Journal of Information Systems Engineering and Management
2025, 10(25s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Integrating Enhanced Clustering Algorithms and Ensemble Techniques for Bigdata

Anand N. Gharu[1], Shyamrao V. Gumaste[2]
[1]Research Scholar, Department of Computer Engineering, MET's Institute of Engineering, Adgaon Nashik, Maharashtra, India.
(Affiliated to Savitribai Phule Pune University)
Email:gharu.anand@gmail.com
[2]Professor, Department of Artificial Intelligence and Data Science, MET's Institute of Engineering, Adgaon Nashik, Maharashtra, India.
Email:svgumaste@gmail.com

ARTICLE INFO

Received: 19 Dec 2024

Revised: 10 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

The proposed system introduces an enhanced clustering algorithm optimized for big data analytics, addressing challenges such as scalability, heterogeneity, and high dimensionality. Utilizing an ensemble clustering approach combined with a voting mechanism, the system generates robust and accurate cluster outcomes suitable for diverse applications. The invention leverages cloud-based platforms for efficient processing of large-scale datasets while ensuring adaptability to numerical, categorical, and mixed data types. Comprehensive performance evaluation metrics, including silhouette score and Davies-Bouldin index, provide insights into clustering quality and efficiency. The system's design supports real-time applications in healthcare, finance, IoT, and AI, emphasizing scalability and precision. This innovation contributes to the advancement of artificial intelligence and data mining technologies by delivering an adaptable, efficient, and scalable clustering solution tailored for big data environments.

Keywords : Data Mining, Clustering Algorithms, Machine Learning, Bigdata.

I. INTRODUCTION

The proposed system lies in the domain of data mining, artificial intelligence, and big data analytics with a focus on enhancing clustering techniques for efficient decision-making in cloud-based environments. It integrates advanced data mining algorithms to address the challenges of handling massive, heterogeneous datasets often encountered in big data applications. The invention targets the optimization of clustering methods by utilizing ensemble approaches that combine multiple clustering algorithms, ensuring improved accuracy and adaptability to various data types and scenarios. By incorporating a robust voting mechanism, the system ensures precise cluster finalization tailored to specific use cases. This invention is particularly significant for applications involving both numerical and non-numerical attributes, enabling versatile and scalable solutions for AI-driven decisions in diverse domains such as healthcare, finance, IoT, and smart systems. Additionally, the system emphasizes performance evaluation and comparative analysis, establishing benchmarks for efficiency and effectiveness in big data clustering. Even though there are many surveys for clustering algorithms available in the literature [1], [2], [3], and [4] for a range of domains (such as machine learning, data mining, information retrieval, pattern recognition, bio-informatics, as well as semantic ontology), it is difficult for users to decide a priori which algorithm would be the most appropriate for a given big dataset. This is because there are already certain limitations on the survey: This is due to three factors: (i) the algorithms' characteristics have not been fully examined; (ii) the field has produced numerous new algorithms that were not considered in these surveys; and (iii) no comprehensive empirical analysis has been conducted to ascertain the advantages of one algorithm over another.

Proposed System Architecture :

- Input Dataset is pre-processed.
- A novel EECA algorithms consist five algorithm i.e. Enhanced K-means, Enhanced K-Medoids, Enhanced Fuzzy C Means, Enhanced DBSCAN and Enhanced EM Clustering. It generates clusters for each algorithms for given dataset.
- (Proposed) Majority Voting techniques will be applied to cluster generated by each algorithms to identify efficient cluster.
- Finally, generates optimal cluster as final output.

