International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Analysis of Clustering Algorithms in Machine Learning for Healthcare Data

Rekha B H[1], Vaishak S Iyengar[2], Vinay V S[3]

Assistant Professor ,Department of IS&E, BIET, Davanagere,
Karnataka, India
rekhabh12@gmail.com
Department of Information Science and Engineering Davangere Karnataka India
Vaishak2303@gmail.com,svvinay10@gmail.com

ABSTRACT

One of the most widely used data analysis methods in machine learning is the clustering algorithm, which is used to accurately assess the enormous amounts of healthcare data from electronic health records, hospitals, clinics, body sensor networks, and internet of things devices. By dividing up the data of comparable patients according to their pertinent characteristics, clustering algorithms are always essential in the prediction of diseases. Numerous clustering methods have been created thus far for the analysis of various healthcare data sets.

I. INTRODUCTION

In the present day, the amount of healthcare data records generated on a daily basis is growing exponentially. A deluge of data has been generated by the proliferation of medical sensors, IoT devices, and digital medical records, which usually end up in various medical storage repositories. After that, a variety of activities, including analytical, process, and retrieval ones, are carried out to draw insightful conclusions from the unprocessed data. Physicians or other healthcare professionals will be able to make well-informed treatment decisions at the appropriate moment with the use of real-time warnings. Big data analytics technologies can therefore be utilised to increase treatment efficiency, save lives, give insight much more quickly, and ultimately save money.

Healthcare data is collected from a variety of sources, including hospitals, clinical settings, medical research, electronic records, and authorised websites. They are saved in a variety of formats, including text, video, audio, image, impala complicated kinds, and sequence files , making it challenging to process and analyse all of the data. One significant technique for addressing this analytic difficulty is to organise or cluster large amounts of health data into more compact formats. In such cases, clustering algorithms play an important role in analysing vast amounts of healthcare data as discrete segments in a distributed manner and efficiently aggregating all of these data across different clusters to acquire the final processed medical data. Several clustering algorithms have been created to analyse data, but it remains a difficult process to determine which technique delivers the best and ideal number.

II. LITERATURE SURVEY

Concept of cluster analysis:

Known as an autonomous learning method, clustering is one of the most popular algorithms in the machine learning field [2]. When there are no class labels available to process the datasets, are the clustering of technique is significant because it breaks the enormous volume of data into smaller groupings of data. Every cluster comprises a collection of data points, with the primary purpose of the clustering method being to categorise and organise each data point into a certain cluster. Additionally, data points in the same cluster ought to possess comparable features and/or qualities, but data points in other clusters should have extremely different features and/or properties [1]. Several clustering methods, including K-means, K-Medoids, or Partitioning Around Medoids (PAM), and Hierarchical, have been introduced in the research works that have already been completed [1–2] for the analysis of healthcare data sets.

K-means Clustering Algorithm:

K-means is a straightforward and widely applicable clustering algorithm that is mostly utilised for classifying an unlabeled dataset. Finding comparable clusters, denoted by variable k, is the primary goal of this algorithm. The mean or centroid is a statistic that this algorithm utilises to describe the cluster for this purpose. A centroid, which may or may not be a member of the dataset, is a data point that represents the cluster's centre. Thus, it separates the n data points into k clusters, and then assigns each data point to the cluster that has the closest possible centroid. The Euclidean distance is then precisely computed from each data point to the centroid in an equilateral cluster. The centroid of a cluster is always assigned data points based on their lowest euclidean distance from it.

K-medoids Clustering Algorithm:

The algorithm K-medoid, commonly known as Partition Around Medoids (PAM), is a variant type of algorithm. Within a cluster that is centred and has a small dispersion over all of the data points in the cluster, a data point functions as a medoid in this algorithm. As a result, this medoid can serve as a cluster representative for other data points. The primary principle of PAM is to organise the collection of medoids, calculate important data points as a medoid in a particular cluster, and then assign each data point to the closest medoid in that cluster. Also, there are typically two stages to this algorithm: the construction phase and the swap phase. The first phase's job is to identify the data point with the lowest mean dissimilarity throughout the entire dataset, or the first medoid.

