

A novel comprehensive investigation for enhancing cluster analysis accuracy through ensemble learning methods

H. N. Lakshmi¹, Thaduri Venkata Ramana², LNC Prakash K³, L. Kiran Kumar Reddy⁴,
Kachapuram Basava Raju⁵

¹Department of Emerging Technologies, CVR College of Engineering, Hyderabad, India

²Department of Computer Science and Engineering (Data Science), CVR College of Engineering, Hyderabad, India

³Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), CVR College of Engineering, Hyderabad, India

⁴Department of Computer Science and Engineering (Data Science), Geethanjali College of Engineering and Technology Cheeryal, Hyderabad, India

⁵Department of Artificial Intelligence, Anurag University, Hyderabad, India

Article Info

Article history:

Received Jan 1, 2024

Revised Jul 9, 2024

Accepted Jul 17, 2024

Keywords:

Accuracy

Clustering

Ensemble methods

Machine learning

Patterns

ABSTRACT

Ensemble learning stands out as a widely embraced technique in machine learning. This research explores the application of ensemble learning, including ensemble clustering, to enhance the precision of cluster analysis for datasets with multiple attributes and unclear correlations. Employing a majority voting-based ensemble clustering approach, specific techniques such as k-means clustering, affinity propagation, mean shift, BIRCH clustering, and others are applied to defined datasets, leading to improved clustering results. The study involves a comprehensive comparative analysis, contrasting ensemble clustering outcomes with those of individual techniques. The process of improving cluster identification accuracy encompasses data collection, pre-processing to exclude irrelevant elements, and the application of standard clustering algorithms. The task includes defining the optimal number of groups before comparing clustering models. Additionally, a combined model is constructed by merging BIRCH clustering and mean shift clustering, leveraging their advantages to enhance overall clustering strength and accuracy. This research contributes to advancing ensemble learning and ensemble clustering methodologies, offering improved accuracy, and uncovering hidden patterns in complex datasets.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Thaduri Venkata Ramana

Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), CVR College of Engineering

Hyderabad, India

Email: meetramana1204@gmail.com

1. INTRODUCTION

Clustering arranges elements in a collection so that some are similar and others are different. Although its meaning is unclear, clustering is used in many fields of science and life. The abundance of unlabeled data compared to labelled data makes clustering crucial today. The rapid increase and accessibility of data sources and processing advances highlight the need to generate usable intuitions from data. This supports appropriate analysis based on realization methods. Machine learning (ML) investigations are usually supervised or unsupervised. Supervised learning creates a function from an input to an output based on input-output pairings. Thus, these acquisition methods require tagged data with the desired yield [1]. Tagged data suggests a conventional result, but it is expensive and hard to get. In intrusion detection, zero-day attacks

are rare and expensive to label. Thus, unsupervised learning is used when dataset labels are unavailable [2]. In a realizing setting, the procedure seeks inferences from the dataset without knowing its labels. Semi-supervised learning is prevalent and includes approaches between supervised and unsupervised learning [3].

Exploring patterns and correlations within data is a crucial aspect of current research. Cluster analysis is a highly effective data mining technique used for this purpose. Similarities or differences between data points in a dataset can be used to detect important groupings or clusters, making this unsupervised learning technique highly valuable. Cluster analysis reveals potential hidden patterns, structures, or connections in the data through the combination of similar data points. This versatile approach is utilized in a wide range of industries, such as market research, photo identification, anomaly detection, and customer segmentation. The goal of clustering is to efficiently divide the data into clusters based solely on internal characteristics such as density or proximity. Various clustering strategies involve algorithms like k-means, hierarchical clustering, and density-based spatial clustering of applications with noise (DBSCAN). By disregarding external information and combining different clustering methods, the objective is to pinpoint a singular data partition that accurately represents the inherent structure of the data. When it comes to individual clustering, there are certain factors that can affect its accuracy, such as noise, initialization, and the selection of clustering parameters. In these situations, the performance of traditional clustering methods may be compromised, leading to less-than-optimal results. Ensemble clustering has garnered significant interest among data mining and machine learning researchers as a solution to the challenge of locating actual clusters. Ensemble clustering has emerged as a powerful tool that combines multiple clustering results to create a more reliable and effective consensus clustering. This is achieved by merging results from different algorithms or the same algorithm with different parameter settings. Unlike the typical approach that uses a single algorithm to generate one clustering result, there has been a growing interest in ensemble clustering, which involves employing multiple clustering approaches to produce more accurate results [4].

This research is motivated by the need to enhance the precision and reliability of cluster analysis in many applications like weather forecasting. Correct weather estimates are crucial for disaster awareness, resource sharing, and long-term weather planning. Traditional clustering algorithms often struggle with the complexity of relationships and patterns in weather forecasting datasets. To address this, this research proposes leveraging ensemble learning techniques to improve clustering accuracy and elevate the quality of weather predictions. Weather forecasting involves analyzing various meteorological attributes like temperature, humidity, wind speed, and precipitation to predict future conditions. Traditional clustering algorithms [5], face challenges in capturing intricate, nonlinear relationships between these attributes. Applying ensemble learning to weather forecasting datasets aims to overcome these limitations by combining multiple clustering models, producing more accurate and robust results. This approach mitigates the impact of noisy or incomplete data, improves cluster separability, and improves inclusive clustering performance. Ensemble learning, a prominent machine learning approach, enhances predictive accuracy and generalization by amalgamating the insights of multiple models. The fundamental concept involves aggregating predictions from diverse models, harnessing their collective strength to yield outcomes that surpass those of individual models in terms of accuracy and robustness. Demonstrating efficacy across diverse domains such as classification, regression, and anomaly detection, ensemble learning has proven its versatility.