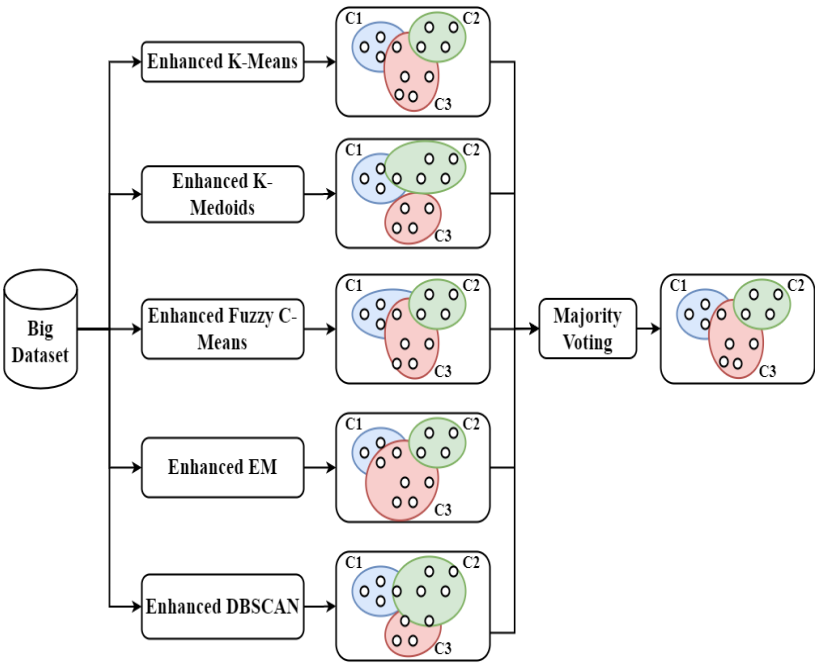

Figure 1: Proposed System Architecture

II. RELATED WORK

The increasing prevalence of big data in today's digital era has posed significant challenges to traditional data processing and analysis techniques. With the exponential growth of structured, semi-structured, and unstructured data, industries across various sectors demand robust mechanisms to extract meaningful insights efficiently. Among the many facets of big data analytics, clustering plays a crucial role as an unsupervised machine learning technique, enabling the segmentation of vast datasets into meaningful groups or clusters. This capability is indispensable in a wide array of applications, including market segmentation, anomaly detection, recommendation systems, and predictive analytics. However, as datasets grow in size and complexity, traditional clustering algorithms often fail to deliver the required performance in terms of accuracy, scalability, and adaptability. These challenges necessitate the development of innovative approaches capable of addressing the unique requirements of big data environments.

One of the critical challenges in clustering large datasets is handling the inherent heterogeneity of big data. Unlike conventional datasets that may consist predominantly of numerical attributes, big data often includes mixed data types, such as categorical, textual, and temporal data. Traditional clustering algorithms are typically optimized for specific data types, resulting in suboptimal performance when applied to diverse datasets. Furthermore, the high dimensionality of big data presents additional obstacles, as it increases computational overhead and can obscure meaningful patterns due to the curse of dimensionality. These challenges call foradvanced clustering techniques capable of managing heterogeneity and high dimensionality effectively while maintaining computational efficiency.

Another significant limitation of traditional clustering algorithms is their sensitivity to initial conditions and parameter selection. Algorithms such as k-means, for instance, rely heavily on the choice of initial centroids, which can lead to inconsistent results and suboptimal clustering outcomes. Similarly, hierarchical clustering methods are often constrained by the choice of linkage criteria, making them less adaptable to varying dataset characteristics. Such limitations highlight the need for ensemble-base approaches that leverage the strengths of multiple clustering algorithms, thereby mitigating the shortcomings of individual methods and enhancing overall clustering performance.

A notable innovation of the proposed system is its use of ensemble clustering combined with an efficient voting mechanism for cluster finalization. Ensemble clustering involves generating multiple clustering solutions using diverse algorithms or parameter settings and subsequently combining these solutions to produce a consensus clustering result. This approach enhances the robustness and accuracy of clustering outcomes by reducing the impact of individual algorithmic biases and errors. The proposed system further refines this process by employing a voting mechanism that selects the most appropriate clusters based on predefined criteria, ensuring that the final clustering result aligns with the specific requirements of the application scenario.

Performance evaluation is a cornerstone of the proposed invention, emphasizing the importance of measurable improvements in clustering outcomes. The system incorporates comprehensive performance metrics to assess the quality and efficiency of clustering results, including measures such as silhouette score, Davies-Bouldin index, and computational time. Additionally, the proposed system facilitates comparative analysis of clustering performance across datasets with numerical and non-numerical attributes, enabling a holistic understanding of its effectiveness in diverse contexts. By benchmarking the

proposed system against existing clustering techniques, the invention provides valuable insights into its advantages and limitations, paving the way for further refinement and adaptation.