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. SUMMARY OF LITERATURE SURVEY

Hierarchical Clustering:

A unique kind of unsupervised machine learning algorithm known as Hierarchical Cluster Analysis (HCA) is called hierarchical. Using a tree-based framework, hierarchical cluster analysis aims to group comparable unlabeled data points into a number of clusters. The data points at the base of the tree create a collection of clusters, each of which is unique from the others. Additionally, the data points in a given cluster are typically the same as those in other clusters within the dataset. Based on the hierarchy, this technique generates a dendrogram, or tree-type structure. Agglomerative hierarchical clustering, also known as AGNES (Agglomerative Nesting), and Divisive hierarchical clustering, also known as DIANA (Divisive Analysis), are the two main categories of hierarchical clustering algorithms. Each of these algorithms is precisely the opposite of the other. Table displays the summary of different algorithms according to different characteristics.

Table 1. Summary of clustering algorithms

Algorithm	Big data			Computation speed	Modifications corrections	Cluster shape	Results interpretation
	Size of data set	Type of data	Complexity				
K-means	Large	Numerical	O(nkd)	Fast	Flexible	Non convex	Easy
K-medoids	Small	Categorical	$O(n^2dt)$	Moderate	Difficult	Non convex	Difficult
Hierarchical	Large	Numerical	O(n)	Slow	Flexible	Non convex	Easy

Validation Measures:

Various internal and stability validation indicators are used to assess the performance of unsupervised learning algorithms. Assessing the proper number of clusters and determining the quality of the suitable clustering technique depend heavily on the internal metrics. Without utilising any external data, this measure computes a cluster's quality only based on its internal data. Three categories comprise the fundamental internal validation measurements [2]: Silhouette, Dunn index, and Connectivity. The internal validation indices for a physiological data set are briefly presented in this section.

1.Internal Measures

Connectivity: This measure represents the total number of rows *n* (data points or observations) and columns *m* in a dataset. The values are always considered as numeric (e.g., a physiological parameter's values). Let $Y_{ni}(j)$ and $x_iY_{ni}(j)$ be the j^{th} nearest neighbor of data point *i* and zero, respectively, if both *i* and *j* are in the same cluster, and then ^1_j otherwise. The connectivity is measured for a particular cluster C = {C1,C2 ....C*k*} with *n* data points using the below equation

$$\mathcal{C} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_i Y_{ni}(j) \quad (1)$$

Where *p* represents a parameter value and if the connectivity measure has a value between 0 and ∞, it should always be decreased.

Dunn Index: This is an important metric that presents the ratio of the lowest distance between the data points which is not available in the same cluster and the highest distance in the intra-cluster. The index value can be obtained as

$$D(4) = \min_{\mathcal{C}_k, \mathcal{C}_l \in \mathcal{C},} \frac{\left(\min_{i \in \mathcal{C}_k, j \in \mathcal{C}_l} dist(i,j) \right)}{CC_k = C_l} \max_{\mathcal{C}_m \in \mathcal{C}} d(\mathcal{C}_m)$$

Where *d*(C*m*) indicates a cluster C*m* with maximum distance and this index has a value between 0 and ∞, and it should always be increased.

Average Distance (AD):

By taking into account the two aforementioned scenarios, the AD measure's primary purpose is to forecast the average distance between data points that are grouped together. The lesser values are always taken into account while evaluating the findings if the AD is between zero and ∞. To compute AD, use the supplied expression below.

$$AD(K) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \frac{1}{n\left(\mathcal{C}^{i,l} \cap \mathcal{C}^{i,0}\right)} \left[\sum_{i \in \mathcal{C}^{i,0}, j \in \mathcal{C}^{i,l}} dist(i,j) \right]$$

Average Distance Between Means (AM): This measure's primary goal is to determine the average distance between data points under the two previously mentioned scenarios that are displayed in the same cluster. Only the Euclidean distance, which has smaller values between 0 and ∞, is used, nevertheless, and is always favoured. Let x i,0 represent the cluster containing the average data point I , and x I , l the cluster containing the data point C i, without the removal of column l C. Then, the following formula is used to compute.

$$ADM(K) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \frac{1}{n\left(\mathcal{C}^{i,l} \cap \mathcal{C}^{i,0}\right)} dist\left(\bar{x}_{\mathcal{C}^{i,l}}, \bar{x}_{\mathcal{C}^{i,0}}\right)$$