The remainder of the paper is organized as follows. In section 2, relevant literature is exhibited. Section 3 specifies an outline of the recommended clustering ensemble approach which outlining its distinct phases. Moving on to section 4, we elaborate on the empirical investigations and assess performance using authentic datasets. Concluding the presentation in section 5, we ultimately consider conclusions and possible improvements for the future.

2. LITERATURE REVIEW

The current literature on cluster analysis can be broadly characterizes two major streams: traditional clustering approaches and commonly employed techniques. Traditional approaches, exemplified by k-means, hierarchical grouping, and DBSCAN, have played a focal role in the field and find extensive application across diverse domains. Their attractiveness stems from their simplicity and practicality, making them preferred selections for clustering tasks. Noteworthy within these traditional methods are specific techniques celebrated for their unique advantages, k-means for its efficiency and ease of implementation, hierarchical clustering for its hierarchical gathering illustration, and DBSCAN for its expertise in density-based clustering. While these methods have provided substantially to address a diverse array of clustering challenges, they are not without limitations. Therefore, it becomes imperative to study different clustering strategies that can surmount these shortcomings and augment the accuracy and adaptability of cluster analysis. A clustering ensemble tries to merge several clumping models to generate a superior outcome compared to distinctive clustering methods in terms of both uniformity and superiority.

An advanced ensemble clustering methodology that depends on the quick distribution of cluster-wise comparisons through casual walks. At first, a graph is generated to represent the similarities between clusters. The clusters act as nodes, and the Jaccard coefficient is used to determine the weights of the edges connecting them. With the help of a structured graph, a passage probability matrix is created to support the random walk activity. This matrix allows for the spread of structural information throughout the graph. Through an analysis of the paths originating from different points, a new matrix for comparing clusters is created by considering the relationships between these paths. Afterwards, the recently obtained cluster-wise similarity matrix is converted from the cluster-level to the object-level, resulting in an improved co-association matrix. This matrix effectively captures both the associations between objects and the relationships between clusters at multiple scales. An analysis is conducted using visual clustering methods to determine the number of groups. The cluster quantity is visually represented by these techniques as dark blocks with square shapes running diagonally. The representation of clusters is depicted by a series of black, square blocks, and the techniques work exceptionally well on smaller datasets. However, the accuracy of grouping has decreased on large databases, and the processing size has increased [6].

By analyzing various clustering algorithms, this research aimed to assist vendors in identifying and prioritizing the most profitable market segments, while disregarding the less lucrative ones. The goal was to determine the most precise method for clustering customer behavior. This type of study is crucial for business growth as it helps in retaining customers and increasing company profits. Trades categorize their customers based on similar behavioral characteristics. This method ensures maximum exposure of online offers to capture the attention of potential customers. Two different learning methods, k-means and hierarchical clustering, are used on a customer dataset to determine which approach yields the most accurate clustering results. Grouping plays a crucial role in numerous data-driven applications and is considered a fascinating and significant task in machine learning. It is also studied in statistics, pattern differentiation, computational geometry, bio-informatics, optimization, and in a wide range of other disciplines [7], [8]. Clustering techniques such as possibilistic fuzzy C-mean achieve a fuzzy subdivision of data points using probability. These techniques find applications in various areas, including image subdivision and others [9] [10]. Many researchers have delved into the clustering processes for large datasets. An in-depth analysis of large datasets is conducted, and the algorithms for classification and clustering that are based on MapReduce are discussed and examined [11]. Analyzing the attributes of different clustering algorithms and addressing the challenges of handling large datasets, conducting a comprehensive analysis of the main clustering procedures [12]. Examining the concept, possibilities, and challenges of large datasets and providing a brief overview of the cutting-edge techniques used to analyze this information [13].

The study [14] categorizes the present nonparallel cluster formation techniques and conducts experiments to examine the correctness, scalability, and efficiency of various algorithm types in handling huge datasets. The purpose of ensemble clustering is to integrate the data from a diversified range of components in order to create a presumably enhanced partition. The fundamental approach comprises constructing a cluster of algorithms $A = \{A_{(1)}, A_{(2)}, \dots, A_{(M)}\}$, designated as ensemble members, and consolidating the resultant partitions $P = \{P_{(1)}, P_{(2)}, \dots, P_{(M)}\}$, through the application of a consensus function [15], [16]. It is crucial to stress that consensus clustering is not exclusively applied for getting a partition; it can also serve goals such as analyzing the number of groups in the information [17], examining their consistency, or conducting a hyperparameter investigation [18]. An alternate name synonymous with ensemble clustering is "consensus clustering," as reported [19], and provided a marginally different approach. They advised employing resampling without replacement in their investigation. This concept was based on the premise that clusters that demonstrated better resistance to resampling might be actual clusters. They also proposed a principle for identifying the correct number of groups. Müller *et al.* [20] provide extensive assessments of several clustering techniques. They describe the various qualities of each algorithm and conduct a complete study of their belongings. The outcomes are assessed using both real-world and simulated samples. Using numerous clustering evaluation criteria, they contrasted these findings with classic clustering algorithms across different areas [21], [22].