III. ALGORITHMS

1. **Enhanced K-Means Clustering**:
    1. Initialize k cluster centroids by randomly selecting k points from the dataset X.
    2. Repeat until convergence or maximum iterations:
       a. Assign each data point to the cluster with the nearest centroid.
       b. Update the centroids by calculating the mean of all data points in each cluster.
    3. Return the final cluster assignments and centroids.
2. **Enhanced K-Medoids Clustering**:
    1. Initialize k cluster medoids by randomly selecting k data points from the dataset X.
    2. Repeat until convergence or maximum iterations:
       a. Assign each data point to the cluster with the nearest medoid.
       b. Update the medoids by selecting the data point in each cluster that minimizes the sum of distances to all other data points in the cluster.
    3. Return the final cluster assignments and medoids.
3. **Enhanced Fuzzy C-Means (E-FCM) Clustering**:
    1. Initialize the fuzzy partition matrix U with random values.
    2. Repeat until convergence or maximum iterations:
       a. Calculate the cluster centroids using the current fuzzy partition matrix U.
       b. Update the fuzzy partition matrix U using the calculated cluster centroids.
    3. Return the final cluster centroids and fuzzy partition matrix.
4. **Enhanced Expectation-Maximization EM Clustering**:
    1. Initialize the Gaussian Mixture Model (GMM) with n_clusters components, using a full covariance matrix and a higher maximum iteration limit.
    2. Fit the GMM to the dataset X using the Expectation-Maximization algorithm.
    3. Obtain the cluster labels by predicting the cluster assignments for each data point.
    4. Return the final cluster labels and the GMM parameters (means, covariances).
5. **Enhanced DBSCAN (E-DBSCAN) Clustering**:
    1. For each data point in the dataset X:
       a. Compute the radius of the neighborhood around the data point that contains at least min_samples points.
       b. If the number of points in the neighborhood is greater than or equal to min_samples, assign the data point to a cluster.
       c. If the number of points in the neighborhood is less than min_samples, assign the data point as noise.
    2. Return the final cluster labels.

IV. ENSEMBLE CLUSTERING

In the field of ensemble learning, cluster ensemble is the correct alternative for solving so many issues related to traditional cluster model. Ensemble cluster model aims to combine the different partitions obtained from several base models. The results show that ensemble model provides high accuracy compared to single base model. In [21] presented an ensemble model to merge the results generated by several base cluster models. To obtain best consensus partition they adopted majority voting principle. They selected simple K-means algorithm as base learning model applied on different random data samples. Later they combined using voting approach refereed as voting-k-means algorithm. Because of the ensemble approach, the proposed algorithm able to recognize different cluster shapes other than hyper-spherical. Another powerful consensus clustering approach is the Voting-Merging method proposed by authors in [20]. The proposed approach will be performed in two stages. In the first stage, apply voting process by considering votes from each cluster model. Later, it combines all the votes to produce better partition.

In [21] presented a popular consensus function to generate best partition i.e. voting active clusters (VAC). They provide an adaptive voting method used to increase the cluster quality by frequent updation of votes. This model have an additional capability like, it is able to collect the data from different sources to cluster the data. In this approach an individual cluster processing model produces many base learner models to handle some set of data. At last, all the individual cluster models results will be pooled with voting process to produce best consensus partition. The results show that, if the data is noise free; the proposed method provides high accuracy compared with classical ensemble cluster models. In [22] proposed a new ensemble cluster model based on co-association (CA) matrix. To obtain more reliable results, they proposed hierarchical cluster model to obtain the quality clusters based on normalized edges. The performance analysis shows that, the proposed approach produces better accuracy compared with individual runs of classical techniques and some ensemble approaches. Authors in [23] presented an ensemble cluster model based on popular optimization technique called genetic algorithm. They proposed an

ensemble approach based on the information theory referred as IT-GA. In this algorithm they adopted correspondence matrix construction to derive consensus partition. Later, to re label the partition they adopted Hungarian method. Finally, they compared results with traditional ensemble models. The proposed model outperforms in performance measures.

V. RESULTS

1. K-means clustering :

Output of K-Means :

Cluster 1 Centroid: [-6.8840719 -8.83396959 7.33269085 2.0397464 4.19345145]
Cluster 2 Centroid: [-9.60286527 9.38703798 6.6694459 -5.76120151 -6.35551522]
Cluster 3 Centroid: [-2.52117634 9.02250119 4.63531694 1.97306693 -6.90190802]

2. Enhanced K-means clustering :

Output of Enhanced K-Means :

Cluster 1 Centroid: [-9.60286527 9.38703798 6.6694459 -5.76120151 -6.35551522]
Cluster 2 Centroid: [-2.52117634 9.02250119 4.63531694 1.97306693 -6.90190802]
Cluster 3 Centroid: [-6.8840719 -8.83396959 7.33269085 2.0397464 4.19345145]

VI. CONCLUSION

The proposed system represents a significant advancement in the field of data mining and big data analytics, offering a novel solution to the challenges of clustering large-scale and heterogeneous datasets. Its innovative ensemble clustering approach, combined with a robust voting mechanism and cloud-based scalability, ensures high performance, adaptability, and accuracy. With its broad applicability across various domains, the proposed system holds promise for driving innovation and efficiency in big data analytics, contributing to the advancement of artificial intelligence and machine learning technologies. As the digital landscape continues to evolve, the invention stands as a testament to the transformative potential of advanced clustering techniques in enabling intelligent and efficient data-driven decision-making.

