Data Clustering: Application and Trends

Gbeminiyi John Oyewole[a]
Department of Engineering and Technology Management,
University of Pretoria, South Africa

George Alex Thopil
Department of Engineering and Technology Management,
University of Pretoria, South Africa

Abstract

Clustering has primarily been used as an analytical technique to group unlabeled data for extracting meaningful information. The fact that no clustering algorithm can solve all clustering problems has resulted in the development of several clustering algorithms with diverse applications. We review data clustering, intending to underscore recent applications in selected industrial sectors and other notable concepts. In this paper, we begin by highlighting clustering components and discussing classification terminologies. Furthermore, specific, and general applications of clustering are discussed. Notable concepts on clustering algorithms, emerging variants, measures of similarities/dissimilarities, issues surrounding clustering optimization, validation and data types are outlined. Suggestions are made to emphasize the continued interest in clustering techniques both by scholars and Industry practitioners. Key findings in this review show the size of data as a classification criterion and as data sizes for clustering become larger and varied, the determination of the optimal number of clusters will require new feature extracting methods, validation indices and clustering techniques. In addition, clustering techniques have found growing use in key industry sectors linked to the Sustainable Development Goals (SDGs) such as manufacturing, transportation and logistics, energy, and healthcare, where the use of clustering is more integrated with other analytical techniques than a stand-alone clustering technique.

Keywords: clustering, clustering classification, clustering components, Industry applications clustering algorithms

[a] Corresponding author

1 Introduction

Clustering has been defined as the grouping of objects in which there is little or no knowledge about the object relationships in the given data (Jain et al., 1999), (Liao, 2005), (Bose and Chen, 2015),(Grant and Yeo, 2018), (Samoilenko and Osei-Bryson, 2019), (Xie et al., 2020). Clustering also aims to reveal the underlying classes present within the data. Besides, clustering is referred to as a technique that groups unlabeled data with little or no supervision into different classes. The grouping is such that objects that are within the same class have similarity characteristics and are different from objects within other classes. Clustering has also been described as an aspect of machine learning that deals with unsupervised learning. The learning lies in algorithms extracting patterns from datasets obtained either from direct observation or simulated data. Schwenker and Trentin (2014) described the learning process as attempts to classify data observations or independent variables without knowledge of a target variable.

The grouping of objects into different classes has been one of the outcomes of data clustering over the years. However, the difficulty of obtaining a single method of determining the ideal or optimal number of classes for several clustering problems has been a key clustering issue noted by several authors such as Sekula et al. (2017), Rodriguez et al. (2019), Baidari and Patil (2020). Authors have referred to this issue as the subjectivity of clustering. Sekula et al. (2017), Pérez-Suárez et al. (2019) and Li et al. (2020a) described this subjectivity as the difficulty in indicating the best partition or cluster. The insufficiency of a unique clustering technique in solving all clustering problems would imply the careful selection of clustering parameters to ensure suitability for the user of the clustering results. Jain et al. (1999) specifically noted the need for several design choices in the clustering process which have the potential for the use and development of several clustering techniques/algorithms for existing and new areas of applications. They presented general applications of clustering such as in information filtering and retrieval which could span across several industrial/business sectors. This work however discusses applications of clustering techniques specifically under selected industrial/business sectors with strong links to the United Nations Sustainable Development Goals (SDGs). We also note some new developments in clustering such as in techniques and datatype over the years of the publication of Jain et al. (1999).

This review aims to give a general overview of data clustering, clustering classification, data concerns in clustering and application trends in the field of clustering. We present a basic description of the clustering component steps, clustering classification issues, clustering algorithms, generic application of clustering across different industry sectors and specific applications across selected industries. The contribution of this work is mainly to underscore how clustering is being applied in industrial sectors with strong links to the SDGs. Other minor contributions are to point out clustering taxonomy issues, and data input concerns and suggest the size of input data is useful for classifying clustering algorithms. This review is also useful as a quick guide to practitioners or users of clustering methods interested in understanding the rudiments of clustering.

Clustering techniques have predominantly been used in the field of statistics and computing for exploratory data analysis. However, clustering has found a lot of applications in several industries such as manufacturing, transportation, medical science, energy, education, wholesale, and retail etc. Furthermore, Han et al. (2011) and Landau et al. (2011), Ezugwu

et al. (2022) indicated an increasing application of clustering in many fields where data mining or processing capabilities have increased. Besides, the growing requirement of data for analytics and operations management in several fields has increased research and application interest in the use of clustering techniques.

To keep up with the growing interest in the field of clustering over the years, general reviews of clustering algorithms and approaches have been observable trends (Jain et al., 1999),(Liao, 2005), (Xu and Wunsch, 2005), (Alelyani et al., 2013),(Schwenker and Trentin, 2014),(Saxena et al., 2017). Besides, there has been a recent trend of reviews of specific clustering techniques such as in Denoeux and Kanjanatarakul (2016), Baadel et al. (2016) Shirkhorshidi et al. (2014), Bulò and Pelillo (2017), Rappoport and Shamir (2018), Ansari et al. (2019), Pérez-Suárez et al. (2019), Beltrán and Vilariño (2020), Campello et al. (2020). We have also observed a growing review of clustering techniques under a particular field of application such as in Naghieh and Peng (2009), Xu and Wunsch (2010), Anand et al. (2018), Negara and Andryani (2018), Delgoshaei and Ali (2019). However, there appears not to be sufficient reviews targeted at data clustering applications discussed under the Industrial sectors. The application of clustering is vast, and as Saxena et al.,(2017) indicated, might be difficult to completely exhaust.

To put this article into perspective, we present our article selection method, a basic review of clustering steps, classification and techniques discussed in the literature under section 2. Furthermore, we discuss clustering applications across and within selected business sectors or Industries in section 3. A trend of how clustering is being applied in these sectors is also discussed in section 3. In section 4 we highlight some data issues in the field of clustering. Furthermore, in section 5, we attempt to discuss and summarize clustering concepts from previous sections. We thereafter conclude and suggest future possibilities in the field of data clustering in section 6.

2 Components and Classifications for data clustering

Our article selection in this work follows a similar literature search approach of Govender and Sivakumar (2020) where google scholar (which provides indirect links to databases such as science direct) was indicated as the main search engine. In addition to key reference word combinations, they used such as "clustering", "clustering analysis", we searched the literature using google scholar for clustering techniques", "approaches", "time series", "clustering sector application", "transportation", "manufacturing", "healthcare" and "energy". More search was conducted using cross-referencing and the screening of abstracts of potential articles. We ensured that the articles with abstracts containing the keywords earlier indicated were selected for further review while those not relevant to our clustering area of focus were excluded. Figure 1 below further illustrates the process of our article selection using the Prisma flow diagram (Page et al., 2021) which aims to show the flow of information and summary of the screening for different stages of a systematic review.

