Analysis of Clustering Algorithms in Machine Learning for Healthcare Data

Chapter · July 2020

DOI: 10.1007/978-981-15-6634-9_12

2 authors:

Ambigavathi Munusamy
Anna University, Chennai
19 PUBLICATIONS **296** CITATIONS
SEE PROFILE

D. Sridharan
Anna University, Chennai
79 PUBLICATIONS **965** CITATIONS
SEE PROFILE

Analysis of Clustering Algorithms in Machine Learning for Healthcare Data

M. Ambigavathi(✉) and D. Sridharan

Anna University, CEG Campus, Chennai, India
ambigaindhu8@gmail.com

Abstract. Clustering algorithm is one of the most popular data analysis technique in machine learning to precisely evaluate the vast number of healthcare data from the body sensor networks, internet of things devices, hospitals, clinical, medical data repositories, and electronic health records etc. The clustering algorithms always play a crucial role to predict the diseases by partitioning the similar patient's data based on their relevant attributes. The vast number of clustering algorithms have been developed for analyzing several healthcare data sets so far. However, the algorithms presented in the literature may achieve a better result with a particular type of data set but may fail or provide poor results with the data set of other types. Many of the research studies considered specific or multiple data sets for clustering analysis. But there are only a few studies used mixed type of data for analyzing and verifying the optimal number of clusters. To alleviate these issues, this paper aims to inspect various clustering algorithms from the theoretical and experimental perspectives. The experimental results elucidate the best algorithm from each categories using a physiological data set. The efficiency of each clustering algorithm in machine learning is validated using a number of internal as well as stability measures. Finally, this paper highlights the future directions with a proper clustering algorithm for handling high dimensional healthcare data sets.

Keywords: Machine learning · Clustering algorithms · Unsupervised learning algorithms · Big data · Healthcare applications

1 Introduction

The numerous records of healthcare data generated every day are increasing astronomically in today's modern era [1]. The explosion of medical sensors, internet of things devices, and digitalization of medical records have created a flood of data typically landing in different medical storage repositories. Then, various kinds of operations such as analytical, process, and retrieval are performed to extract valuable insights from the raw data [2]. With the help of real-time alerts, doctors or medical practitioners will take perspective decisions about treatment at the right time [3, 4]. Therefore, big data analytics solutions can be used to save human lives, provide analysis much faster, ultimately save money and improve the efficiency of treatment [5].

M. Singh et al. (Eds.): ICACDS 2020, CCIS 1244, pp. 117–128, 2020.
https://doi.org/10.1007/978-981-15-6634-9_12

The healthcare data is captured from various sources that include [6] hospitals, clinical, medical research, electronic records, and authorized websites respectively. They are stored in different formats such as text, video, audio, image, impala complex types, and sequence file respectively [7] and also make it very difficult to process and analyze all pieces of data effectively. One key strategy to solve this analytic issue is to group or cluster the big health data in a more compact format. In such a case, clustering algorithms contribute a major role to analyze the massive volume of healthcare data as small segments in a dispersed way and effectively aggregate all these data across different clusters to obtain the final processed medical data [8]. There are several clustering algorithms developed [9] to analyze the data but still, it is a challenging task which algorithm provides the best and the optimal number of clusters with respect to different data sets. Many authors have evaluated the clustering algorithms using different medical data sets with unique validation metrics [10–13]. Only a few authors [14] have been used synthetic data sets with real-time data sets to assess the variations and performance of three distinct clustering algorithms. Each data set is unique in its own way. No studies have been considered so far to estimate various clustering algorithms using the mixed type of physiological data. This type of analysis on vital parameters must require in the near future to identify the time-critical data than normal data. Therefore, this work considers only a synthetic data set instead of real-time data sets to evaluate the best number of clusters for healthcare data analysis. Moreover, the value of the raw healthcare data collected from hospitals or patients in real-time may be similar or slightly different from our synthetic data set. But the minimum and maximum values of vital data may only deviate from the considered ranges.

