International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Evaluation of K-Means, Hierarchical Clustering, and DBSCAN in Blood Donor Segmentation

Srilekha. S[1], Priyadharshini.P[2], Adhilakshmi. M[3]

[1,2,3]Sri Venkateswaraa College of Technology

Abstract:
Clustering techniques are pivotal in the fields of data analysis and pattern recognition, offering significant insights by grouping data points with similar characteristics. This study aims to perform a comprehensive comparison of three widely used clustering algorithms—K-Means, Hierarchical Clustering, and DBSCAN—on a dataset of blood donors. The objective is to determine which algorithm achieves the most precise and effective clustering of the data, taking into account factors such as donor location, blood type, and donation frequency. The study presents a novel approach by integrating a web-based platform that allows blood donors to register online. This platform not only facilitates the real-time updating of the dataset but also enhances the overall relevance and applicability of the clustering model by continuously incorporating new data entries. By leveraging such a dynamic dataset, the clustering algorithms can adapt to evolving patterns and trends, ensuring more accurate and meaningful insights over time.To rigorously evaluate the performance of each clustering method, several well-established metrics are employed, including the Silhouette Score, which assesses how similar each data point is to its own cluster compared to other clusters; the Davies-Bouldin Index, which evaluates the average similarity ratio of each cluster with its most similar cluster; and the Calinski-Harabasz Index, which measures the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters. The results of this study indicate that the K-Means algorithm consistently outperforms both Hierarchical Clustering and DBSCAN in terms of accuracy and the clarity of cluster definitions. The findings underscore the robustness of K-Means for applications involving blood donor data, where capturing precise donor groupings can have substantial implications for healthcare logistics and resource allocation. These insights pave the way for further research into the optimization of clustering techniques in dynamic datasets and their practical applications in medical and other domains.

1. Introduction:

Clustering is an essential unsupervised machine learning technique used to identify natural groupings in data. This technique is particularly useful in scenarios where there is no predefined label or category, such as in healthcare data analysis, customer segmentation, and pattern recognition. In the context of blood donation, clustering can help identify groups of donors based on geographic location, blood type, and donation frequency, which can significantly enhance the efficiency of blood collection and distribution.
In this study, we explore three widely used clustering algorithms—K-Means, Hierarchical Clustering, and DBSCAN—on a blood donor dataset. The goal is to evaluate which algorithm provides the most accurate clustering and can adapt effectively to dynamic changes in the dataset. By using a web-based platform for

donor registration, we ensure the dataset remains current, allowing for more responsive and relevant clustering outcomes.

2. Literature Review:

Clustering techniques have been extensively studied across various domains. K-Means is known for its simplicity and efficiency in partitioning data into a pre-defined number of clusters. However, it is sensitive to the initial placement of centroids and can struggle with clusters of varying densities and non-globular shapes. In contrast, Hierarchical Clustering does not require specifying the number of clusters beforehand and is capable of revealing the nested structure within data through dendrograms. Yet, it is computationally intensive and not well-suited for large datasets. DBSCAN has the advantage of identifying clusters of arbitrary shape and handling noise effectively, but its performance can be sensitive to the choice of parameters ϵ\epsilonϵ (the neighborhood radius) and minPtsminPtsminPts (minimum number of points in a neighborhood).

Prior research has applied these clustering methods to various datasets, but a direct comparison using blood donor data has not been thoroughly explored. This study aims to fill this gap by comparing the performance of these algorithms using a robust set of evaluation metrics.