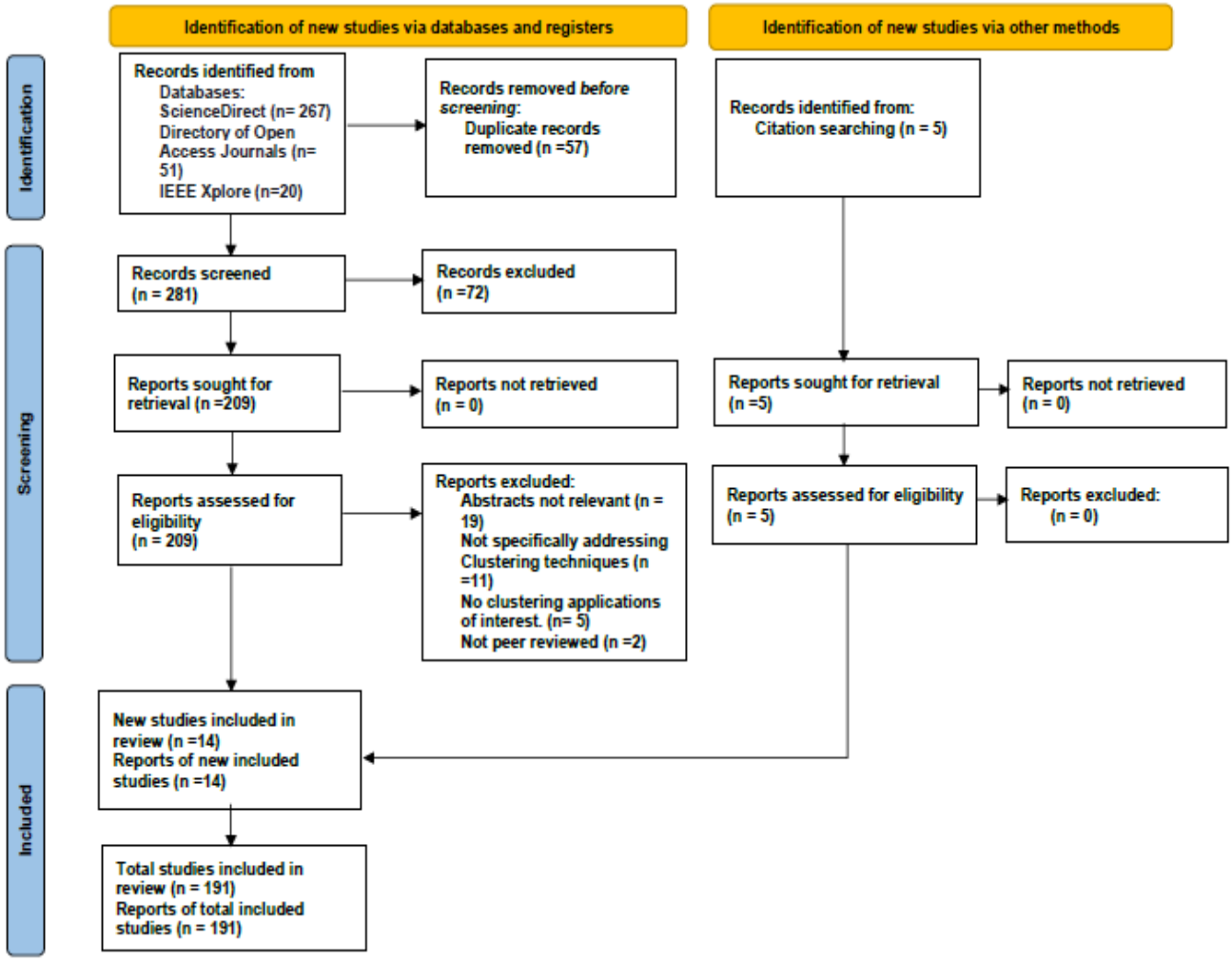

Figure 1 Article selection process using PRISMA 2020 Flow diagram

The components of data clustering are the steps needed to perform a clustering task. Different taxonomies have been used in the classification of data clustering algorithms Some words commonly used are approaches, methods or techniques (Jain et al., 1999), (Liao, 2005), (Bulò and Pelillo, 2017), (Govender and Sivakumar, 2020). However, clustering algorithms have the tendency of being grouped or clustered in diverse ways based on their various characteristics. Jain et al. (1999) described the tendency to have different approaches as a result of cross-cutting issues affecting the specific placement of clustering algorithms under a particular approach. Khanmohammadi et al. (2017) noted these cross-cutting issues as a non-mutual exclusivity property of clustering classification. We follow the logical perspective of Khanmohammadi et al. (2017) using the term criteria to classify data clustering techniques or approaches. The clustering techniques or approaches are subsequently employed to classify clustering algorithms.

2.1 Components of a Clustering task

Components of data clustering have been presented as a flow from data samples requirement through clustering algorithms to cluster formations by several authors such as Jain et al. (1999), Liao (2005), and Xu and Wunsch (2010). According to Jain et al. (1999), the following were indicated as the necessary steps to undertake a clustering activity: Pattern representation (feature extraction and selection), similarity computation, grouping process and cluster representation. Liao (2005) suggested three key components of time series clustering which are the clustering algorithm, similarity /dissimilarity measure and performance evaluation. Xu and Wunsch (2010) presented the components of a clustering task as consisting of four major feedback steps. These steps were given as feature selection/ extraction, clustering algorithm design/selection, cluster validation and result interpretation. According to Alelyani et al. (2013) components of data clustering was illustrated as consisting of the requirement of unlabeled data followed by the operation of collating similar data objects in a group and separation of dissimilar data objects into other groups. Due to the subjective nature of clustering results, the need to consider performance evaluation of any methods of clustering used has become necessary in the steps of clustering.
Taking these observations into consideration, we essentially list steps of clustering activity below and present them also in figure 2:

1) Input data requirement.
2) Pattern representation (feature extraction and selection).
3) Clustering or grouping process (Clustering algorithm selection and similarity/ dissimilarity computation).
4) Cluster formation.
5) Performance evaluation (clustering validation).
6) Knowledge extraction.

Out of the six steps highlighted above, component steps (2), (3), and (5) practically appear to be critical. This is because if the components steps (2), (3), and (5) are not appropriately and satisfactorily conducted during clustering implementation, each step or all steps (2), (3) (5) including (4) might need to be revisited. We briefly discuss these vital steps.

2.1.1 Pattern Representation (Step 2)

Jain et al.,(1999) defined pattern representation as the " number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm". They indicated that pattern representation could consist of feature extraction and/or selection. On one hand, feature selection was defined as “the process of identifying the most effective subset of the original features to use in the clustering process”. On the other hand, “Feature extraction is the use of one or more transformations of the data input features to produce new salient features to perform the clustering or grouping of data.” We refer readers to Jain et al.,(1999), Parsons et al. (2004), Alelyani et al. (2013), Solorio-Fernández et al. (2020) for a comprehensive review of pattern representation, feature selection and extraction.

2.1.2 Clustering or grouping process (Step 3)

This step is essentially described as the grouping process by Jain et al.,(1999) into a partition of distinct groups or groups having a variable degree of membership. Jain et al.,(1999) noted that Clustering techniques attempt to group patterns so that the classes thereby obtained reflect the different pattern generation processes represented in the pattern set. As noted by Liao (2005) clustering algorithms are a sequence of procedures that are iterative and rely on a stopping criterion to be activated when a good clustering is obtained. Clustering algorithms were indicated to depend both on the type of data available and the particular purpose and application. Liao (2005) discussed Similarity/dissimilarity computation as the requirement of a function used to measure the similarity between two data types (e.g., raw values, matrices, features-pairs) being compared. Similarly, Jain et al. (1999) presented this as a distance function defined on a pair of patterns or groupings. Several authors such as Jain et al. (1999), Liao (2005), Xu and Wunsch (2010), Liu et al. (2020) have noted that similarity computation is an essential subcomponent of a typical clustering algorithm. We further discuss some similarity/dissimilarity measures in section 2.4.