Despite the vast number of analysis for clustering algorithms using various healthcare data sets including heart rate [15], brain [16], body temperature [17], emotions [18], cancer [19], blood pressure, ambulatory, and emergency respectively, available in the literature. In such a case, it is very difficult for handlers to decide in advance which algorithm is most suitable one for identifying the abnormality in a given big health dataset. There are still many limitations exist in the literature that need to be addressed: (i) the unique attributes of various clustering algorithms especially for physiological data set are not analysed carefully, (ii) several clustering algorithms have been developed for healthcare domain but they were not deliberated any mixed type of vital information and (iii) only experimental analysis has been carried out to specific healthcare data set to study the significance of one algorithm over another. The aforementioned reasons are highly motivated us to inspect various clustering algorithms, especially for the mixed type physiological data set. The main contributions are outlined as follows:

- To study three distinct types of clustering algorithms based on the theoretical perspectives.
- To validate the different clustering algorithms using internal and stability metrics.
- To analyze the most optimal clustering algorithm with respect to clinical perspectives.

Therefore, this article provides readers with a sufficient analysis of particular clustering algorithms by theoretically and experimentally comparing them on the synthetic physiological data set. Other sections of this paper are described as follows: The theoretical details of clustering algorithms are summarized in Sect. 2. Section 3 describes the

internal and stability validation measures for various clustering algorithms. The experimental and comparative analysis of different clustering algorithms are explained in Sect. 4. Finally, Sect. 5 concludes the paper with appropriate clustering algorithm with future scope.

2 Analysis of Clustering Algorithms

Clustering is one of the best known algorithm in machine learning domain, named as an unsupervised learning algorithm [20]. The significance of clustering algorithm is to divide the large volume of data into smaller groups of data when there is no class labels available to process the datasets. Each cluster contains a set of data points where clustering algorithm mainly used to classify and group each data point into a particular cluster. Besides, the data points within the same cluster should have similar properties, while data points in the different cluster should have highly dissimilar properties and/or features [21]. Many clustering algorithms for analyzing healthcare data sets have been introduced in the existing research works [22–26]: K-means, K-Medoids or Partitioning Around Medoids (PAM), and Hierarchical. The main procedures of these algorithms are classified as follows.

2.1 K-means Clustering Algorithm

K-means is a simple and most general clustering algorithms which is mainly used to classify the given dataset that is unlabeled. This algorithm mainly aims to find similar clusters represented by variable k. For this purpose, this algorithm uses the mean or centroid as a metric to characterize the cluster. A centroid is a data point that indicates the center of the cluster, and it might not necessarily be a member of the dataset. So, it divides n data points into k number of clusters and then each data point n belongs to appropriate cluster with the nearest possible centroid. Next, the Euclidean distance is accurately calculated from each data point n to the centroid in a given cluster. Always, the data points in a cluster are assigned to the centroid depending on the minimum euclidean distance from that centroid point. When there no data point is available to assign, an early grouping is considered. In such case, ‘c’ new centroids are re-calculated, thus new iteration continues until the ‘c’ centroids stop changing their position.

2.2 K-medoids Clustering Algorithm

K-medoid is a variant type of algorithm which is also termed as Partition Around Medoids (PAM). In this algorithm, data point act as a medoid within a cluster that are centrally located whose disparity over all data points in the cluster is minimal. Therefore, this medoid can be used as a representative of other data points within a cluster. The main core idea of PAM is to first calculate major data point as a medoid in a specific cluster, group the set of medoids, and then each data point is assigned to the nearest medoid in a given cluster. Moreover, this algorithm generally follows two phases: build and swap phase. The role of the first phase is to select the first medoid as the data point with the lowest mean dissimilarity with respect to the whole dataset. Likewise, in the second

phase, given the current set of 'k' medoids, all the neighbor data points are evaluated. A new medoid is created by exchanging data points in the old medoid with the data points in a new non-medoid.

2.3 Hierarchical Clustering

Hierarchical is a special type of unsupervised machine learning algorithm, also referred as Hierarchical Cluster Analysis (HCA). The goal of hierarchical cluster analysis is to cluster similar unlabeled data points into number of clusters using tree based structure. The data points in the end of tree forms a set of clusters, where each and every cluster is distinct from other clusters. Besides, the data points within a specific cluster is mostly identical to other clusters in the data set. This algorithm uses a tree-type structure (dendrogram) based on the hierarchy. Basically, there are two types of hierarchical clustering algorithms include Agglomerative hierarchical clustering or AGNES (Agglomerative Nesting) and Divisive hierarchical clustering or DIANA (Divisive Analysis). Both this algorithm is exactly the reverse of each other. The summary of various algorithms with respect to various characteristics are listed in Table 1.

