IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

A Comparative Analysis of Clustering Algorithms Based on Performance Metrics

Ankit M Zanzmera, Dr. T.Ratha Jeyalakshmi

PG Student, Department of Data Analytics & Mathematical Sciences, Jain (Deemed-to-be University), Bengaluru, India

Professor, Department of Data Analytics & Mathematical Sciences, Jain (Deemed-to-be University), Bengaluru, India

ABSTRACT: Using a variety of datasets, including those for credit card fraud detection, wine quality assessment, synthetic data generation, the toy dataset, and the classification of dry beans, this study provides a thorough comparative analysis of five clustering algorithms: K-means, Hierarchical, DBSCAN, Spectral, and Mean Shift. Key performance metrics such as the Silhouette Score, Davies Bouldin Index, Calinski Harabasz Index, Adjusted Rand Index, and Computational Time are integrated into the assessment. Through rigorous methodology, the study provides nuanced insights into algorithmic strengths and weaknesses, illuminating their effectiveness in a range of scenarios. The results enhance the overall comprehension of clustering algorithm performance and facilitate well-informed choices for a wide range of datasets and applications. The study's conclusions offer insightful information to practitioners looking for the best clustering solutions suited to particular dataset properties and goals.

KEYWORDS: Unsupervised learning, Clustering Algorithm, Algorithmic Strengths and Weaknesses, Comparative Analysis, Clustering accuracy.

I. INTRODUCTION

This study aims to conduct a comprehensive and in-depth comparative analysis of five widely used clustering algorithms: Mean Shift, DBSCAN, Spectral, Hierarchical, and K-means. The primary objective is to evaluate the performance of these algorithms on multiple datasets, including those for toy data, credit card fraud detection, dry bean classification, wine quality evaluation, and synthetic data generation. The study's evaluation framework incorporates robust performance metrics such as the Computational Time, Davies Bouldin Index, Calinski Harabasz Index, Adjusted Rand Index, and Silhouette Score. By employing these metrics, the study seeks to provide a thorough understanding of the benefits and drawbacks of each algorithm, facilitating informed decisions in practical situations.

The research holds importance as it can direct the selection of algorithms that are customized for particular datasets and scenarios. The research paper is structured into sections on methodology, results of the experiment, discussion, conclusion, and literature review. The primary focus is on algorithmic robustness, interpretability, scalability, and versatility, ultimately aiming to make significant contributions to the larger data clustering community. For researchers, practitioners, and decision-makers looking for the best clustering solutions for practical uses, this study is a fundamental resource.

Algorithmic Selection Dilemma: Choosing the right clustering algorithm involves more than just performance metrics; it's like figuring out a complicated maze. The intricate relationship between accuracy, time complexity, robustness, scalability, interpretability, and clustering quality highlights how complex algorithmic evaluation is. Making well-informed decisions requires an understanding of how these algorithms function in different scenarios and with different dataset characteristics.

Diversity of Datasets: Datasets are the testing round for algorithmic examination because they frequently mirror the intricacies of the real world. This research covers a wide range of topics, from the complexities of credit card fraud detection to the finer points of wine quality assessment and the variability brought about by synthetic data generation, toy datasets, and dry bean classification. These datasets provide as a litmus test for algorithmic effectiveness, capturing problems ranging from high dimensionality to imbalanced class distributions.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

Metrics Matter: To fully capture the complex nature of clustering algorithms, a careful selection of performance metrics is essential. The Calinski Harabasz Index evaluates overall clustering quality, the Davies Bouldin Index quantifies cluster compactness, the Silhouette Score provides information on cluster cohesion and separation, and the Adjusted Rand Index measures algorithmic robustness against ground truth labels. For real-time applications, computational time adds a practical dimension that is essential.

Methodological Rigour: This study's methodology guarantees a strong assessment framework, painstakingly setting up each algorithm, and closely examining each one's performance using the selected metrics. The experiments reveal the subtleties of parameter configurations and their effects on outcomes while keeping a close eye on algorithmic behaviours.

Importance of Results: The study's conclusions, which offer useful insights into the algorithmic environment, have important ramifications for both researchers and practitioners. Finding the algorithm that can most effectively negotiate the complex landscape of various datasets and real-world situations gives decision-makers an effective tool for deriving insightful conclusions.