The innovative method called density conscious subspace clustering (DENCOS) This method identifies clusters by examining the relative density in a subspace and represents them as regions. Our approach at DENCOS involves using a split-and-conquer strategy to effectively identify clusters with different density thresholds and subspace cardinals. Based on extensive research and analysis, it has been proven that DENCOS is highly accurate in identifying groups in various subspaces. In fact, it outperforms other clustering techniques in terms of precision. Researchers have chosen specific approaches for ensemble clustering based on the intended applications, particularly when dealing with high-dimensional data. A feature subset [23] was generated by selecting various sets of objects, and this approach was called "object distribution". The process of projecting things into different subspaces is carried out in a seemingly random manner, lacking any discernible pattern or rationale [24].

3. PROPOSED METHOD

Ensemble clustering merges a group of clustering methods using the same dataset, resulting in a final clustering. The objective of ensemble clustering is to enhance the quality of individual data clustering. In this study, majority voting-based ensemble clustering approach is utilized; to achieve this, we employ specific clustering techniques on defined datasets to obtain individual outcomes. Following the implementation, the ensemble clustering outcomes are contrasted with those of individual techniques. The findings from ensemble clustering indicate improved clustering results and Figure 1, provides a detailed summary of a research study aimed at improving the accuracy of cluster identification. Data collection is the first step, and then there are many stages of data preparation and organization before clustering. In the pre-processing phase, irrelevant elements such as null entries, unidentified values, and negligible attributes, which exert little or no noticeable influence on the results, are excluded, and anomalies are filtered out to construct a proficient clustering model.

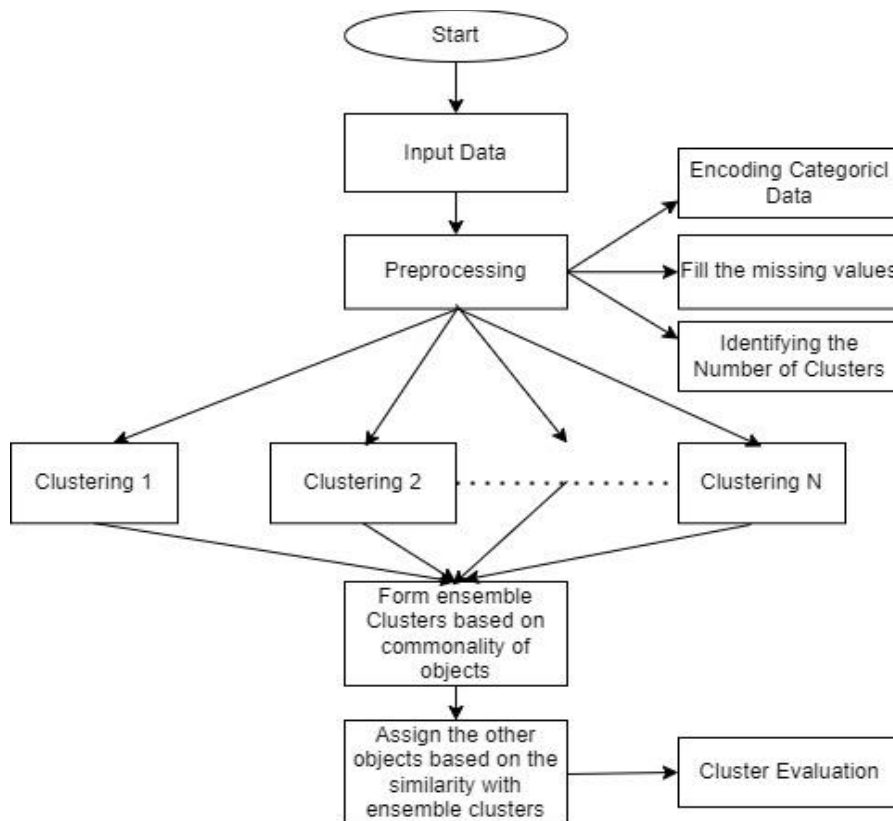


Figure 1. Architecture of the proposed methodology

Finding the ideal number of groups needed to split the dataset into meaningful groups is the first step in using different clustering algorithms to train the dataset [25]. The dataset is subjected to common clustering methods such as k-means clustering, affinity propagation, mean shift, BIRCH clustering, and others before the clustering models are compared. Basic clustering results are produced using these techniques, and the suggested strategy is then contrasted with them. To create an embedded model for this study, BIRCH clustering and mean shift clustering, the two distinct clustering techniques are merged to create a combined model. This combination strategy improves the overall strength and accuracy of clustering by using the benefits of each grouping technique.

3.1. Mean shift clustering

A non-parametric grouping approach named mean shift does not necessitate a grasp on the number of clusters beforehand. It operates by repeatedly moving data points in the direction of the supporting probability distribution's mode, or peak. The following is a detailed process for mean shift clustering:

- Input: consider the data points $X = \{x_1, x_2 \dots x_n\}$, kernel function K with bandwidth h and convergence threshold ϵ .