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

3 Validation Measures

The performance of unsupervised learning algorithms is evaluated using different internal, and stability validation metrics. The internal measures are very important for evaluating the right number of clusters and computing the quality of the appropriate clustering algorithm. This measures consider only the internal information to calculate the quality of a clusters without using any external information. The basic internal validation measurements [27] are classified into three types: Connectivity, Silhouette and Dunn index. This section briefly presents the internal validation indices used for a physiological data set.

3.1 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 $x_i Y_{ni}(j)$ 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 $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2 \ldots . \mathcal{C}_k\}$ with n data points using the below equation

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

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

Silhouette Coefficient. This coefficient is a very useful metric for evaluating the performance of clustering results. This value measures how data points are grouped and computes the average distance available between the different clusters. The width of this coefficient always lies in the following interval $[-1, 1]$ that implies the super grouped data points with values near to 1 and lower grouped data points with values near to -1. Therefore, the coefficient for data point i is defined as

$$S(i) = \frac{(y_i - x_i)}{max(y_i, x_i)} \tag{2}$$

Where x_i and y_i denote the average distance between the data points in the same cluster and the average distance between the data points in the nearest neighboring clusters which can be expressed as

$$y_i = \min_{\mathcal{C}_k \in \frac{\mathcal{C}}{\mathcal{C}_i}} \sum_{j \in \mathcal{C}_k} \frac{dist(i,j)}{n\mathcal{C}_k} \tag{3}$$

Where $\mathcal{C}_i$ indicate a cluster with data point i, $dist(i, j)$ presents the distance between the data points i and j, then $n\mathcal{C}_k$ implies cardinality of the cluster $\mathcal{C}$.

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

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

Where $d(\mathcal{C}_m)$ indicates a cluster $\mathcal{C}_m$ with maximum distance and this index has a value between 0 and ∞, and it should always be increased.

3.2 Stability Measures

The stability measure is a special type of validation measure to individually evaluate the cluster results from the overall analysis by removing each column in the data set. This type of measure is very significant especially when the physiological raw data are highly correlated with others. For this purpose, this study uses stability measures to compare the consistency of raw data in the medical synthetic data set. Generally, the stability measures [28] are broadly classified into four different groups: (i) Average Proportion of Non-overlap (APN), (ii) Average Distance (AD), (iii) Average Distance between Means (ADM), and (iv) Figure of Merit (FOM).

Average Proportion of Non-overlap (APN). This measure is used to calculate the average proportion of data point that is not located in the same cluster with a particular or single column removed. Let consider $\mathcal{C}^{i,0}$ be the cluster with data point i using the original cluster and $\mathcal{C}^{i,l}$ be the cluster with column l removed in the data set. Then, APN value is always varied between the following interval [0, 1]. If the APN values close to 0 that indicates the highly consistent results. For the total number of cluster set K, the APN value is measured using given formula

$$APN(K) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \left(1 - \frac{n\left(\mathcal{C}^{i,l} \cap \mathcal{C}^{i,0}\right)}{n\left(\mathcal{C}^{i,0}\right)} \right) \tag{5}$$

Average Distance (AD). The main function of AD measure is to predict the average distance between the data points that are placed in the same cluster by considering the aforementioned two cases. If the AD has a value between zero and ∞, and then the smaller values are always considered to evaluate the results. The following given expression is used to compute AD,

$$AD(K) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \frac{1}{n\left(\mathcal{C}^{i,l} \cap \mathcal{C}^{i,0}\right)} \left[\sum_{i \in \mathcal{C}^{i,0}, j \in \mathcal{C}^{i,l}} dist(i,j) \right] \tag{6}$$

Average Distance Between Means (AM). The main objective of this measure is to calculate the average distance between data points that are presented in the same cluster under the aforementioned two cases. However, only it uses the Euclidean distance with smaller values between 0 and ∞ is always preferred. Let $\bar{x}_{\mathcal{C}^{i,0}}$ denote cluster contains average data points i and $\bar{x}_{\mathcal{C}^{i,l}}$ indicate the cluster contains data point i with column l removed. Then, it is computed using the below formula,

$$ADM(K) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{l=1}^{M} \frac{1}{n\left(\mathcal{C}^{i,l} \cap \mathcal{C}^{i,0}\right)} dist\left(\bar{x}_{\mathcal{C}^{i,l}}, \bar{x}_{\mathcal{C}^{i,0}}\right) \tag{7}$$