Figure of Merit (FOM): A factor of multiplicity (FOM) plays a crucial role in grouping data by estimating the average variance of the deleted columns in each of the several clusters. It also uses the average number of clusters to compute the mean error rate, with smaller values between 0 and ∞ being favoured. Next, FOM utilises the provided formula to anticipate a certain left-out column l.

$$FOM(l, K) = \sqrt{\frac{1}{N}\sum_{k=1}^{k}\sum_{i\in C^{k,l}} dist\left(x_{i,l}, \bar{x}_{C_k(l)}\right)}$$

V. METHODOLOGY USED IN EXISTING SYSTEM

By using two packages that are defined in the R programming tool, the clustering techniques are verified. The Nb Clust package [3] and the c Valid package [2] are the two main packages utilised in this investigation. For a given data collection, both programmes are essential for figuring out the ideal number of data clusters and validating the useful outcomes of the clustering study. Euclidean distance is a parameter in the NbClust function used in this analytical investigation. The range of important parameters, as illustrated in Fig. 1, is used to measure the frequency of occurrence of time-critical.

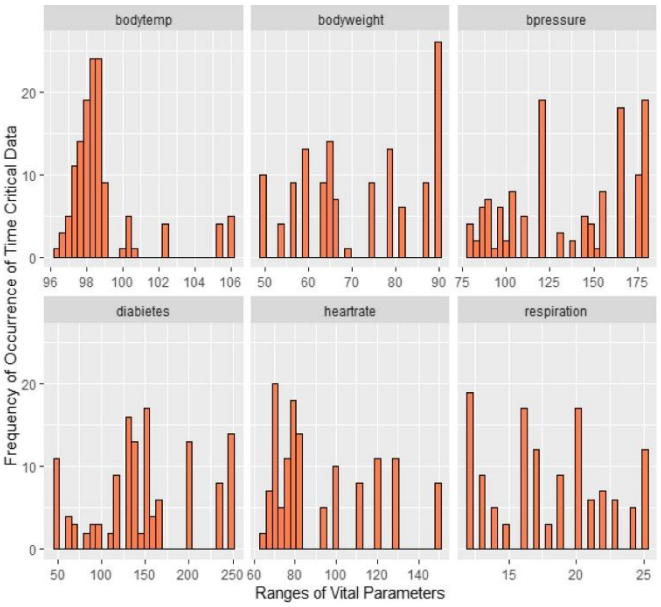
Set of data: The core data set for this experiment research is the statlog heartrate real-world data set, which consists of 3 variables and 130 instances from the UCI machine learning repository. Using the same 130 examples, this paper adds 5 more variables to validate the benefits of the synthetic dataset. There are only numeric values with various properties in the data set. Depending on the patient's conditions, the vital ranges of each attribute—normal, moderate, and extremely high—are included. Table 2 lists the various attributes of synthetic and real-world healthcare data sets.

Table 2. Various characteristics of healthcare data sets

Name of data set	Type of dataset	Type of data	No of instances	No of attributes
Heart rate	Real world	Multivariate	130	3
Physiological data	Synthetic data	Numerical	130	8

Comparative Analysis: The comparative analysis's goals are to determine the degree to which each algorithm can correctly classify similar health records from the mixed physiological data set, as well as to determine the ideal number of cluster sizes for each algorithm and identify the best performing algorithm. The analysis results of three distinct clustering algorithms are validated through the use of both internal and stability measures.

Evaluating Validity: Table 3 displays the analytical findings of different clustering techniques based on the internal validity metrics. Algorithms are first tested with cluster sizes ranging from k = 2 to k = 10.



The K-means algorithm with two clusters outperforms the hierarchical clustering algorithm in terms of connectivity and Dunn index measures, but the latter outperforms the former in terms of silhouette validity. As a result, it is the second best-known clustering algorithm in terms of internal validity. Additionally, the comparative analysis revealed that the K-medoids produce no clustering.

VI. COMPARISION

The ideal number of clusters and their scores are evaluated using two key metrics: internal and stability. The best scores for each algorithm are shown in Table 4. Based on the observations,Dunn index, and silhouette. In contrast, among all the clustering techniques investigated, the hierarchical approach with various clusters frequently produces the highest stability values for APN, ADM, and FOM, with the exception of AD.