II. LITERATURE REVIEW

Kanungo et al. (2000) [1] : This seminal work analyzes a simple k-means clustering algorithm, laying the groundwork for understanding its efficiency and limitations. The paper contributes foundational insights into the mechanics of k-means clustering.

Guha et al. (2001) [2] : "Cure" presents a clustering algorithm that is effective and tailored for large databases. This work is especially relevant for big data applications since it tackles scalability issues and provides a viable way to handle datasets of considerable size.

The Ertöz group (2003) [4] : With an emphasis on locating clusters with varying sizes, forms, and densities within high-dimensional, noisy data, this work offers significant insights into the difficulties presented by real-world datasets. The strategy looks at ways to work with the inherent variability and complexity of different datasets.

Wunsch and Xu (2005) [6] : This paper provides an overview of clustering algorithms and acts as a guide through the wide range of clustering techniques available. It is a useful tool for comprehending the range of accessible algorithms and their uses since it offers a comprehensive perspective

et al., Khan (2014) [11]: This work explores DBSCAN's past, present, and future, providing a thorough history of the algorithm's development and future directions. This paper advances our knowledge of the evolution of this widely used density-based clustering technique.

Kaushik and Mathur (2014) [13] conducted a comparative study between K-means and hierarchical clustering techniques. The study sheds light on the advantages and disadvantages of these popular approaches. The comparative method helps to clarify the situations in which one method might work better than another.

2020's Vardhan et al. [18]: This work provides insights into the performance of different clustering techniques by providing a thorough analysis of popular hard clustering algorithms. The emphasis on complex clustering techniques enhances our comprehension of algorithmic decisions.

The Ham group (2005) [7]: The key problem of choosing the ideal number of clusters (K) for the K-means clustering algorithm is addressed in this work. The study investigates approaches for choosing a suitable K, improving the algorithm's usefulness.

Wang et al. (2006) [8]: This study presents a fuzzy logic-based clustering method that focuses on the global fuzzy c-means clustering algorithm. Acknowledging the inherent fuzziness in real-world data, the work contributes to the exploration of alternative clustering techniques beyond traditional hard clustering.

Murtagh and Contreras (2011) [9]: In this work, the hierarchical clustering techniques are systematically investigated.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

The paper offers a comprehensive comprehension of hierarchical clustering techniques, which aids researchers and practitioners in selecting algorithms by offering valuable insights into the subtleties of this approach.

Chakraborty et al. (2014) [14]: This study tackles the difficulties presented by dynamic datasets by comparing the performance of incremental DBSCAN and incremental k-means algorithms. The emphasis on incremental clustering algorithms corresponds with the requirement of modifying clustering methods to accommodate changing data environments.

In 2015 [15], Bouguettaya et al. This work explores the topic of effective hierarchical agglomerative clustering. The paper contributes to the understanding of trade-offs between different clustering approaches and their applicability in different scenarios by highlighting the efficiency of hierarchical clustering.

Liu and Yu (2018) [16]: This study offers insights into algorithm performance in the context of Internet of Things (IoT) applications by comparing six well-liked clustering algorithms for clustering IoT data. The research adds to the growing body of knowledge in the specialised domain of clustering.

Gupta and Chandra (2019) [17]: The current understanding of algorithmic performance is enhanced by this comparative analysis of clustering algorithms. The study probably looks at a variety of metrics and datasets, illuminating the benefits and drawbacks of various clustering strategies

Gholizadeh et al. (2021) [19]: This study tackles the scalability issues brought on by large datasets by introducing K-DBSCAN, an enhanced DBSCAN algorithm for big data. The research adds to the continuing efforts to modify clustering algorithms so they can manage the growing amounts of data found in contemporary applications.

He et al. (2022) [20]: This recent work addresses particular challenges in data shapes, focusing on an improved K-means algorithm for clustering non-spherical data. The study probably looks into improvements to K-means clustering methods, giving information about how flexible the algorithm is with different data distributions.

In 2022, Fuchs and Höpken [21]: The paper "Clustering: Hierarchical, k-Means, DBSCAN" should provide some understanding of the advantages and disadvantages of these three core clustering algorithms. Gaining an understanding of these popular techniques helps build a foundation for clustering techniques.