2.1.3 Performance Evaluation (Step 5)

This step is done to confirm the suitability of the number of clusters or groupings obtained as the results of clustering. Liao (2005) discussed this as validation indices or functions to determine the suitability or appropriateness of any clustering results. Sekula et al.,(2017) indicated that the high possibility of clustering solutions is dependent on the validation indices used and suggests the use of multiple indices for comparison.

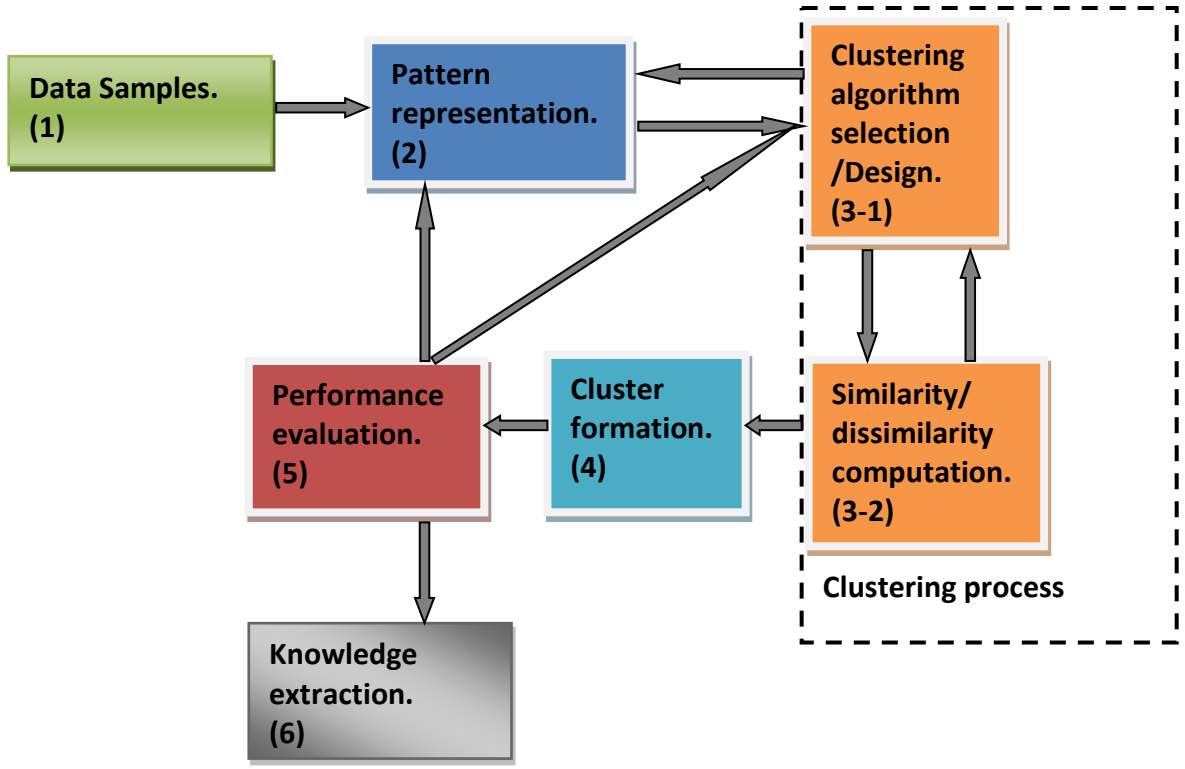

Figure 2 Typical clustering steps (1 to 6)

2.2 Clustering classification

There have been different terminologies for data clustering classification in the literature. This variety of classifications was indicated by Samoilenko and Osei-Bryson (2019), Rodriguez et al. (2019) as a means to organize different clustering algorithms in the literature. Some have used the word approaches, methods, and techniques. However, the term techniques and methods appear to have been widely used to depict the term clustering algorithms.

Liao (2005) also segmented time-series data clustering using three main criteria. These criteria referred to the manner of handling data as either in its raw form or transforming the raw data into a feature or parameters of a model. Saxena et al.,(2017) used the terminology of clustering approaches and indicated linkage to the reason for different clustering techniques. This is due to the reason for the word "cluster" not having an exact meaning for the word. Bulò and Pelillo (2017) also discussed the limitation of hard or soft classifications of clustering into partitions and they suggested an approach to clustering which was referred to as the game-theoretic framework that simultaneously overcomes limitations of the hard and soft partition approach of clustering. Khanmohammadi et al. (2017) indicated five criteria in the literature for classifying clustering algorithms which are the nature of data input, the measure of proximity of data objects, generated data cluster, membership function style and clustering strategy. These criteria have resulted in different classifications of clustering algorithms.

We present in Figure 3 below a summary of the classification criteria presented by Khanmohammadi et al. (2017). We extend the classification criteria by adding a criterion that can also be used to classify clustering algorithms. This is the size of input data. The size of data was presented as a factor that affects the selection of clustering algorithm by Andreopoulos et al. (2009), Shirkhorshidi et al. (2014) and more recently Mahdi et al. (2021). They observed that some clustering algorithms perform poorly and sacrifice quality when the size of data increases in volume, velocity, variability and variety. On another hand, some other clustering algorithms can increase scalability and speed to cope with the huge amount of data. Another possible criterion that could be added is what Bulò and Pelillo (2017) described as a framework for clustering. However, this appears to be a clustering strategy. They described this as a perspective framework of the clustering process that is different from the traditional approaches of obtaining the number of clusters as a by-product of partitioning. They referred to this as a clustering ideology which can be thought of as a sequential search for structures in the data provided. Figure 3 below categorizes the approaches or criteria and the sub-approaches or sub-criteria that can be useful in classifying clustering algorithms.