3. Methodology:

3.1 Dataset Description:

The blood donor dataset used in this study includes a variety of features: donor names, age, gender, blood type, medical conditions, allergies, blood pressure, height, weight, contact details, geographical information (latitude and longitude), and donation history, including recency, frequency, and volume of donations. This rich dataset provides a comprehensive view of each donor, allowing for meaningful clustering based on both geographical and behavioral attributes.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Names	Age	Gender	Blood Typ	Medical Cr	Allergies	Blood Pres	Height (cm	Weight (kg	Contact N	Country	State	District	Location	Pincode	Latitude	Longitude	Recency (r	Frequency	Monetary	Time (months)
2	John Smith		20 Male	O+	Hypertens	Pollen	120/80	175	70	1.23E+09	India	Assam	Tinsukia	Adarsngao	786174	27.6133	95.1425	2	50	12500	98
3	Emma Joh		25 Female	A+	Diabetes	Peanuts	130/85	160	65	1.99E+09	India	Assam	Tinsukia	Agbandha	786187	27.5847	95.2436	0	13	3250	28
4	Michael W		32 Male	B+	No	Dust mites	110/70	180	80	1.12E+09	India	Assam	Tinsukia	Akabari	786181	27.5472	95.3279	1	16	4000	35
5	Sophia Brc		45 Female	AB+	No	Shellfish	140/90	165	55	1.56E+09	India	Assam	Tinsukia	Amarpur	786157	27.4961	95.3842	2	20	5000	45
6	James Jon		37 Male	O-	No	Cat dander	125/82	170	75	1.44E+09	India	Assam	Tinsukia	Amguri	786160	27.4428	95.4104	1	24	6000	77
7	Olivia Davi		49 Female	A-	No	Penicillin	115/75	172	68	1.67E+09	India	Assam	Tinsukia	Arolmuria	786154	27.5029	95.2995	4	4	1000	4
8	William M		22 Male	B-	Anemia	No	135/88	168	72	1.89E+09	India	Assam	Tinsukia	Arunodaya	786160	27.5254	95.4398	2	7	1750	14
9	Ava Wilsor		29 Female	AB-	No	Pollen	122/81	163	69	2E+09	India	Assam	Tinsukia	Ashok Nag	786170	27.4886	95.2812	1	12	3000	35
10	Alexander		35 Male	O+	Hypothyro	No	128/84	176	76	1.22E+09	India	Assam	Sonitpur	Cherelia	784167	26.8698	92.9685	3	9	2250	22
11	Mia Martir		42 Female	A+	No	Dogs	118/79	171	74	1.33E+09	India	Assam	Sonitpur	Chitalmari	784110	26.9015	92.8539	5	46	11500	98
12	Daniel And		20 Male	B+	Diabetes	No	132/87	158	63	1.45E+09	India	Assam	Sonitpur	Da Besseri	784150	26.9204	92.8137	4	23	5750	58

Figure 3.1

3.2 Clustering Algorithms:

3.2.1 K-Means Clustering:

K-Means is a partition-based clustering algorithm that divides a dataset into KKK clusters, where each cluster is represented by the mean of its points, called the centroid. The algorithm iteratively adjusts the positions of the centroids to minimize the within-cluster variance (WCSS):

$$WCSS = \sum_{i=1}^{K} \sum_{x \epsilon C_i} ||x - \mu_i||^2$$

where C_i represents the i-th cluster, x is a data point, and μ_i is the centroid of cluster C_i. The algorithm continues until the centroids no longer change significantly, indicating that the data points have been effectively clustered.

3.2.2 Hierarchical Clustering:

Hierarchical Clustering builds a multilevel hierarchy of clusters, which can be visualized using a dendrogram. It operates in two modes: agglomerative (bottom-up) and divisive (top-down). In this study, we focus on agglomerative clustering, which starts with each data point as a single cluster and merges them iteratively based on a chosen linkage criterion (e.g., single linkage, complete linkage, average linkage). The goal is to minimize the dissimilarity between clusters being merged.

$$D(C_i, C_j) = \min_{\substack{x \in C_i \\ y \in C_i}} \| x - y \|$$

Here, $d(C_i, C_j)$ represents the distance between clusters C_i and C_j. The choice of linkage method significantly impacts the clustering results.

3.2.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

DBSCAN is a density-based clustering algorithm that groups together points that are closely packed together, marking points that lie alone in low-density regions as outliers. It requires two parameters: ϵ (the maximum radius of a neighborhood) and *minPts* (the minimum number of points in an ϵ - neighborhood). A point is considered a core point if it has at least *minPts* neighbors within a radius of ϵ. DBSCAN is effective for identifying clusters of arbitrary shapes and handling noise, but choosing optimal parameters is crucial for its performance.

4. Evaluation Metrics:

To assess the performance of the clustering algorithms, we use several well-established metrics:

4.1 Silhouette Score:

The Silhouette Score evaluates how well a data point fits within its assigned cluster relative to other clusters. This metric ranges from -1 to 1, with a higher value signifying that the data point is more appropriately clustered and better separated from neighboring clusters.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where *a(i)* is the average distance between point *i* and all other points in the same cluster, and *b(i)* is the smallest average distance between point *i* and points in a different cluster. A higher score indicates that data points are well matched to their own clusters and poorly matched to neighboring clusters.