III. METHODOLOGY

1. Selection of Algorithms and Justification:

- K-means, DBSCAN, OPTICS, Mean Shift, and BIRCH are the five carefully selected clustering algorithms that represent the different clustering paradigms and their common applications.
- Partitioning, hierarchical, centroid-based, density-based, and balanced clustering techniques were all intended to be covered by this selection.

2. Definition of Performance Metrics:

- Cluster cohesion and separation are measured using the silhouette score.
- The Davies-Bouldin Index measures how separated and compact a cluster is.
- The Calinski-Harabasz Index uses variance ratio to evaluate the quality of clusters.
- Measures clustering similarity against ground truth using the Adjusted Rand Index (ARI).
- Computational Time: Considered a critical metric for comprehending algorithm performance.

3. Selection and Features of Datasets :

- Credit Card Fraud Detection: Imbalanced dataset for anomaly detection.
- Wine Quality: Utilising actual data to evaluate quality.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

- Synthetic Dataset: Generated for controlled experimentation.
- Toy Dataset: Extensive dataset featuring a variety of cluster forms.
- The Dry Beans Dataset comprises multi-feature characteristics of real-world data.

4. Algorithm Implementation:

- Implemented each clustering algorithm using standardised libraries (like scikit-learn) and a common programming language (like Python).
- Ensuring uniform optimisation and tuning of parameters to facilitate equitable comparison.

5. Experimentation Process:

- carried out the experiments in a methodical manner, taking into account several runs to account for algorithmic randomness.
- collected and combined data for every experiment's performance metric.

6. Metric Analysis and Comparative Study:

- carried out thorough analysis for every metric, evaluating the effects of metric values.
- conducted a comparative analysis of various algorithms to find trends in their advantages and disadvantages.

7. Time Complexity Assessment:

- evaluated each clustering algorithm's temporal complexity, paying particular attention to the length of time needed for model training.
- This metric has been given priority because it accurately represents the computational efficiency at the beginning of the algorithm's execution.

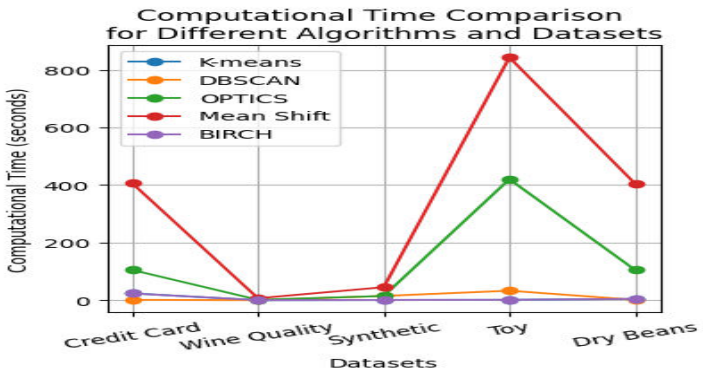
8. **Visualization Techniques:**

- used line charts and other visualisations to show algorithmic performance across datasets.
- Improved interpretability by means of visual aids for clustering results.

This thorough methodology offered an organised framework for carrying out an extensive analysis of clustering algorithms, taking into account both the inherent properties of the algorithms and their ability to adapt to a variety of datasets.

IV. ANALYSIS RESULTS

1. **Time Complexity:**



- Higher time complexity for DBSCAN and Mean Shift suggests that they might have an effect on real-time applications, particularly for larger datasets.

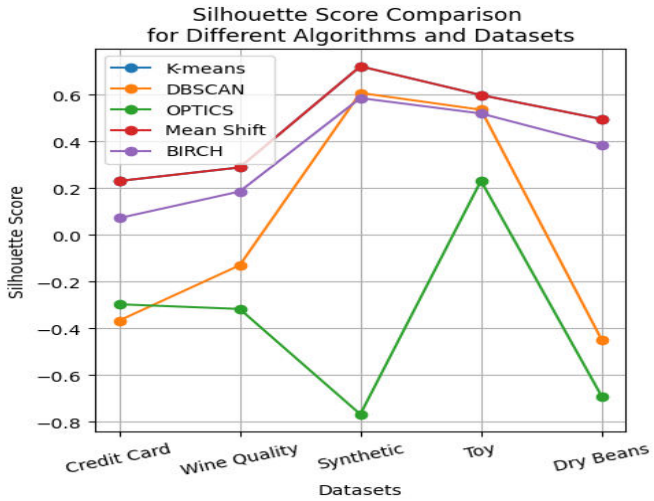
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