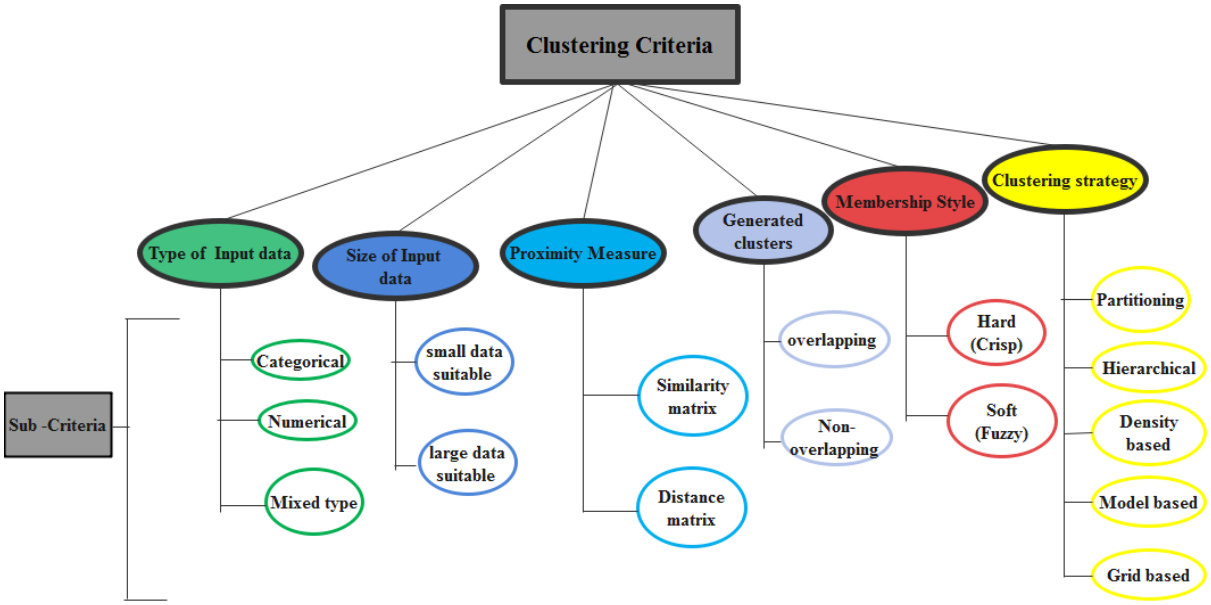

Figure 3 Criteria and sub-criteria for classifying clustering algorithms

2.3 Clustering algorithms

The criteria/sub-criteria described in the previous section can be used in classifying clustering algorithms. However, clustering algorithms have traditionally been classified as either having a partitioning (clusters obtained are put into distinctive groups) or hierarchical (forming a tree linkage or relationships for the data objects being grouped) strategy to obtain results. Jain et al. (1999) indicated the possibility of having additional categories to the traditional classification. Some authors have since then indicated the classification of clustering algorithms using five clustering strategies such as in Liao(2005), Han et al. (2011). Using the clustering criteria described earlier we demonstrate the classification of selected 21 clustering algorithms out of several clustering algorithms in the literature. These are (1) k-means, (2) k-mode, (3) k-medoid, (4), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), (5) CLustering In QUEst (CLIQUE), (6) Density clustering (Denclue), (7) Ordering Points To Identify the Clustering Structure (OPTICS), (8) STatistical INformation Grid (STING), (9) k-prototype, (10) Autoclass (A Bayesian approach to classification) (11) fuzzy k-means, (12) COOLCAT (An entropy based algorithm for categorical clustering), (13) Cluster Identification via Connectivity Kernels (CLICK), (14) RObust Clustering using linK (Sandrock), (15) Self Organising Map (SOM), (16) Single-linkage (17) Complete-linkage (18) Centroid-linkage, (19) Clustering Large Applications Based upon Randomized Search (CLARANS), (20) Overlapped k-means, (21) Model-based Overlapping Clustering (MOC).

We summarize these classifications in Tables 1 and 2 below and include selected references for extensive reading.

Table 1 Classifications of clustering algorithms based on identified clustering criteria and sub-criteria

Clustering Criteria	Sub criteria	Description	Applicable scenario(s)	Grouping of selected clustering algorithms
Type of input data: (Banerjee et al., 2005) (Andreopoulos et al., 2009), (Khanmohammadi et al., 2017)	Categorical type	Data points are usually described as qualitative data (having characteristic attributes).	Customer information such as gender, payment method etc.	k-mode(2), COOLCAT(12), CLICK(13), ROCK(14),
	Numeric type	Data points are usually described as quantitative data (measurable in numbers).	Gene expression dataset (gene vs tissue), grouping potential customers in sales and marketing.	k-means (1), k-medoid (3), DBSCAN (4), Denclue(6), OPTICS(7), Sting(8), SOM(15), CLARANS(19), Overlapped k-means (20)
	Mixed type	Data points could have numerical or categorical (discrete) descriptive attributes.	Disease data (Patient, sex, age, group).	CLIQUE (5), k-prototype (9), Autoclass(10), Fuzzy k-means(11), Single linkage(16),complete-linkage(17), centroid –linkage(18), MOC (21),
Generated clusters: (Andreopoulos et al., 2009), (N'Cir et al., 2015) (Khanmohammadi et al., 2017), (Beltrán and Vilariño, 2020)	Overlapping	Data points can belong to more than one cluster (membership either hard or fuzzy).	Social network analysis, information retrieval (e.g., several topics for a document).	Fuzzy k-means (11), Overlapped k-means (20), MOC (21)
	Non-overlapping	Data points can only belong to one of the various identified clusters (exclusive).	Clustering of movies are done by content e.g., AA, A, B, B15, C and D.	k-means(1), k-mode(2), k-medoid(3),DBSCAN(4), CLIQUE(5), Denclue(6), OPTICS(7), STING(8), k-prototype(9), Autoclass(10) COOLCAT(12), CLICK(13), ROCK(14), SOM(15), Single-linkage(16), complete-linkage(17), centroid –linkage(18), CLARANS(19)
Membership style: (Khanmohammadi et al., 2017 (Beltrán and Vilariño, 2020)	Soft (Fuzzy)	Probability membership where a data point can belong to a cluster with some degree of membership between 0 and 1.	Clustering of a range of million colours.	Fuzzy k -means (11)
	Hard (Crisp)	Binary membership where a data point can either belong or doesn't belong to a cluster (0 or 1 membership).	Group work (Grouping 12 students in a group of 4 having 3 students per group).	k-means(1), k-mode(2), k-medoid(3), DBSCAN(4), Clique(5), Denclue(6), Optics(7), Sting(8), k-prototype(9),Autoclass(10), COOLCAT(12), Click(13), Rock(14), SOM(15), Single-linkage(16), complete-linkage(17), centroid –linkage(18), CLARANS(19), Overlapped k-means (20), MOC(21)
Proximity measure: (Andreopoulos et al., 2009) (Xu and Wunsch, 2005) (Xu and Wunsch, 2010) (N'Cir et al., 2015) Khanmohammadi et al., 2017	Similarity matrix	Data points are grouped into different clusters according to their resemblance to one another or not (usually for qualitative variables)	Common in Document clustering, and gene expression data analysis. (e.g., use of cosine similarity, pearson correlation etc.)	k-mode (1), CLIQUE(5), Autoclass(10), COOLCAT(12), CLICK(13), ROCK(14)
	Distance matrix	Data points are grouped into different clusters according to certain distance functions (usually for continuous features)	Clustering using distance functions such as Euclidean, Minkowski, distance, Sup distance, city-block distance etc.	k-means (1), k-medoid (3), DBSCAN(4), Denclue(6), OPTICS(7), STING(8), k-prototype(9), Fuzzy k –means(11), SOM(15), Single-linkage(16), complete-linkage(17), centroid –linkage(18) CLARANS(19), Overlapped k-means(20),MOC(21)

Table 2: Continuation of classification of selected clustering algorithms based on identified clustering criteria and sub-criteria