Table 3. Internal validation of clustering algorithms

Type of measures	Clustering method	Validity measures	Cluster size			
			2	3	4	5
Internal validation metrics	Hierarchical	Connectivity	7.5556	10.4845	10.7845	14.5425
		Dunn	0.2943	0.2971	0.3140	0.3140
		Silhouette	0.3075	0.2093	0.2390	0.2261
	K-means	Connectivity	2.1940	41.1071	25.3369	35.9258
		Dunn	0.3450	0.1761	0.2356	0.1950
		Silhouette	0.2470	0.1743	0.2496	0.2345
	Pam	Connectivity	19.5016	45.1821	67.7214	67.7167
		Dunn	0.0763	0.0508	0.0330	0.0429
		Silhouette	0.2147	0.2094	0.1867	0.2152

However, the best appropriate cluster size for a physiological data set is 2, and it has been demonstrated that it is suitable for dealing with high-dimensional physiological datasets. Finally, this investigation revealed that the Pam algorithm did not yield the ideal number of clusters on a synthetic data set with high problem dimensionality, as shown in Table 4.

Table 4. Stability validation of clustering algorithms

Type of measures	Clustering method	Validity measures	Maximum cluster size			
			2	3	4	5
Stability validation metrics	Hierarchical	APN	0.0648	0.3074	0.0389	0.1035
		AD	3.5190	3.5114	3.0791	2.9625
		ADM	0.2221	0.9328	0.4820	0.4237
		FOM	0.9695	0.9520	0.9304	0.8599
	K-means	APN	0.1824	0.3675	0.1987	0.1940
		AD	3.6036	3.4353	2.9575	2.7903
		ADM	1.2480	1.2474	0.7353	0.7205
		FOM	0.9779	0.9509	0.9023	0.8712
	Pam	APN	0.1932	0.2424	0.3472	0.3391
		AD	3.4350	3.2164	3.1585	2.9340
		ADM	0.6666	0.8308	1.1647	1.0535
		FOM	0.9535	0.9196	0.9141	0.8984

VII. CONCLUSION

This work provides a complete theoretical overview of clustering algorithms, specifically for healthcare data processing, from both theoretical and experimental viewpoints. Numerous clustering methods have been proposed in existing studies for analysing healthcare data sets and validated using various metrics. However, it is extremely difficult to predict which clustering technique will be most appropriate for a specific data collection, as well as the optimal number of clusters from a given set. Based on these perspectives, this study examined various clustering algorithms from a therapeutic perspective and evaluated them using internal and stability metrics. The observed results provided a better solution for developing unique clustering algorithms and recommending a specific technique for a large amount of physiological data.

VIII. REFERENCES

1. Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S.: Big data in healthcare: management,analysis and future prospects. J. Big Data **6**(54), 1–25 (2019). https://doi.org/10.1186/s40 537-019-0217-0
2. Thasni, K.M., Haroon, R.P.: Application of big data in health care with patient monitoring andfuture health prediction. In: Smys, S., Senjyu, T., Lafata, P. (eds.) ICCNCT 2019. LNDECT, vol. 44, pp. 49–59. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37051-0_6
3. Dautov, R., Distefano, S., Buyya, R.: Hierarchical data fusion for smart healthcare. J. BigData **6**(19), 1–23 (2019). https://doi.org/10.1186/s40537-019-0183-6
4. Prosperi, M., Min, J.S., Bian, J., Modave, F.: Big data hurdles in precision medicine andprecision public health. BMC Med. Inform. Decis. Mak. **18**(139), 1–15 (2018)
5. Zillner, S., Neururer, S.: Big data in the health sector. In: Cavanillas, J.M., Curry, E., Wahlster,W. (eds.) New Horizons for a Data-Driven Economy, pp. 179–194. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-21569-3_10
6. Ambigavathi, M., Sridharan, D.: A survey on big data in healthcare applications. In: Choudhury, S., Mishra, R., Mishra, R.G., Kumar, A. (eds.) Intelligent Communication, Control and Devices. AISC, vol. 989, pp. 755–763. Springer, Singapore (2020). https://doi.org/10.1007/ 978-981-13-8618-3_77
7. Ambigavathi, M., Sridharan, D.: Big data analytics in healthcare. In: 2018 Tenth InternationalConference on Advanced Computing (ICoAC), India, pp. 269–276. IEEE (2018)