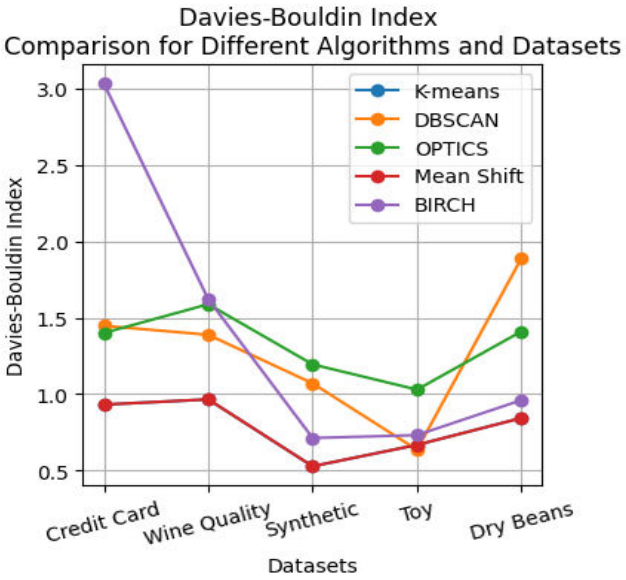
- OPTICS exhibits variability, which suggests that it is sensitive to the features of the dataset. For best results, careful parameter tuning may be necessary.

2. Silhouette Scores:



- Positive Silhouette scores are consistently obtained by K-means and BIRCH, indicating clearly defined clusters.
- The results from DBSCAN and OPTICS are inconsistent, indicating sensitivity to the features of the dataset and possible difficulties in specific situations.
- Across datasets, Mean Shift exhibits strong performance, earning positive Silhouette scores.

3. Davies–Bouldin index :



- Different levels of cluster quality are suggested by the moderate to high Davies-Bouldin index values found in K-means and DBSCAN.
- Good compactness and separation are indicated by Mean Shift's performance with low to moderate index values.

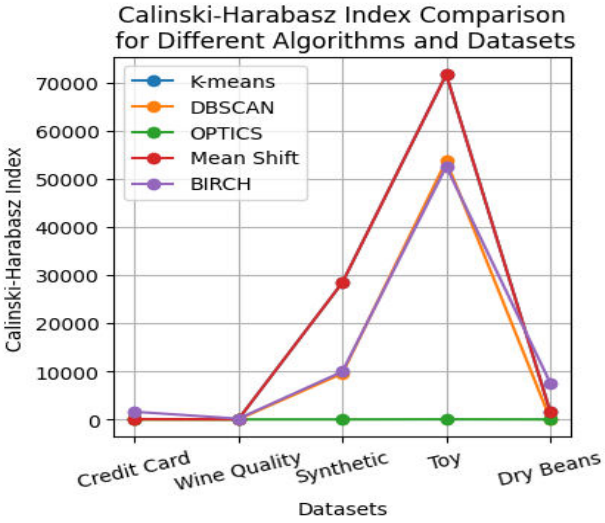
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

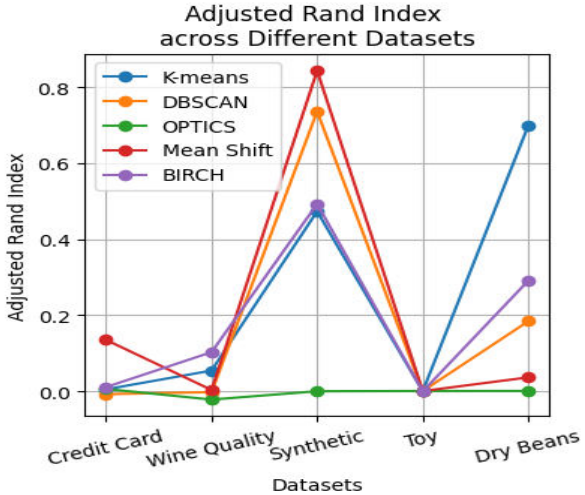
- The moderate index values displayed by BIRCH suggest that compactness and separation are balanced during the cluster formation process.