Figure of Merit (FOM). The decisive role of a FOM is to estimate the average variance of the deleted columns in different clusters and grouping is performed based on the remaining (undeleted) columns. The smaller values between 0 and ∞ are mostly preferred and also it computes the mean error rate using average number of clusters. Then, FOM predicts a particular left-out column l using the given formula

$$FOM(l,K) = \sqrt{\frac{1}{N} \sum_{k=1}^{k} \sum_{i \in \mathcal{C}^{k,l}} dist\left(x_{i,l}, \bar{x}_{\mathcal{C}_k(l)}\right)} \tag{8}$$

Where $x_{i,l}$ presents the value of i^{th} observation in the l^{th} column and $\overline{x}_{C_k(l)}$ denote the average of a cluster. Generally, FOM uses only Euclidean distance and also it is multiplied by the following adjustment factor $\sqrt{\frac{N}{N-K}}$, to decrease the amount of cluster expansions.

4 Experimental Results

The clustering algorithms are validated by including two packages defined in R programming tool. The two major packages used in this study are clValid [29] package and NbClust package [30], respectively. Both packages are very significant to determine the best optimal number of data clusters for a given data set and validate the effective results from the clustering analysis. This analysis study uses Euclidean distance as a parameter in NbClust function. The frequency of occurrence of time-critical data is measured with respect to the range of vital parameters, which are shown in Fig. 1.

4.1 Data Set

This experiment study uses statlog heartrate real-world data set (i.e., UCI machine learning repository) as a basic data set, which consists of 130 instances and 3 variables. To validate the advantages of the synthetic dataset, this work includes 5 additional variables by utilizing the same 130 instances. The data set contains only numerical values with different attributes. The vital ranges of each attribute are incorporated based on the conditions of the patient such as normal, moderate and extremely high. The various characteristics of both real world and synthetic healthcare data sets are mentioned in Table 2.

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

4.2 Comparative Analysis

The aim of comparative analysis is to choose how accurately each and every algorithm can able to group similar health records from the mixed physiological data set. Further, to analyze the optimal number of the cluster's size for every algorithm and predict which algorithm performs better than others. The analysis results of three different clustering algorithms are validated using both internal and stability measures.

Evaluating Validity. The analysis results of various clustering algorithms based on the internal validity measures are presented in Table 3. Initially, algorithms are validated with the varying cluster size from k = 2 to k = 10.