- Initialization: Initialize cluster centers $C = X$ and shift vector $Shift = Zeros$.
- Mean shift iterations: repeat until convergence:

for (each $x_i \in X$):

Compute the mean shift using

$$Shift(x_i) = \frac{\sum_{j=1}^n K\left(\frac{x_i - x_j}{h}\right) * x_j}{\sum_{j=1}^n K\left(\frac{x_i - x_j}{h}\right)}. \quad (1)$$

Update cluster centers

$$C = C + Shift \quad (2)$$

- Convergence check: If $(\|Shift\| < \epsilon)$, end the for loop.
- Assign each data point x_i to the group whose center it converged.
- Return the clusters C .

The kernel function K is typically a Gaussian kernel, but other choices can be used depending on the characteristics of the data. The bandwidth h controls the size of the neighborhood used to estimate the local density and influences the shape and number of clusters. The convergence threshold ϵ determines when the mean shift iterations should stop. The shift vector represents the direction and magnitude of the mean shift for each data point. The algorithm alliteratively shifts data points towards the mode of the underlying density until convergence.

3.2. BIRCH clustering

The BIRCH clustering technique is a procedure applied to arrange and group data points into clusters. It handles large datasets effectively by arranging data into hierarchical form. The process involves the following steps:

- Initialization: Begin by initializing the tree structure and setting parameters such as the branching factor and the threshold for the number of points in a sub cluster.
- Insertion: Add data points to the tree structure, creating sub clusters at the leaf nodes. If a sub cluster exceeds the specified threshold, it is split to maintain the hierarchical structure.
- Clustering: As data points are inserted, the tree structure is continuously updated, and clusters are formed at different levels of the hierarchy.
- Merging: Sub clusters with similar characteristics are merged to create a more compact representation of the data.
- Output: The resulting hierarchy of clusters represents the organized structure of the data, with each leaf node corresponding to a sub cluster.

The BIRCH clustering procedure is particularly useful for applications where efficiency and scalability are essential, creating it appropriate for big and high-dimensional databases.

3.3. Embedded approach

In this strategy, genuine groupings are established by incorporating the BIRCH and mean shift methodologies. Initially, the association between each pair of entities in the database is verified to determine if they belong to the identical cluster across all methodologies. If they consistently belong to the same cluster in every technique, they are grouped into that cluster; otherwise, an assessment for an alternative cluster is conducted at a subsequent stage. Following the allocation of data objects to a predefined number of clusters, an examination is carried out for any remaining entities in the database. For any objects in the database not assigned to any clusters, they are allocated to one of the existing clusters based on the resemblance of the tuples to the centroids of those groups. The outlined approach is depicted in Algorithm 1.

According to the algorithm Let D be the dataset,
 $M = \{M_i \setminus M_i \text{ is a clustering method}\}$, which is the set of clustering methods,
 C be the set of all groups that is $C = \{C_{ij} \setminus C_{ij} \text{ is } j^{\text{th}} \text{ cluster in } i^{\text{th}} \text{ method}\}$, and
 $K = \{K_i \setminus i = 1, 2, \dots, \text{number of clusters}\}$, be the set of resulting clusters.

In this research, k-means clustering, mean shift clustering, agglomerative clustering, spectral, OPTICS clustering, BIRCH clustering are employed and obtained results individually on the described

datasets. For the ensemble clustering, mean shift clustering and BIRCH clustering are utilized to enhance the clustering accuracy.

Algorithm 1.

Input: Dataset, resulting clusters of all cluster methodologies.

Output: Real clusters after embedded method.

Consider,

$M = \{M_i \mid M_i \text{ is a clustering method}\}$

$C = \{C_{ij} \mid C_{ij} \text{ is } j^{\text{th}} \text{ cluster in } i^{\text{th}} \text{ method}\},$

$K = \{K_l \mid l = 1, 2, \dots \text{number of clusters}\},$

1. Initialize $l = 1, \text{count} = 0.$ // Initialization of variables.
2. *while* ($l \neq |K|$)
3. $K_l = \emptyset.$
4. *for* (each $r_x \in D$)
5. *for* (each $r_y \in D, x \neq y$) // Considering each pair of objects
6. *for* ($i = 1$ to $|M|$)
7. *for* ($j = 1$ to $|C|$) // For each cluster from each method
8. *if* ($\{r_x, r_y\} \in C_{ij}$) // checking for each pair of the objects belongs to same cluster
9. $\text{Count} = \text{count} + 1$
10. *End of sep 8.*
11. *End of sep 7.*
12. *End of sep 6.*
13. *if* ($\text{count} = |M|$)
14. $K_l = K_l \cup \{r_x, r_y\}.$ // Assigning couple of tuples to same cluster if they belong to same cluster in all methods.
15. *End of sep 13.*
16. *End of sep 5.*
17. *End of sep 4.*
18. $\text{count} = 0.$
19. $l = l + 1.$
20. $D = D - K_l.$
21. *End of sep 2.*
22. *if* ($D \neq \emptyset$)
23. Find the centroid of each cluster $K_l.$
24. *for* (each $r_x \in D$ and K_l) // Assigning the remaining objects to corresponding clusters
25. $K_l = \{K_l \cup r_x \mid \text{distance from } r_x \text{ to the centroid of } K_l \text{ is minimum}\}.$
26. *End of sep 24.*
27. *End of sep 22.*
28. Return the ensemble set of clusters,
29. $K = \{K_l \mid l = 1, 2, \dots |C|\}.$