Clustering Criteria	Sub criteria	Description	Applicable scenario(s)	Clustering algorithms
Clustering strategy: (Jain et al., 1999) (Han et al., 2012) (Khanmohammadi et al., 2017) (Govender and Sivakumar, 2020), **(Ezugwu et al., 2022)** (Andreopoulos et al., 2009) (Han et al., 2011) (Campello et al., 2020) (Wang et al., 1997) (Hireche et al., 2020) (Andreopoulos et al., 2009) (Hudson et al., 2011) (Bouveyron and Brunet-Saumard, 2014)	Partitioning	Given a number of partitions e.g., k-partitions, n-data objects are organized into such partitions by optimizing a partitioning criterion e.g distance function. Each partition contains at least one object such that k ≤n.	e.g., grouping post graduate students with different supervisors.	k-mean (1), k-mode(2) , k-medoids(3), k-prototype(9), fuzzy k-means(11), COOLCAT(12), CLARANS(19), Overlapped k-means(20), MOC(21)
	Hierarchical	This method works by grouping data objects into a tree of clusters. This could be agglomerative or divisive.	Clusters have different levels e.g. text mining (Subtopics of Mathematics could be algebra, calculus, trigonometry etc.	Rock (14), single-linkage (16), complete –linkage(17),centroid linkage(18) ,
	Density-based clustering	The central idea is to continue growing a cluster as long as the density (number of objects or data points) in the "neighbourhood" exceeds some threshold. Rather than producing a clustering explicitly.	e.g. bioinformatics for locating the densest subspaces in interactome networks	DBSCAN (4), Denclue(6), OPTICS(7), CLICK(13).
	Grid-based clustering	This method quantizes the object space into a finite number of cells that form a grid structure on which all the operations for clustering are performed. It clusters based on the cells rather than data objects.	Useful in facilitating several spatial queries (e.g. listing hotspot of crime within a specific distance of geographical region)	CLIQUE (5), STING (8)
	Model-based clustering	The method assumes a model for each of the clusters and attempts to best fit the data to the assumed model. Statistical and Neural networks are two approaches	In protein sequencing, bioinformatics, synchronisation of flowering (Eucalypt flower records)	Autoclass(10) , SOM(15), MOC(21),
Size of data : (Andreopoulos et al., 2009) (Khanmohammadi et al., 2017) (Shirkhorshidi et al., 2014)	Suitability for large data (High dimensional). Data	As data points increase clustering quality is minimally compromised due to scalability and speedup of the algorithm (small $O(\cdot)$ complexity)	For example social networking websites with billions of subscribers, Microarray gene expression data etc)	k-mode(2), CLIQUE(5), STING(8), SOM(15), CLARANS(19), Overlapped k-means(20)
	Non-suitability for large data (low dimensional data)	As data points increase clustering quality are largely compromised due to the high complexity of data and computational cost (large $O(\cdot)$ complexity)	extraction of knowledge from data having bytes sizes less than 10^8 bytes	k-means(1), k-medoid(3), DBSCAN(4), Denclue(6), OPTICS(7), k-prototype(9), Autoclass(10), Fuzzy k –means(11), COOLCAT(12), CLICK(13), ROCK(14), Single-linkage(16), complete-linkage(17), centroid –linkage(18) MOC(21),

$O(\cdot)$ useful in describing the effect of the size of data on clustering algorithm speed and scalability (The higher the values the slower the clustering algorithm(Andreopoulos et al., 2009)

2.3.1 Traditional clustering strategies

In this section, a basic description of clustering algorithms that represent the traditional clustering strategy of partitioning and hierarchical clustering algorithms is provided.

We present the common partitioning algorithm (k-means) and generic hierarchical clustering algorithm due to their basic usage and importance in being foundational for other clustering algorithms. This is as discussed by Xu and Wunsch (2010) and Sekula (2015), (James et al., 2015) with some modifications to aid comprehension.

Given the following notations:

n: number of observations of the data to the cluster (number of data objects).

K: the number of clusters (selected randomly or obtained through statistical tests such as in the function NBclust in the statistical program R).

C_k: Cluster centroid for each kth cluster, where k ranges from 1 to K

a) K-means algorithm

1 Randomly assign a number from 1 to K to each of the n observations. (Initial cluster assignment.

2 Iterate until the cluster assignment stops changing.

(a) For each of the kth clusters, compute the cluster centroid C_k.

(b) Assign each observation to the cluster whose centroid is closest (Where closest is defined using distance measures such as Euclidean distance).

3 Iteration and cluster assignment ends when the total within-cluster variation summed over all K clusters is as small as possible.

b) Generic Hierarchical Agglomerative Clustering

1 Begin with n observations and a distance/dissimilarity measure (such as Euclidean distance) of all $n(n-1)/2$ pairwise dissimilarities (Each observation is treated as its cluster).

2 Compute pairwise inter-cluster dissimilarities.

(a) Examine all pairwise inter-cluster dissimilarities among the individual clusters and identify the pair of clusters that are least dissimilar (Dissimilarities computed depend on the type of linkages such as complete, single, or average and the type of dissimilarity measure such as correlation-based distances, Euclidean distances).

(b) Combine these two clusters.

(c) Compute the new pairwise inter-cluster dissimilarities among remaining clusters.

3 Iteration proceeds until all n- observations belong to one cluster.

c) Generic Hierarchical divisive Clustering

The hierarchical divisive clustering is the reverse of the hierarchical agglomerative clustering.

1 Begins with one cluster (all n observations in a single cluster).

2 Split this single (large) cluster in a hierarchy fashion into new smaller clusters using a dissimilarity measure and appropriate linkage.

3 Iteration proceeds till all n observations have been allocated.

2.3.2 Traditional clustering strategy variants

Denoeux and Kanjanatarakul (2016) and Saxena et al. (2017)presented clustering algorithms as basically having either hierarchical or partitioning strategies. The density-based, grid-based, and model-based clustering strategies were indicated by them to exhibit the spirit of either the hierarchical or partitioning strategy. The classification of clustering algorithms based on one of the five clustering strategies as presented in Table 2 above appears to be widely used by several authors. Therefore, we limit our further discussions of clustering algorithms based on clustering strategy.

Some other clustering algorithms have been noted by Han et al. (2011) and Campello et al.,(2020) to possess characteristics that make them difficult to exclusively classify under one of the five clustering strategies. As a result, different classification strategies have been given in the literature to account for this (Saxena et al., 2017). Recently there have been additional clustering strategies developed such as discussed in Ezugwu et al. (2022) . Some are partly to overcome limitations of the traditional clustering techniques such as in Bulò and Pelillo (2017), Valls et al. (2018),He et al. (2020). Others have resulted from the need to apply clustering in new fields of application. Saxena et al.,(2017) also acknowledged the division of clustering algorithms into the five previous classifications above. However, they indicated other clustering methods such as multiobjective clustering, collaborative fuzzy clustering, search based clustering technique as variants of the two broad clustering methods earlier indicated. Based on their review we present a brief description.

We summarize the description of Saxena et al., (2017) and suggest other references of recent articles that have extended the selected clustering variants for detailed studies in Table 3 below.