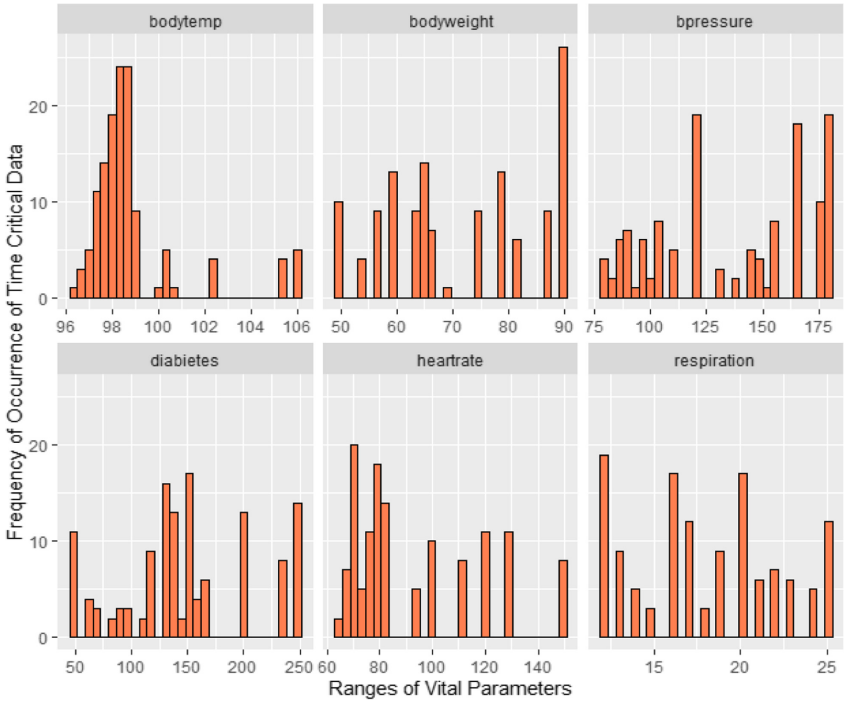

Fig. 1. Frequency of occurrence of vital data

From the cluster analysis, it is observed that the K-means algorithm with two clusters provides better results using connectivity and Dunn index measures as compared to the hierarchical clustering algorithm. However, the hierarchical algorithm achieves better output according to the silhouette validity measure. Therefore, it is the second best known clustering algorithm in terms of internal validity. Moreover, the comparative analysis suggested that the K-medoids yield no clustering results in comparison to K-means and hierarchical algorithms.

Evaluating Stability. The stability of three different clustering algorithms is validated to predict any variations in the clustering outputs based on the removal of one column in a given data set. The achieved results of stability for each clustering algorithm are displayed in Table 3. From the assessments, it is noticed that the hierarchical algorithm almost approaches the lower stability values based on the APN, ADM, and FOM respectively. Though it achieved better stability values with all three measures it is failed to provide the best result for AD measure. Further, the maximum stability value of K-means algorithm indicates that the algorithm is not able to yield better values. Likewise, the Pam algorithm is ineffective to give stable outputs in terms of stability measures. Hence the hierarchical clustering algorithm contributes the highest stability results in every aspects as compared with K-means and Pam algorithms.

Evaluating Optimal Scores. The optimal number of clusters and their scores are evaluated using two important measures such as internal and stability. The best optimal

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

scores of every algorithms are depicted in Table 4. Based on the observations, it is clearly shown that the K-means algorithm with two optimal clusters can provide the best results in terms of connectivity, Dunn index and silhouette, respectively. In contrast, the hierarchical algorithm with different clusters often yields the highest stability values for APN, ADM, and FOM except for AD among all considered clustering algorithms.

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

Table 5. Optimal scores for various clustering algorithms

Type of validity measure	Name of validity metric	Optimal score	Clustering method	Optimal number of clusters
Internal	Connectivity	2.1940	k-means	2
	Dunn	0.3450	k-means	2
	Silhouette	0.3075	Hierarchical	2
Stability	APN	0.0389	Hierarchical	4
	AD	2.7903	K-means	5
	ADM	0.2221	Hierarchical	2
	FOM	0.8599	Hierarchical	5

However, the best optimal cluster size for a physiological data set is 2 and also it is significantly confirmed that the suitability for dealing with high-dimensional physiological datasets. Finally, this analysis suggested that the Pam algorithm failed to produce the optimal number of clusters on synthetic data set with high problem dimensionality, as mentioned in Table 5.

5 Conclusion and Future Work

This study provided a detailed theoretical view on clustering algorithms especially for healthcare data analysis from both theoretical and experimental perspectives. There are numerous clustering algorithms deliberated in the existing studies for analyzing healthcare data sets and also validated with different metrics. However, it is very hard to decide in advance which clustering algorithm would be the most suitable for a particular data set and what would be the best optimal number of clusters from a given a set. Based on these perceptions, this study analysed various clustering algorithms in clinical point of view and validated using internal and stability measures. The observed results reported a better solution to develop novel clustering algorithm and to recommend a specific algorithm for huge volume of physiological data set. The grouping of abnormal variations from different columns of data sets is the most significant requirement rather than grouping the normal variations when using the mixed or complicated vital data sets. In future, this study will further extend the analysis for big pandemic healthcare data sets with respect to similarity score, condition-specific, and then generic preference-based measures.