4. EXPERIMENTAL ANALYSIS AND PERFORMANCE EVALUATION

This segment elaborates on the experimental examination, chosen datasets, and resultant findings derived from the assessment. The comprehensive experiments were conducted on the weather history dataset and weather prediction dataset. In this research the results carried out using a machine with an Intel Core i3 CPU and 4 GB of RAM using the Python programming language. In python the existing clustering method mean shift clustering is implemented using the MeanShift class and the method BIRCH clustering is implemented using the BIRCH class from scikit-learn. The proposed method of ensemble clustering is implemented according to Algorithm 1.

4.1. Data set description

The historical weather archive furnishes past weather information pertaining to different locations. It encompasses comprehensive details regarding weather circumstances documented throughout a particular time frame. Comprising a grand total of 96,453 entries, each signifies a distinct timestamp alongside correlated weather metrics. After analyzing visualizations and extracting insights, the most suitable number of clusters for this dataset has been ascertained to be four. This implies that, in accordance with the specified criteria, segmenting the data into four clusters yields the most significant and indicative categorization of data points within the provided dataset and problem domain. The weather prediction dataset comprises meteorological records sourced from 18 distinct European locations. Spanning the years 2000 to 2010, the dataset encompasses 3,654 daily observations. It involves diverse variables such as average temperature, maximum temperature, minimum temperature and so on. The optimal cluster count for the dataset has been determined to be 2. This implies that, in accordance with the chosen criteria, partitioning the data into two clusters yields the most meaningful and representative categorization of data points within the specified dataset and problem domain.

4.2. Study on performance

In this section, we examine the results of the experiment and evaluate the advantages of the proposed model in comparison with other current models selected for the studies. Two components of the extensive trials were carried out. Initially, an assessment was conducted on the accuracy of clustering using well-known techniques including k-means, Affinity propagation, mean shift, and BIRCH clustering. To extract useful information from the database, this evaluation required using an appropriate distance metric. The clustering procedure was used in the second experiment to emphasize the significance of the suggested strategy. This required using an ensemble approach that uses BIRCH clustering and mean shift. Metrics like the Silhouette score and cluster Davis Bouldin score were used to assess the relevance of the clusters that were produced by the suggested ensemble approach versus traditional clustering methods. The weather history dataset is shown in Figure 2 using the elbow method and Calinski-Harabasz score. The data was segmented into four clusters, which produced the most significant and suggestive grouping of data points within the given dataset and issue area. In a similar vein, Figure 3 illustrates the weather prediction dataset using the elbow approach and silhouette score. Dividing the data into two clusters produces the most representative and significant grouping of data points.

Grounded on the conclusions existing in Table 1, it is apparent that the collective clustering approach has produced a reduced Davies-Bouldin score [26], and an elevated Silhouette score [27], when contrasted with all alternative conventional clustering algorithms for weather history dataset. These outcomes indicate that the collective model applied to the weather history dataset demonstrates superior clustering efficacy concerning both the distinctiveness and coherence of clusters, as assessed through these metrics.

Based on the findings presented in Table 2, the ensemble clustering approach has produced a significantly reduced Davies-Bouldin score when compared to various conventional clustering algorithms for weather prediction dataset. These outcomes indicate that, when assessed with these metrics, the ensemble model applied to the weather prediction dataset demonstrates superior performance in terms of both the separation and cohesion of clusters.

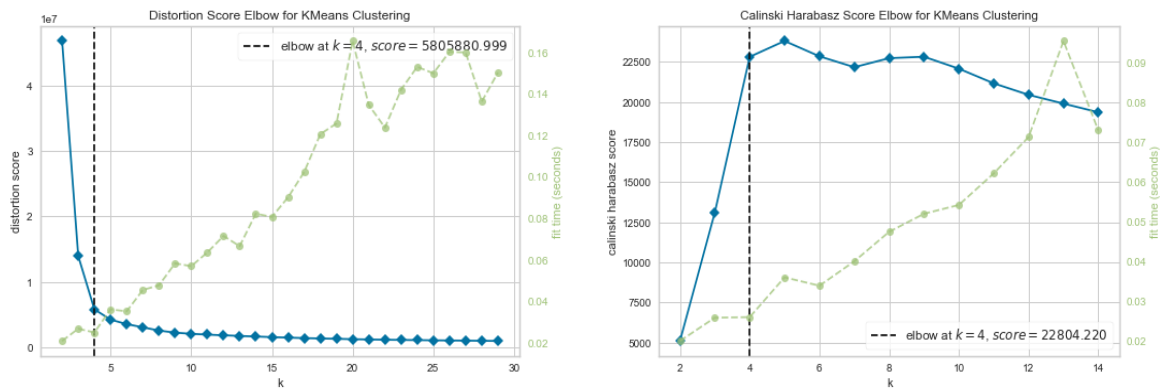


Figure 2. Optimal number of clusters identified by elbow method and Calinski-Harabasz score