Table 3 Clustering algorithms based on extended clustering strategy

Extended clustering strategy	Description	Clustering algorithms	Selected References
Graph (theoretic) clustering	A method that represents clusters using graphs. Graph clustering involves the task of dividing nodes into clusters so that the edge density is higher within clusters as opposed to across clusters.	complete link; minimum cut; information-theoretic; normalized cut etc.	(Matula, 1977),(Hu et al., 2009) (Das et al., 2020),(Chen et al., 2020)
Spectral clustering	This constructs affinity matrix in terms of similarity between data points before performing the clustering task. e.g Un-normalized and Normalized spectral clustering. A special case of graph-theoretic clustering. Obtaining the quality of affinity matrix and spectral vectors determination are major steps.	Traditional spectral; Spectral clustering using normalized laplacian;multi-view spectral clustering.	(Ng et al., 2002),Saxena et al., (2017),(Du et al., 2020) (Sharma and Seal, 2020)
Dominant Set Clustering (DSC)	This is based on a stepwise search for patterns or structures in data with the clustering ideology in mind similar to solution search in optimization theory, game –theory and graph theory. Another special case of graph-theoretic clustering.	DSC based on Frank-Wolfe algorithms, DSC based on replicator dynamics.	(Bulò and Pelillo, 2017),(Johnell and Chehreghani, 2020)
Evolutionary Approaches Based Clustering (EABC)	The population of solutions corresponds to the K-partitions of the data. Partitions with a large fitness value corresponding to a small square error are retained after the evolutionary operation.	EABC on particle swarm optimization, EABC on Genetic Algorithm; EABC on Ant colony optimization. Whale optimization algorithm, Crow search algorithm. Emperor Penguin Optimizer	(Jain et al., 1999) Saxena et al., (2017) (Ezugwu et al., 2022)
Search-Based Clustering Approaches (SBCA)	This comprises stochastic and deterministic techniques. The stochastic techniques are similar to the evolutionary-based approach and may not guarantee an optimal solution while the deterministic seeks to obtain optimal solutions .	SBCA on simulated annealing, SBCA on Tabu search	Saxena et al., 2017 (Bandyopadhyay et al., 2008) (Nakayama and Kagaku, 1998)
Collaborative fuzzy clustering	This is relatively recent compared to other clustering techniques. subsets of patterns can be processed together to find a structure that is common to all of them.	horizontal or vertical type.	(Hu et al., 2020) (Pedrycz, 2002) (Zhao et al., 2020)
Multi-objective clustering	Clustering criteria are jointly optimized.	MOCK; MOCA-SM	(Ramadan et al., 2020) (Kessira and Kechadi, 2020)
Overlapping clustering or overlapping community detection	Objects belong to more than one cluster or group in overlapping clustering. Overlapping community is aimed at identifying such multiple groups.	MOC; SBK; ADCLUS; OKM; DClustR;OCDC; MCLC	(Banerjee et al., 2005) (Beltrán and Vilariño, 2020) (Xie et al., 2013)
Evidential clustering (EVCLUS)	This a soft clustering technique based on determining mass functions for data objects.	EK-NNclus; EVCLUS; ECM	(Denoeux and Kanjanatarakul, 2016, Masson and Denoeux, 2008) (Denoeux, 2020)
Subspace clustering	This is an extension of feature selection that attempts to find clusters in different subspaces of the same dataset.	CLIQUE,ENCLUS, DOC, CBF,Multi-view subspace clustering	(Parsons et al., 2004),(Huang et al., 2016), (Rong et al., 2020)

We present basic steps of selected variants of the traditional clustering strategy as discussed by Saxena et al. (2017) including examples of clustering algorithms of selected clustering strategy variants as discussed by Jain et al. (1999), Pedrycz (2002), Johnell and Chehreghani (2020), Ramadan et al. (2020) with little modifications to aid basic comprehension. Given n number of observations of data the goal is to form clusters using different representations and approaches.

a) Grid-based clustering

1. Define a set of grid cells.
2. Assign observations to the appropriate grid cell and compute the density of each cell.
3. Eliminate cells, whose density is below a certain threshold.
4. Form clusters from contiguous groups of dense cells.

b) Spectral clustering

1. Construct a similarity graph between n observations or objects to be clustered.
2. Compute the associated graph Laplacian matrix (This is obtained from the weighted adjacency matrix and diagonal matrix of the similarity graph).
3. Compute the first K- eigenvectors of the Laplacian matrix to define a feature vector for each object (K implies the number of clusters to construct).
4. Organize objects into K classes by running the k-means algorithm on the features obtained.

c) Evolutionary based clustering

1. Generate (e.g., randomly) a population of solution S. Each solution S corresponds to valid K-partitions or clusters of n observations.
2. Assign a fitness value with each solution.
3. Assign a probability of selection or survival (based on the fitness value) to each solution.
4. Obtain a new population of solutions using the evolutionary operators namely selection (e.g., roulette wheel selection), recombination (e.g., crossover) and mutation (e.g., pairwise interchange mutation).
5. Evaluate the fitness values of these solutions.
6. Repeat steps 4 to 5 until termination conditions are satisfied.

d) Dominant set clustering

This follows the iterative procedure to compute clusters according to (Johnell and Chehreghani, 2020)

1. Compute a dominant set using the similarity matrix of the available n observations or data objects.
2. Remove the clustered observation from the data.
3. Repeat until a predefined number of clusters has been obtained.

e) Collaborative fuzzy clustering (Pedrycz, 2002)

This is achieved through two stages namely: (A) Generation of clusters without collaboration and (B) Collaboration of the clusters.

1. Given subsets of patterns (patterns are obtained from n observations)
2. Select distance function, number of fuzzy clusters, termination criterion, and collaboration matrix.
3. Initiate randomly all partition matrices based on the number of patterns.
4. *Stage A: Generation of clusters without collaboration*.
   4.1 Compute prototypes (centroids) and partition matrices for all subsets of patterns. (The results of clustering for each subset of patterns come in the form of a partition matrix and collection of prototypes).
   4.2 Computation is done until a termination criterion has been satisfied.
5. *Stage B: Collaboration of the clusters.*
   5.1 Given the computed matrix of collaboration
   5.2 Compute prototypes (e.g., using Lagrange multipliers technique) and partition matrices (e.g., using weighted Euclidean distances between prototype and pattern)
   5.3 Computation is done until a termination criterion has been satisfied.

f) Multi-objective clustering

The multi-objective clustering is described using the k-means modified to work on two objective functions according to Ramadan et al. (2020).

1. The data (consisting of n observations) is divided into a number of sets. The number of sets may depend on the number of distributed machines or the number of threads to be used.
2. x value (mean) and y value (variance) are computed for each set of data.
3. k-means clustering is applied to each set. K $(\text{number of clusters})$ is selected either heuristically or based on the number of records in each set.
4. At the global optimizers, Pareto optimality is applied to the clusters' centroids and nondominated centroids.
5. for nondominated clusters, the distance between a point x and the cluster centre is computed as well as the Silhouette scores between x and the nearest cluster centre. Then, the k-means algorithm is used to re-cluster those points.
6. A window W is used to extract the most effective clusters based on the required points. Pareto optimality could be applied once more for better results.

g) Search-based clustering

The search-based clustering is described using the Simulated Annealing example presented by Jain et al. (1999).

1. Randomly select an initial partition P_0 for the data (comprising n observations) and compute the squared error value termed E_{P0}.
2. Select values for the control parameters, initial and final temperatures T_0 and T_f respectively.
3. Select a neighbour partition(P_1) of P_0 and compute its squared error value termed E_{P1}.
4. If E_{P1} is larger than E_{P0}, then assign P_1 to P_0 with a temperature-dependent probability. Else assign P_1 to P_0.

5. Repeat step 3 for a fixed number of iterations.
6. Reduce the value of T_0, i.e. $T_0 = c\,T_0$, where c is a predetermined constant.
7. If T_0 is greater than T_f, then go to step 3. Else stop.