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/s40537-019-0217-0

2. Thasni, K.M., Haroon, R.P.: Application of big data in health care with patient monitoring and future 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. Big Data **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 and precision 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 International Conference on Advanced Computing (ICoAC), India, pp. 269–276. IEEE (2018)
8. Van Hieu, D., Meesad, P.: Fast K-means clustering for very large datasets based on MapReduce combined with a new cutting method. In: Nguyen, V.-H., Le, A.-C., Huynh, V.-N. (eds.) Knowledge and Systems Engineering. AISC, vol. 326, pp. 287–298. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-11680-8_23
9. Fahad, A., et al.: A survey of clustering algorithms for big data: taxonomy and empirical analysis. IEEE Trans. Emerg. Top. Comput. **2**(3), 267–279 (2014)
10. Hatamlou, A.: Heart: a novel optimization algorithm for cluster analysis. Prog. Artif. Intell. **2**(3), 167–173 (2014). https://doi.org/10.1007/s13748-014-0046-5
11. Khalid, S., Prieto-Alhambra, D.: Machine learning for feature selection and cluster analysis in drug utilization research. Curr. Epidemiol. Rep. **6**, 364–372 (2019). https://doi.org/10.1007/s40471-019-00211-7
12. Zhao, W., Zou, W., Chen, J.J.: Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinform. **15**, 1–11 (2014)
13. Wei, P., He, F., Li, L., Shang, C., Li, J.: Research on large data set clustering method based on MapReduce. Neural Comput. Appl. **32**, 93–99 (2020). https://doi.org/10.1007/s00521-018-3780-y
14. Patil, C., Baidari, I.: Estimating the optimal number of clusters k in a dataset using data depth. Data Sci. Eng. **4**, 132–140 (2019). https://doi.org/10.1007/s41019-019-0091-y
15. Asril, H., Mousannif, H., Al Moatassime, H.: Reality mining and predictive analytics for building smart applications. J. Big Data **6**(66), 1–25 (2019). https://doi.org/10.1186/s40537-019-0227-y
16. Durieux, J., Wilderjans, T.F.: Partitioning subjects based on high-dimensional fMRI data: comparison of several clustering methods and studying the influence of ICA data reduction in big data. Behaviormetrika **46**, 271–311 (2019). https://doi.org/10.1007/s41237-019-00086-4
17. Obermeyer, Z., Samra, J.K., Mullainathan, S.: Individual differences in normal body temperature: longitudinal big data analysis of patient records. Bio Med. J. **359**, 1–9 (2017)
18. Sharma, K., Castellini, C., van den Broek, E.L., Albu-Schaeffer, A., Schwenker, F.: A dataset of continuous affect annotations and physiological signals for emotion analysis. Nat. Sci. Data **6**(196), 1–13 (2019)
19. Papachristou, N., Miaskowski, C., Barnaghi, P., Maguire, R., Farajidavar, N.: Comparing machine learning clustering with latent class analysis on cancer symptoms' data. In: IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), UK, pp. 1–5. IEEE (2016)

20. Nerurkara, P., Shirkeb, A., Chandanec, M., Bhirudd, S.: Empirical analysis of data clustering algorithms. In: 6th International Conference on Smart Computing and Communications, ICSCC 2017, India, pp. 770–779. Elsevier (2018)
21. Tambe, S.B., Gajre, S.S.: Cluster-based real-time analysis of mobile healthcare application for prediction of physiological data. J. Ambient Intell. Hum. Comput. **9**(429), 1–17 (2017)
22. Praveen Kumar, D., Amgoth, T., Annavarapu, C.S.R.: Machine learning algorithms for wireless sensor networks: a survey. Inf. Fusion **49**, 1–25 (2019)
23. Rokach, L.: A survey of clustering algorithms. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-09823-4_14
24. Pérez-Suárez, A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: A review of conceptual clustering algorithms. Artif. Intell. Rev. **52**(2), 1267–1296 (2018). https://doi.org/10.1007/s10462-018-9627-1
25. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. Ann. Data Sci. **2**(2), 165–193 (2015). https://doi.org/10.1007/s40745-015-0040-1
26. Barbakh, W.A., Wu, Y., Fyfe, C.: Review of clustering algorithms. In: Non-standard Parameter Adaptation for Exploratory Data Analysis. Studies in Computational Intelligence, vol. 249. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04005-4_2
27. Palacio-Nino, J.-F., Berzal, F.: Evaluation metrics for unsupervised learning algorithms **1**, 1–9 (2019)
28. von Luxburg, U.: Clustering Stability: An Overview, pp. 1–41. Now Publishers Inc., Hanover (2010)
29. Brock, G., Pihur, V., Datta, S., Datta, S.: clValid: an R package for cluster validation. J. Stat. Softw. **25**(4), 1–22 (2008)
30. Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust: an R package for determining the relevant number of clusters in a data set. J. Stat. Softw. **61**(6), 1–36 (2014)