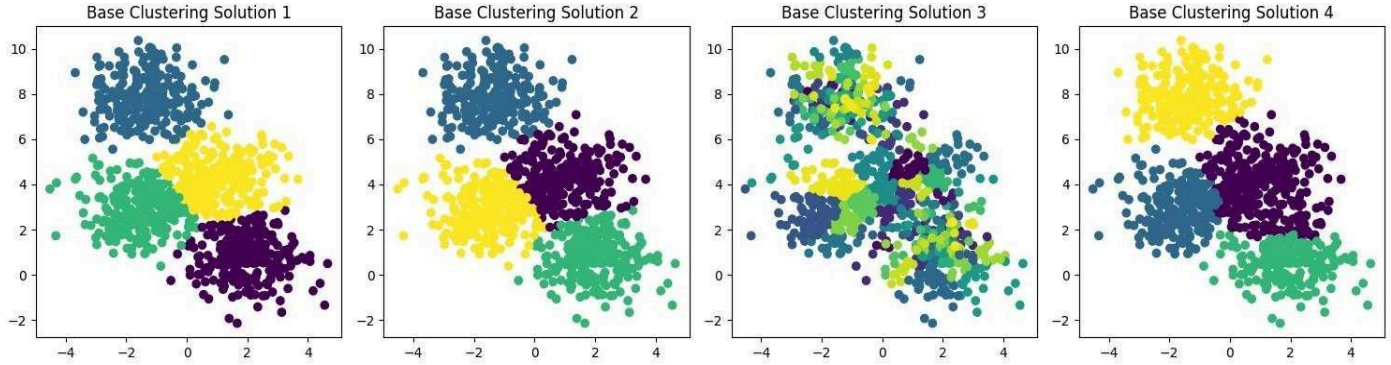
5.Ensemble Clustering:

- We apply agglomerative hierarchical clustering to the co-association matrix. This leverages the information from all base clusterings, encoded within the CAM, to produce a final ensemble clustering solution. The hierarchical clustering process iteratively merges clusters based on a similarity measure, ultimately resulting in a single hierarchy of clusters.

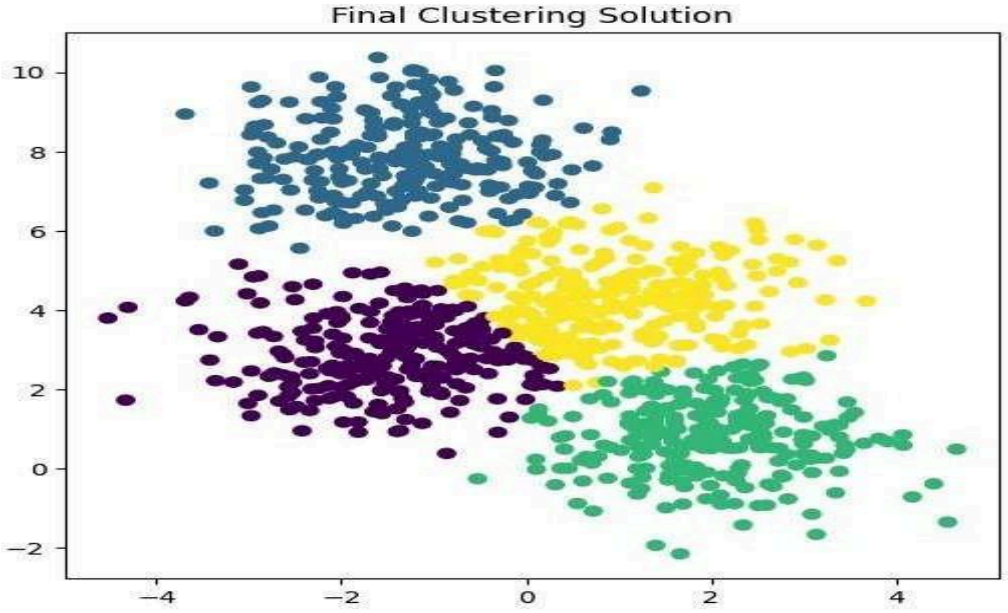
6.Evaluation:

- To assess the quality and cohesiveness of the final clustering solution, we employ appropriate metrics like silhouette score or Calinski-Harabasz index. These metrics evaluate factors such as intra-cluster similarity and inter-cluster separation, providing insights into the effectiveness of the ensemble clustering approach.

Result of Diverse Base Clustering using different



Clustering Algorithm:



Evaluation Metric:

Parameter Used:-

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

  The Silhouette Coefficient for a sample is (b - a) / max(a, b) .
  a=mean intra-cluster distance
  b=mean nearest-cluster distance for each sample

- 2.**Calinski-Harabasz Score:** It measures the ratio of between-cluster dispersion to within-cluster dispersion, indicating the compactness and separability of clusters. The CH score generally favors higher number of clusters

  Calinski-Harabasz score = (n-cn)*sum(diag(B))/((cn-1)*sum(diag(W))) .
  x - data matrix or data frame.
  cn - integer. Number of clusters.
  B being the between-cluster means, and W being the within-clusters covariance matrix.