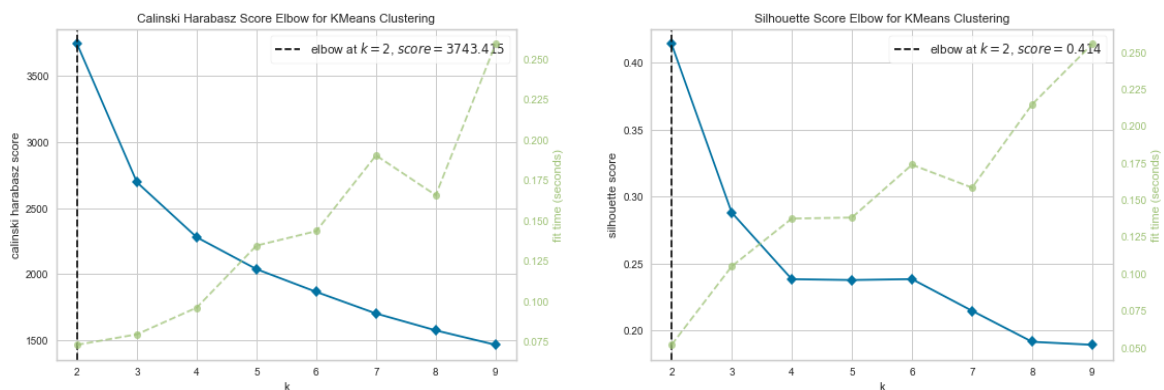


Figure 3. Optimal number of clusters identified by elbow method and silhouette score

Table 1. Comparing ensemble clustering to traditional models on weather history data

Algorithm	Number of clusters	Davis Bouldin score	Silhouette score
KMeans clustering	4	0.401	0.608
Mean shift clustering	4	0.435	0.867
Agglomerative clustering	4	0.405	0.588
Spectral clustering	4	0.401	0.605
OPTICS clustering	4	1.870	-0.561
BIRCH clustering	4	0.405	0.028
Ensemble clustering	4	0.184	0.873

Table 2. Comparing ensemble clustering to traditional models on weather prediction dataset

Clustering algorithm	Number of clusters	Davis Bouldin score	Silhouette score
KMeans	2	0.937	0.414
Mean shift	3	0.955	-0.002
Agglomerative	2	1.021	0.354
Spectral	2	0.939	0.412
OPTICS	2	1.300	-0.255
BIRCH	2	0.970	0.378
Ensemble	3	0.683	0.427

The findings indicate that the ensemble model is formed by combining the results of two well-established clustering models, namely mean shift and BIRCH, utilizing a voting technique to obtain the ultimate outcomes. These two models are favored for proficiency in recognizing dense regions within the data and perform exceptionally well in scenarios characterized by clusters with diverse shapes and sizes, rendering it a fitting option for capturing complex structures present in the data by the mean shift method and BIRCH which provides scalability and computational effectiveness. It adeptly manages sizable datasets and exhibits lower computational intensity in comparison to certain alternative algorithms. This renders a pragmatic option for scenarios where there is a consideration for computational resources.

We provide a model of an ensemble clustering algorithm to solve the issue that the conventional clustering algorithms have trouble handling the set-based data along with nonlinear data based on the Davis Bouldin score and the Silhouette score. Initially, we establish the categories based on how comparable the data is. Determining the ideal number of groups needed to split the dataset into meaningful groups is the problem at hand [28]. The dataset is subjected to common clustering methods such as affinity propagation, k-means clustering, BIRCH clustering, mean shift, and others before the clustering models are compared. Basic clustering results are produced using these techniques, and the suggested strategy is then contrasted with them. To create an embedded model for this study, BIRCH clustering and mean shift clustering, the two distinct clustering techniques are merged to create a combined model, as shown. This combination strategy improves the overall strength and accuracy of clustering by using the benefits of each grouping technique. In contrast to the mean shift technique, which is based on the absolute and relative distances of the granular vectors [29]. In conclusion, our suggested ensemble clustering method outperforms conventional clustering algorithms, as shown by the trials, which show an average enhancement of 6.8% and 4.2% in the modeling results, respectively.

4.3. Discussion

In this study, ensemble clustering performs better at clustering on numerous data sets than MeanShift along with additional clustering methods. Mean shift performs somewhat worse in the weather dataset than the British mixture along with agglomerative clustering methods, although the difference is not very great. In contrast to the conventional clustering algorithm, an algorithm looks for structural innovations using neighborhood granulation technology. This enhances the algorithm's clustering performance and yields superior results for a variety of datasets. Overall, it is discovered that from Tables 1 and 2, ensemble clustering performs better and is more adaptable. The measurements used for accuracy comparison are the Silhouette score and the Davis Bouldin score, both of which have shown promising outcomes when combined with ensemble clustering. This can be further used to improve the accuracy of the research [30], [31].

5. CONCLUSION

This study presents an ensemble model that is intended to handle the complexity and size of modern datasets, to solve the issue that classic clustering algorithms struggle to handle set-based data and nonlinear data. To create this model, the results of the mean shift and BIRCH algorithms were carefully combined. A voting approach was used to ensure that each algorithm's advantages were effectively included.