2.4 Similarity and dissimilarity measures

As indicated by Jain et al.,(1999) similarity measures are the actual strategies that clustering algorithms utilize in grouping data objects to fall within a class or cluster. The dissimilarity measures are used to differentiate a data grouping or cluster from one another. Saxena et al., (2017) also emphasized the important role similarity of objects within a cluster plays in a clustering process. According to Jain et al.,(1999), many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects and also they gave conditions for any valid distance measure. Xu and Wunsch (2010) emphasized the conditional requirement for computing similarity/dissimilarity function between any two data pairs of objects when using the distance measure. They stated that a valid similarity function or measure must satisfy the symmetry, positivity triangular inequality and reflexivity conditions. We present some of the similarity functions noted in the literature in Table 4 and suggest references to readers for more comprehensive studies. Other similarity functions or measures that have been discussed in the literature are city-block distance, sup distance, squared Mahalanobis, point symmetry distance. Xu and Wunsch (2010), Niwattanakul et al. (2013), Saxena et al. (2017) and Kalgotra et al. (2020) provide additional discussions on other similarity functions not included in this article.
Basic mathematical definitions of some of these measures as discussed by Xu and Wunsch (2010) are presented below. It is assumed that dataset $\boldsymbol{X}$ consists of $\boldsymbol{n}$ data objects or observations and $\boldsymbol{d}$ features. Notation $D(.,.) =>$ Distance function between two objects in the dataset. $\mathbf{S}(.,.) =>$ Similarity function between two objects in the dataset.

a) Minkowski distance:

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(\sum_{l=1}^{d} \left| x_{il} - x_{j\,l} \right|^{p}\right)^{\frac{1}{p}}$$

$p =>$ a generic numeric value

b) Euclidean distance:

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(\sum_{l=1}^{d} | x_{il} - x_{j\,l} |^{2}\right)^{1/2}$$

Special case of minkowski $p = 2$

c) Cosine similarity:

$$\boldsymbol{S(x_i, x_j) = cos\,\alpha = \frac{x_i{}^T\, x_j}{\|x_i\|\,\|x_j\|}}$$

d) Extended Jaccard measure

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\boldsymbol{x_i}^T \boldsymbol{x_j}}{\|\boldsymbol{x_i}\|^2 + \|\boldsymbol{x_j}\|^2 - \boldsymbol{x_i}^T \boldsymbol{x_j}}$$

Table 4 Selected use of some similarity and dissimilarity measures

Measure	Suitability	Selected Reference
Minkowski:	For numeric attributes. The similarity between data pairs corresponds to the closeness of distance between data pairs.	(Xu and Wunsch, 2010) (Saxena et al., 2017) (Xu and Wunsch, 2005)
Euclidean distance	Most commonly used for numeric attributes. A special instance of Minkowski e.g k-means algorithm.	(Thakur et al., 2020) (Qian et al., 2004)
Cosine measure	Varies more with linear transformations than rotational transformations. More commonly used for document clustering.	(Qian et al., 2004) (Ye, 2011)
Pearson correlation measure	Suitable for numeric variables and magnitude difference of two variables. Used for analyzing gene expression data.	(D'haeseleer, 2005) (Xu and Wunsch, 2010)
Jaccard measure	Suitable for information retrieval and word similarity measurement. Can detect a mistake in spellings but cannot detect over-type words	(Niwattanakul et al., 2013) Xu and Wunsch, 2005)
Dice coefficient measure	Similar to the Jaccard measure for information retrieval	(Pandit and Gupta, 2011) Xu and Wunsch, 2005)

2.5 Cluster Optimization and Validation

As indicated in the introduction section, obtaining the optimal number of clusters has been a major output of data clustering and an issue that keeps research in the field of clustering active. It has been widely indicated that no clustering algorithm can always solve all clustering problems. Saxena et al.,(2017) emphasized user control in deciding the number of cluster results, which might either follow a trial and error, heuristic or evolutionary procedure. Fu and Perry (2020) discussed some trial and error and heuristic methods of obtaining the number of clusters and proposed a method that predicts errors and subsequently chooses the smallest error to determine the appropriate number of clusters. Improving the quality of clustering results obtainable from traditional clustering algorithms

and variants have recently been advanced by some authors such as Calmon and Albi (2020), Chen et al. (2020), Ushakov and Vasilyev (2020).

As indicated by Jain et al. (1999) multiple features could be extracted or selected from given data and also performing a pairwise comparison of similarity within clusters for all data values can result in the combinatorial difficulty of clustering with an increase in data sizes. Also, Xu and Wunsch (2005) emphasized that different clustering algorithms could produce different results for a given data and also the same clustering algorithms using different approaches could still result in different clusters formed.

As a result, researchers have validated their search for the optimal number of clusters through techniques that are widely referred to as indices. Two major categories of indices have been highlighted in the literature. These are the internal indices and external indices. Some authors have indicated a breakdown of these validation indices into three categories but as Xu and Wunsch (2005) and Sekula et al.,(2017) indicated these could still be subsumed into the spirit of internal and external indices. According to Baidari and Patil (2020), Internal indices measure the compactness of the clusters by applying similarity measure techniques cluster separability and intra-cluster homogeneity, or a combination of these two Baidari and Patil (2020). External criteria are conducted to match the structure of the cluster to a predefined classification of the instances to validate clustering results. They, however, noted the common use of internal validity with clustering algorithms. Table 5 below shows selected internal and external indices from the literature.

Table 5 Selected Internal and External Validation indices

Major Indices	Other indices linked to major	Examples	Selected References
Internal indices	Stability criteria Sekula et al., 2017	Sum of squared error; Scatter criteria; Condorcet's criterion; The C – criterion; Category utility; Edge cut metrics Calinski and Harabasz (CH) index Krzanowski and Lai (KL) index Silhouette index; Gap index Compact-separate proportion (CSP) index; Index method based on data depth.	(Liu et al., 2010) (Mourer et al., 2020)
External Indices	Relative criteria (Xu and Wunsch, 2005)	Mutual information-based measure F–measure; Biological homogeneity index Biological stability index, Jaccard index, Fowlkes–Mallows index, Confusion matrix	Saxena et al., (2017) (Li et al., 2020b)

We present basic definitions of some of the indices discussed by Xu and Wunsch (2010) with some modifications to aid basic comprehension.