4.2 Davies-Bouldin Index:

The Davies-Bouldin Index evaluates the average similarity between each cluster and its most similar neighboring cluster, aiming to measure how well-separated and distinct the clusters are. A lower index suggests better clustering.

$$DB = \frac{1}{k} \sum_{l=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

where σ_i is the average distance between each point in cluster i and the centroid c_i and $d(c_i, c_j)$ is the distance between the centroids of clusters i and j.

4.3 Calinski-Harabasz Index:

The Calinski-Harabasz Index, or the Variance Ratio Criterion, evaluates the ratio of the sum of between-cluster dispersion and within-cluster dispersion. A higher index indicates better-defined clusters.

$$\text{CH} = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \times \frac{n-K}{K-1}$$

where B_k is the between-cluster scatter matrix, W_k is the within-cluster scatter matrix, n is the number of data points, and K is the number of clusters.

5. Results:

The clustering algorithms were implemented using Python, and the results were evaluated based on the described metrics.

- **K-Means** showed the highest performance with a Silhouette Score of 0.3859, a Davies-Bouldin Index of 1.039, and a Calinski-Harabasz Index of 330.2689, indicating well-separated and cohesive clusters.
- **Hierarchical Clustering** resulted in a lower Silhouette Score of 0.3169, a higher Davies-Bouldin Index of 1.2688, and a Calinski-Harabasz Index of 268.6455, reflecting its sensitivity to outliers and cluster shapes.
- **DBSCAN** produced a moderate Silhouette Score of 0.3519 but a relatively high Davies-Bouldin Index of 1.5354, suggesting its effectiveness in identifying clusters of arbitrary shapes but difficulty in optimizing parameters for this dataset.

	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Means	0.385896	1.038988	330.2689
Hierarchical	0.316866	1.268789	268.6455
DBSCAN	0.017607	1.176017	62.41411

Figure 5.1

6. Discussion:

The results indicate that **K-Means** is the most effective algorithm for clustering blood donor data, particularly in scenarios where the data is well-separated and follows a roughly spherical distribution. Its performance can be attributed to its ability to minimize within-cluster variance, leading to more defined clusters. However, its reliance on the initial centroid placement and the need to specify the number of clusters K can be limitations in some cases.

Hierarchical Clustering provides valuable insights into the nested structure of data and is useful for understanding relationships at different levels of granularity. However, its computational complexity and sensitivity to noise make it less suitable for large, noisy datasets like those involving blood donors.

DBSCAN is effective for datasets with varying densities and noise, making it a good choice for complex, irregular data distributions. However, its sensitivity to parameter selection (ϵ and *minPts*) can result in suboptimal clustering if not carefully tuned.

7. Conclusion:

This study provides a comprehensive comparison of K-Means, Hierarchical Clustering, and DBSCAN for clustering blood donor data. The findings suggest that K-Means is the most suitable algorithm for this application, offering the most precise and meaningful clustering results. Future work could explore the integration of machine learning techniques to automate parameter selection in DBSCAN and investigate hybrid clustering approaches that combine the strengths of multiple algorithms.

By leveraging a web-based platform for real-time donor data updates, this study also highlights the importance of maintaining dynamic datasets to enhance the relevance and applicability of clustering models. These insights are crucial for optimizing healthcare logistics, improving resource allocation, and supporting other domains where effective data clustering is essential.

8. Future Research Directions:

Future research could explore the development of hybrid models that combine the strengths of different clustering algorithms. For example, starting with DBSCAN to identify core clusters and using K-Means to refine these clusters could offer a more robust solution. Additionally, employing advanced techniques like deep learning-based clustering methods on such datasets could further improve accuracy and applicability. Another promising direction is to explore adaptive clustering methods that can dynamically adjust parameters and cluster definitions based on real-time data inputs.

9. References:

1. Maryam Ashoori, and Zahra Taheri. “Using Clustering Methods for Identifying Blood Donors Behavior.” *ResearchGate*, Aug. 2013, https://www.researchgate.net/publication/256455245_Using_Clustering_Methods_for_Identifying_Blood_Donors_Behavior.
2. Shashikala B M, et al. “Machine Learning Clustering Method for Analysis of Blood Donor Deferral.” *ReasearchGate*, Sept. 2021, https://www.researchgate.net/publication/354689020_Machine_Learning_Clustering_Method_for_Analysis_of_Blood_Donor_Deferral.
3. Rashid Mehrabadi, E. and Pedram, M.M. 2010.” Blood Donors Classification and Identifying Future Donors”, The Fourth Iran Data Mining Conference, Sharif University of Technology, Tehran, Iran.