A comprehensive analysis was carried out by contrasting this ensemble model with other well-known conventional clustering models. The results demonstrate how well the planned ensemble model performs in comparison to its conventional alternatives. As a result, the ensemble approach's efficacy in improving clustering accuracy and resilience which makes use of the mean shift and BIRCH clustering algorithms underlines its significance. These revelations further aid in decision-making, make data exploration easier, and promote a deeper examination of the basic patterns in the dataset that is supplied. The goal of this feature is to create a system that is adaptable and flexible enough to be used with ease across various datasets and domains, ensuring its relevance and functioning in a range of data analysis scenarios. In the future, we want to add more complex automation capabilities to make this study compatible with a wider range of data types, increase its scalability to handle bigger datasets, and make it easier for users to adopt by adding these features. In future, to enhance the clustering algorithm's performance, we want to create more sophisticated granular vector distance metrics as distance measures. Applying the suggested granular vectors absolute and relative metrics to different clustering methods is also an intriguing piece of work.

ACKNOWLEDGEMENTS

We express my sincere gratitude to my mentors at CVR College of Engineering, Hyderabad, Department of DS at Geethanjali college of Engineering and Technology, Hyderabad, and the Department of AI at Anurag University for their unwavering support and valuable insights. Their collaborative efforts have significantly shaped the research direction, enhancing the quality and depth of this study.




REFERENCES

- [1] A. T. Mehryar Mohri, Afshin Rostamizadeh, *Foundation of machine learning*. Cambridge, MA, USA, 2012.
- [2] C. M. Bishop, *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press, 1995.
- [3] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwok, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 508–520, Jan. 2021, doi: 10.1109/tsmc.2018.2876202.
- [4] K. R. Prasad, V. Sireesha, M. Mohammed, and K. Jeevitha, "An efficient pre-clusters assessment technique for efficient data partitions," in *Proceedings of the 2nd International Conference on Edge Computing and Applications, ICECAA 2023*, Jul. 2023, pp. 605–610, doi: 10.1109/ICECAA58104.2023.10212335.
- [5] M. Ahmed, N. Choudhury, and S. Uddin, "Anomaly detection on big data in financial markets," *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, Jul. 2017, doi: 10.1145/3110025.3119402.
- [6] M. Ahmed, "Collective anomaly detection techniques for network traffic analysis," *Annals of Data Science*, vol. 5, no. 4, pp. 497–512, Jan. 2018, doi: 10.1007/s40745-018-0149-0.
- [7] C. L. Chowdhary, M. Mittal, K. P., P. A. Pattanaik, and Z. Marszalek, "An efficient segmentation and classification system in medical images using intuitionist possibilistic fuzzy C-Mean clustering and fuzzy SVM algorithm," *Sensors*, vol. 20, no. 14, Jul. 2020, doi: 10.3390/s20143903.
- [8] C. L. Chowdhary and D. P. Acharjya, "Clustering algorithm in possibilistic exponential fuzzy c-mean segmenting medical images," *Journal of Biomimetics, Biomaterials and Biomedical Engineering*, vol. 30, pp. 12–23, Jan. 2017, doi: 10.4028/www.scientific.net/jbbbe.30.12.
- [9] C. L. Chowdhary and D. P. Acharjya, "Segmentation of mammograms using a novel intuitionistic possibilistic fuzzy c-mean clustering algorithm," in *Advances in Intelligent Systems and Computing*, vol. 652, Springer Singapore, 2018, pp. 75–82.
- [10] P. G. Shynu, H. Md Shayan, and C. L. Chowdhary, "A fuzzy based data perturbation technique for privacy preserved data mining," *International Conference on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, Feb. 2020, doi: 10.1109/ic-ETITE47903.2020.244.
- [11] B. Zerhari, A. A. Lahcen, and S. Mouline, "Big data clustering: algorithms and challenges," in *Proceedings of International Conference on Big Data, Cloud and Applications (BDCA'15)*, Tetuan, Morocco, May 2015.
- [12] A. Ben Ayed, M. Ben Halima, and A. M. Alimi, "Survey on clustering methods: towards fuzzy clustering for big data," *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Aug. 2014, doi: 10.1109/socpar.2014.7008028.
- [13] C. L. Philip Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: a survey on Big Data," *Information Sciences*, vol. 275, pp. 314–347, Aug. 2014, doi: 10.1016/j.ins.2014.01.015.
- [14] A. Fahad *et al.*, "A survey of clustering algorithms for big data: taxonomy and empirical analysis," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267–279, Sep. 2014, doi: 10.1109/TETC.2014.2330519.
- [15] S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 3, pp. 337–372, May 2011, doi: 10.1142/S0218001411008683.
- [16] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim, "A comparative study of clustering ensemble algorithms," *Computers & Electrical Engineering*, vol. 68, pp. 603–615, May 2018, doi: 10.1016/j.compeleceng.2018.05.005.
- [17] R. Ünü and P. Xanthopoulos, "Estimating the number of clusters in a dataset via consensus clustering," *Expert Systems with Applications*, vol. 125, pp. 33–39, Jul. 2019, doi: 10.1016/j.eswa.2019.01.074.
- [18] L. Helfmann, J. von Lindheim, M. Mollenhauer, and R. Banisch, "On hyperparameter search in cluster ensembles," *arXiv preprint arXiv:1803.11008*, 2018.
- [19] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1–2, pp. 91–118, Jul. 2003, doi: 10.1023/A:1023949509487/METRICS.