REFRENCES

[1] A. A. Abbasi and M. Younis, ``A survey on clustering algorithms for wireless sensor networks," Comput. Commun., vol. 30, nos. 1415, pp. 28262841, Oct. 2007.
[2] C. C. Aggarwal and C. Zhai, ``A survey of text clustering algorithms,"in Mining Text Data. New York, NY, USA: Springer-Verlag, 2012,pp. 77128.
[3] J. Brank, M. Grobelnik, and D. Mladeni¢, ``Asurvey of ontology evaluation techniques," in Proc. Conf. Data Mining DataWarehouses (SiKDD), 2005.
[4] R. Xu and D. Wunsch, ``Survey of clustering algorithms," IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645678, May 2005.
[5] Xu, Rui & Wunsch, Donald. “Survey of Clustering Algorithms”. Neural Networks, IEEE Transactions on. 16.pp: 645 - 678. DOI:10.1109/TNN.2005.845141.Nov 2005.

[6] Abdolreza Hatamlou,"In search of optimal centroids on data clustering using a binary search algorithm",Article in Pattern Recognition Letters 33(13):1756– 1760 • ,DOI: 10.1016/j.patrec.2012.06.008. October 2012.
[7] Dr. N. Rajalingam, K.Ranjini," Hierarchical Clustering Algorithm - A Comparative Study", International Journal of Computer Applications , Vol. 19– No.3, Apr 2011.
[8] PoojaNagpal and Priyanka Mann, "Survey of Density Based Algorithms" in International Journal of Computer Science and Applications.2011.
[9] JianbingShen, XiaopengHao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", IEEE Transactions, 1057-7149 , 2016.
[10] Ahmad M. Bakr, Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2014, Elsevier Pvt. Ltd.
[11] Park HS, Jun CH. A simple and fast algorithm for K-me- doids clustering. Expert Systems Applications. 2009 Mar; 36(2.2):3336–41.
[12] Huang Z. A fast clustering algorithm to cluster very large cate- gorical data sets in data mining. Proceedings SIGMOD Work- shop Res Issues Data Mining Knowl Discovery; 1997. p. 1–8.
[13] Bezdek JC, Ehrlich R, Full W. FCM: The Fuzzy C-Means Clus- tering algorithm. Computers and Geosciences. 1984; 10(2-3):191–203.
[14] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Trans- actions on Neural Networks. 2005 May; 16(3):645–78.
[15] Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise. Proceedings ACM SIGKDD Conf Knowl Discovery Ad Data Mining (KDD); 1998. p. 58–65.
[16] Cheng CH, Fu AW, Zhang Y. Entropy based sub space cluster- ing for mining numerical data. Proceedings of the fifth ACM SIGMOID International Conference on Knowledge discovery and Data Mining; 1999. p. 84–93.
[17] Berkhin P. Survey of clustering data mining techniques in grouping multidimensional data. Springer. 2006; 25–71
[18] E. Dimitriadou, A. Weingessel and K. Hornik, An ensemble method for clustering, ICANN (2001), pp. 217-224
[19] J.-F. Laloux, N.-A.Le-Khac, and M.-T. Kechadi, "Efficient distributed approach for density-based clustering," Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 20th IEEE International Workshops, pp. 145–150, 27-29 June 2011.
[20] A. Fred, Finding consistent clusters in data partitions, 3rd. Int. Workshop on Multiple Classifier Systems (2001), pp. 309-318.
[21] E. Dimitriadou, A. Weingessel and K. Hornik, An ensemble method for clustering, ICANN (2001), pp. 217-224
[22] Y. Li, J. Yu, P. Hao and Z. Li, Clustering ensembles based on normalized edges Vol. 4426 (Springer-Verlag BerlinHeidelberg, 2007), pp. 664-671.
[23] H. Luo, F. Jing and X. Xie, Combining multiple clusterings using information theory based genetic algorithm, IEEE Int. Conf. Computational Intelligence and Security (2006) 84-89.