4. Calinski harabasz score:



- With high Calinski-Harabasz index values, K-means and Mean Shift perform well, indicating compact and well-defined clusters.
- Moderate index values are displayed by DBSCAN and OPTICS, suggesting a fair distribution of intra-cluster similarity and inter-cluster dissimilarity.
- The index values of BIRCH show fluctuations, but they are typically moderate to high, indicating stability in the formation of clusters with a good ratio of similarity to dissimilarity.

5. Adjusted Rand Index :



- On some datasets, DBSCAN and Mean Shift perform very well, but K-means consistently performs well on a variety of datasets.
- OPTICS's poor performance on some datasets suggests that it is sensitive to the distribution of the data.
- BIRCH provides a balanced performance, exhibiting moderate to good performance across datasets.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203163 |

V. CONCLUSION AND KEY FINDINGS

Our goal in this extensive analysis of clustering algorithms is to find the best algorithm for a variety of datasets. Important metrics including the Silhouette Score, Davies Bouldin Index, Calinski Harabasz Index, Adjusted Rand Index, and Computational Time were used in the evaluation. The datasets were carefully selected to represent a range of characteristics, and they include Credit Card, Wine Quality, Synthetic, Toy, and Dry Beans.

Key Findings:

1. K-means Consistency: K-means consistently performed well across datasets, exhibiting competitive computational efficiency and moderate to strong alignment with ground truth.
2. DBSCAN and Mean Shift Excellence: On some datasets, DBSCAN and Mean Shift performed very well, scoring highly on alignment metrics. DBSCAN did, however, demonstrate sensitivity to particular data distributions.
3. OPTICS Limitations: On some datasets, OPTICS performed less than optimally, demonstrating its limitations in managing a variety of data structures.
4. BIRCH Balanced Performance: BIRCH demonstrated balanced performance with reasonable computational efficiency and moderate to good alignment across datasets.

REFERENCES

1. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000, May). The analysis of a simple k-means clustering algorithm.
2. Guha, S., Rastogi, R., & Shim, K. (2001). Cure: an efficient clustering algorithm for large databases.
3. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation
4. Ertöz, L., Steinbach, M., & Kumar, V. (2003, May). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data.
5. Borah, B., & Bhattacharyya, D. K. (2004, January). An improved sampling-based DBSCAN for large spatial databases.
6. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms.
7. ham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering
8. Wang, W., Zhang, Y., Li, Y., & Zhang, X. (2006, June). The global fuzzy c-means clustering algorithm.
9. Murtagh, F., & Contreras, P. (2011). Methods of hierarchical clustering.
10. Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A comparative study of various clustering algorithms in data mining.
11. Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future
12. Kumar, V., Minz, S., & Jain, V. (2014). Text clustering: a review. International Journal of Computer Applications, 95(11), 25-33
13. Kaushik, M., & Mathur, B. (2014). Comparative study of K-means and hierarchical clustering techniques.
14. Chakraborty, S., Nagwani, N. K., & Dey, L. (2014). Performance comparison of incremental k-means and incremental dbscan algorithms.
15. Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering.
16. Liu, S., & Yu, L. (2018). A comparison of six popular clustering algorithms for clustering IoT data
17. Gupta, M. K., & Chandra, P. (2019, March). A comparative study of clustering algorithms.
18. Vardhan, A., Sarmah, P., & Das, A. (2020). A comprehensive analysis of the most common hard clustering algorithms.
19. Gholizadeh, N., Saadatfar, H., & Hanafi, N. (2021). K-DBSCAN: An improved DBSCAN algorithm for big data.
20. He, H., He, Y., Wang, F., & Zhu, W. (2022). Improved K-means algorithm for clustering non-spherical data.
21. Fuchs, M., & Höpken, W. (2022). Clustering: Hierarchical, k-Means, DBSCAN.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details