- [20] E. Müller, S. Günemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 1270–1281, Aug. 2009, doi: 10.14778/1687627.1687770.
- [21] Y.-H. Chu, J.-W. Huang, K.-T. Chuang, D.-N. Yang, and M.-S. Chen, "Density conscious subspace clustering for high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 1, pp. 16–30, Jan. 2010, doi: 10.1109/tkde.2008.224.
- [22] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 583–617, 2003, doi: 10.1162/153244303321897735.
- [23] Fern, X. Zhang, Brodley, and C. E., "Random projection for high dimensional data clustering," in *Proceedings of the 20th international conference on machine learning*, 2003, pp. 186–193.
- [24] S. Swift *et al.*, "Consensus clustering and functional interpretation of gene-expression data," *Genome biology*, vol. 5, no. 11, 2004, doi: 10.1186/GB-2004-5-11-R94.
- [25] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, Jun. 2005, doi: 10.1109/tpami.2005.113.
- [26] L. N. C. P. K, G. Surya Narayana, M. D. Ansari, and V. K. Gunjan, "Instantaneous approach for evaluating the initial centers in the agricultural databases using k-means clustering algorithm," *Journal of Mobile Multimedia*, Aug. 2021, doi: 10.13052/jmm1550-4646.1813.
- [27] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/tpami.1979.4766909.
- [28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
- [29] B. Rambabu, B. Vikranth, S. Anupkanth, B. Samya, and N. Satyanarayana, "Spread spectrum based QoS aware energy efficient clustering algorithm for wireless sensor networks," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 1, pp. 154–160, Feb. 2023, doi: 10.17762/ijritcc.v11i1.6085.
- [30] V. J. Arputharaj, K. Sankar, A. S. Kumar, M. Sridevi, and D. D. Prasad, "An IoT-Based computational intelligence model to perform gene analytics in paternity testing and comparison for health 4.0," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 14, pp. 5781–5796, Jul. 2023.
- [31] L. K. K. Nekkanti and V. Rao, "A Yolo-based deep learning approach for vehicle class classification," in *Multi-disciplinary Trends in Artificial Intelligence*. Springer Nature Switzerland, 2023, pp. 554–568.

BIOGRAPHIES OF AUTHORS






H. N. Lakshmi    is currently working as professor and head of the Department of Emerging Technologies at CVR college of Engineering-Hyderabad. She has more than 22 years of teaching experience. She completed Ph.D. in computer science from University of Hyderabad in 2016 in the area of Web Services and M.S (software systems) from BITS, Pilani in 1999. Her areas of interest include web services, analysis of algorithms, programming languages, data structures, network security, data mining, machine learning and soft computing. She has authored more than 16 research papers, one book chapter in Springer, out of which 5 papers are DBLP and 9 are Scopus indexed and has around 20 citations. She can be contacted at email: hn.lakshmi@cvr.ac.in.






Thaduri Venkata Ramana    completed M. Tech from CBIT, Osmania University in the year 2008. Completed research from JNTU Hyderabad and achieved Ph.D. in the field of image processing. He worked as one of the directors for a software training and development organization "Crystallite Technologies Pvt Ltd". He is having total 23+ years of experience in corporate and academics. He has published nearly 20 papers in different national and international journals. Currently working as associate professor in Department of CSE-AIML at CVR College of Engineering, Hyderabad, and Telangana. His areas of interests are databases, image processing, network security, cloud computing, project management. For any inquiries or further communication, he can be contacted at email: meetramana1204@gmail.com.






LNC Prakash K    awarded doctorate in computer science and engineering from JNTU Hyderabad, A State Government University, Hyderabad, India, He has more than 24 years of teaching and 15 years of research experience. He has 18 research publications in reputed journals which are indexed by SCI, SCOPUS, and UGC. He is currently working as an associate professor in the Department of Computer Science and Engineering (Data science), CVR College of Engineering, Hyderabad, India. For any inquiries or further communication, he can be contacted at email: klnc.prakash@gmail.com.



L. Kiran Kumar Reddy    currently working as an associate professor and head, Department of CSE (Data science), Geethanjali College of Engineering and Technology, Hyderabad, Telangana. He has received his Ph.D. in computer science and engineering from GITAM University, India. He has total of 21 years of teaching experience at UG and PG level. His main areas of research include databases, data mining, machine learning, and data science. For any inquiries or further communication, he can be contacted at email: kirankumarreddy.cse@gcet.edu.in



Kachapuram Basava Raju    has 22 years of experience in teaching both in graduate and undergraduate level. He received doctorate degree in CSE from JNTU and He received M.Tech. (CSE) from Osmania University, and Undergraduate in computer science from SK University. He worked around 5 years for abroad as assistant professor in IT Dept. Presently he has been working for AI Department of Anurag University from past 2 + Years. His areas of interest to teach is machine learning, image mining, artificial intelligence, Java. He has authored more than 20+ research papers, one book chapter in SCI. He can be contacted at email: kbajuai@anurag.edu.in.