- **3.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. It is bounded – 0 to 1, lower score is better.

  Davies-Bouldin Score= (Si+Sj)/ Dij

  Similarity R for Cluster i to its nearest cluster j is defined as where Si is the average distance of each point in Cluster i to its centroid. Dij is the distance between the centroid of Cluster i and j.

Result of Evaluation Metric:

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

5.1 Proposed Model

We introduce an innovative approach to enhance ensemble clustering by integrating an attention mechanism inspired by neural networks to compute the co-association matrix. The primary objective is to improve clustering performance by capturing intricate relationships and dependencies between data points. The methodology entails four key steps:

Attention Mechanism for Ensemble Clustering

1. **Defining an Attention Mechanism:** We will develop an attention mechanism designed to assess the importance of one data point relative to another. This mechanism will be capable of capturing complex relationships and dependencies, allowing for a nuanced understanding of the data.

2. **Compute Attention Weight:** For each pair of data points, we will compute the attention weight using the defined attention mechanism. The attention weight will be based on the similarity between the data points or other features that capture their relationship.

3. **Construct the Co-Association Matrix:** We will use the computed attention weights to construct the co- association matrix. Each element in the matrix will represent the strength of association between two data points based on the attention weights.

4. **Incorporate into Clustering Algorithm:** We will incorporate the co-association matrix into a clustering algorithm. This can be done by modifying the objective function of the clustering algorithm to incorporate the co-association matrix, or by using it as a similarity measure in the clustering algorithm.

Expected Outcome

We anticipate that leveraging an attention-based approach to compute the co-association matrix will yield significant improvements in clustering performance. The attention mechanism's ability to discern intricate relationships between data points is expected to lead to more accurate and robust clustering results. By capturing nuanced dependencies, the proposed approach has the potential to enhance the effectiveness of ensemble clustering algorithms across various domains

5.2 Proposed Model

Using Self-Organizing Maps (SOM) to improve the co-association matrix in clustering is an interesting approach.Clustering algorithms often rely on the co-association matrix to capture relationships between Data points However, traditional methods for computing the co-association matrix may not fully capture the underlying structure of the data. In this proposed work, we suggest using Self- Organizing Maps (SOM) to enhance the co-association matrix, aiming to improve clustering performance.

Self-Organizing Maps for Enhancing Co-Association Matrix in Clustering

1. **Train a Self-Organizing Map:** We will train a SOM on the input data to learn the underlying structure and topology of the data. The SOM will organize the data points in a low-dimensional grid, preserving the relationships between them

2. **Compute Co-Association Matrix from SOM:** We will use the trained SOM to compute the co- association matrix. Each neuron in the SOM grid will represent a cluster prototype, and the distance between neurons will be used to compute the co-association matrix.

3. **Refine Co-Association Matrix:** We will refine the co-association matrix using information from the SOM By considering the topological relationships between neurons in the SOM grid, we can improve the accuracy and granularity of the co-association matrix.

4. **Incorporate into Clustering Algorithm:** We will incorporate the refined co-association matrix into clustering algorithm. This can be done by modifying the clustering algorithm's objective function to incorporate the refined co-association matrix or by using it as a similarity measure in the clustering algorithm.

Expected Outcome

We expect that using SOM to enhance the co-association matrix will lead to improved clustering performance. The SOM's ability to capture the data's underlying structure and topology can help create a more accurate and meaningful co-association matrix, leading to better clustering results.

6. CONCLUSION

The traditional clustering paradigm often falls short when it comes to complex, real-world data. It's caught between two extremes: focusing on diverse base cluster algorithms to capture intricate data patterns but neglecting the storage burden of their intermediate results, or prioritizing efficient storage while sacrificing the richness of diverse cluster perspectives. This fragmented approach hinders progress towards truly powerful ensemble clustering methods.

By bridging this gap and integrating diversification strategies with storage optimization within single framework, ensemble clustering can be revolutionized. This holistic approach would ensure comprehensive data representation through diverse base clusters while minimizing resource consumption by employing efficient storage techniques. Ultimately, this would lead to significantly more effective and applicable ensemble clustering across various domains.

By adopting a unified method that optimizes both diversification and storage, ensemble clustering can achieve substantial gains in accuracy, efficiency, and quality assessment. This approach ensures a more comprehensive data representation through diverse base clusters, potentially leading to higher quality final results. Additionally, optimized storage strategies can significantly reduce computational costs and improve scalability, especially with large datasets. Furthermore, a unified framework simplifies quality assessment by allowing for a more holistic analysis of the diversified base clusters and their efficiently stored intermediate data, ultimately leading to a more effective and applicable ensemble clustering methodology.

References

[1] 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

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

[4] 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] 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