2.5.1 Description of selected External indices

Given a derived clustering structure C, obtained using a clustering algorithm and linked to dataset X and a prescribed clustering structure P, linked to prior information on dataset X.
$a =$ number of pairs of data objects in X, being a member of the same clusters in C and P.
$b =$ number of pairs of data objects in X, being a member of the same clusters in C and but different clusters in P.
$c =$ number of pairs of data objects in X, being a member of different clusters in C and but same clusters in P.
$d =$ number of pairs of data objects in X, being a member of different clusters in C and P.
$M = n(n-1)/2$ (Total number of pairs of objects within n number of data objects in dataset X.

a) Rand index (R):

$$R = \frac{(a+d)}{M}$$

b) Jaccard coefficient (J):

$$J = \frac{a}{(a+b+c)}$$

c) Fowlkes and Mallows Index (FM):

$$FM = \sqrt{\frac{a}{(a+b)} \quad \frac{a}{(a+c)}}$$

2.5.2 Description of selected Internal Indices:

Also given n data objects in dataset X, with K partitions indexed from $i = 1 \ to \ K$.
Where:
$n_i =$ Number of data objects assigned to cluster C_i
$m_i =$ centroid linked to cluster C_i
$m =$ total centroid (mean) vector of the dataset.
$e_i =$ average error for cluster C_i
$e_j =$ average error for cluster C_j
$D(C_i, C_j) =$ Distance function between clusters C_i and C_j in the dataset

a) Calinski and Harabasz index $(\boldsymbol{CH})$:

$$CH(K) = \frac{T_r\ (S_B\ )}{K-1} / \frac{T_r\ (S_w\ )}{n-K}$$

Where:
$T_r\ (S_B\ ) = \sum_{i=1}^{K} n_i \| m_i - m\|^{\ 2}$ (Trace of between cluster scatter matrix)

$T_r\ (S_w\ ) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \| x_j - m_i\|^{\ 2}$ (Trace within-cluster scatter matrix)
The larger the value of $CH(K)$ the better the quality of the clustering solution obtained.

b) Davies-Bouldin Index $(\boldsymbol{DB})$:

$DB(K) = \frac{1}{K} \sum_{i=1}^{K} R_i$

Where $R_i = \max_{j,\ j \neq i} \left(\frac{e_i + e_j}{\| m_i - m_j\|^{\ 2}} \right)$

The minimum $DB(K)$ indicates the potential K in the data set.

c) Dunn Index $(\boldsymbol{DI})$

$$DI(K) = \min_{i=1,\ldots,K} \left(\min_{\substack{j=1,\ldots,K, \\ j \neq i}} \left(\frac{D(\ C_i, C_j)}{\max_{l=1,\ldots,K} \delta\ (\ C_l)} \right) \right)$$

Where $D(\ C_i\ , C_j) = \min_{x \in C_i\ ,\ y \in C_j} D\ (x, y)$

$\delta(C_l) = \max_{x\ , y \in C_l} D(\ x, y)$

The larger the value of $DI(K)$ the better the estimation of K

3 Applications of clustering

Clustering techniques have been widely used in several fields and areas (Rai et al., 2006), (Devolder et al., 2012), (Bulò and Pelillo, 2017), (Grant and Yeo, 2018), (Nerurkar et al., 2018), (Govender and Sivakumar, 2020). Its relevance has also been shown as an analytical technique on its own (Ray and Turi, 1999), (Lismont et al., 2017), (Motiwalla et al., 2019) and also as a hybrid method with other analytical solution techniques such as in Grant and Yeo (2018), Zhu et al. (2019), Liu and Chen (2019), Jamali-Dinan et al. (2020), Tanoto et al. (2020), Pereira and Frazzon (2020)). We review some field applications of clustering and subsequently review the application of clustering techniques in particular business sectors or fields.

3.1 Field applications

Some of the direct areas of clustering application generally discussed in the literature have been textual document classification, image segmentation, object recognition, character recognition, information retrieval, data mining, spatial data analysis, business analytics, data reduction, and big data mining. Other areas indicated by Saxena et al., (2017), have been sequence analysis (Durbin et al., 1998) (Li et al., 2012), human genetic clustering, (Kaplan and Winther, 2013), (Lelieveld et al., 2017), (Marbac et al., 2019), mobile banking and information system (Motiwalla et al., 2019) (Shiau et al., 2019), social network analysis (Scott and Carrington, 2011), (Shiau et al., 2017), (Khamparia et al., 2020), search result grouping (Mehrotra and Kohli, 2016) (Kohli and Mehrotra, 2016), software evolution (Rathee and Chhabra, 2018) (Izadkhah and Tajgardan, 2019), recommender systems (Petwal et al., 2020), educational data mining (Baker, 2010), (Guleria and Sood, 2020), climatology (Sharghi et al., 2018) (Pike and Lintner, 2020), (Chattopadhyay et al., 2020) and robotics (Khouja and Booth, 1995), (Zhang et al., 2013). In Table 5 below we briefly discuss a few applications as indicated by Saxena et al.,(2017) and also provide references for more detailed studies.

Table 5 Some field applications of clustering techniques

Field	Application of clustering	References
Textual documents, Document storage	Basically, clustering of texts. Efficient document storage and retrieval for many institutions of learning have been noted to be one of the important applications of clustering. In addition, discovering events and sub-events from a sequence of news articles.	(Rasmussen, 1992), (Piernik et al., 2015), (Chan et al., 2016),(Lee et al., 2020), (Celardo and Everett, 2020)
Image segmentation	This is centered around the partition of images for visibility and classification of images based on some properties.	(Forsyth and Ponce, 2002), (Lam and Wunsch, 2014) (Zhang et al., 2020)
Object recognition	3D object grouping has been an area of application.	(Dorai and Jain, 1995)
Character recognition	handwriting recognition has been an important application.	(Connell and Jain, 1998)
Data mining	Widely used in this field both to analyze structured and unstructured databases.	(Hedberg, 1996),(Han et al., 2011)
Spatial and space application	Large data sets from geographical information systems and satellite images have been analyzed using clustering techniques.	(Upton and Fingleton, 1985) (Tahmasebi et al., 2012), (Song et al., 2020) (Zhang et al., 2020)
Business analytics	Operational areas of marketing, demand management and production areas of product development and categorization.	(Kiang et al., 2007), (Fennell et al., 2003), (Pereira and Frazzon, 2020)
Data reduction	Compression of large data into manageable sizes usually saves processing time and cost.	(Jiang et al., 2016) (Huang, 1997)
Big data mining	For databases with a growing capacity of being exponential beyond manageable sizes of conventional database tools.	(Shirkhorshidi et al., 2014) (Russom, 2011) (Ezugwu et al., 2022)
Social networking	Applied in behavioural grouping of people and activities such as e-governance and educational learning sites.	(Cheng et al., 2020), (Khamparia et al., 2020)
Non-numerical openly expressed information	Categorizing verbal information using motivation(push theory) and meaning (pull theory)e.g. in profiling tourists based on motivations for destinations and meanings of destinations to the same tourists.	(Batet et al., 2010), (Valls et al., 2018)

3.2 Selected Industry applications

The application fields or areas of clustering described above have been noted to be in general areas of application that possibly cut across through different industrial and business sectors. Clustering techniques have also found extensive application in certain industries. As indicated by Dalziel et al. (2018) different firms with similar buy-sell characteristics could be grouped under the same industry. Clustering has been used partly as a stand-alone analytical technique and largely as a hybrid technique with other analytical methods to solve industrial problems. According to Jakupović et al. (2010), Dalziel et al. (2018), (Grant and Yeo, 2018) (Xu et al., 2020) and (Ezugwu et al., 2022) several business or industrial sectors exist. They further noted that a unique or universal classification of industries or business sectors is difficult due to the reason that industries or sectors are mostly classified based on the specific needs of the classifier.

According to Citizenship (2016), ten (10) industrial sectors of impact on the SDGs were identified namely Consumer goods, Industrials, Oil and Gas, Healthcare, Basic Materials, Utilities, Telecomms, Financials, Consumer Services and Technology. In addition, the industrial sectors were organised into three namely; the primary sector (raw material extraction and production), Secondary(production of goods from raw materials) and Tertiary (provision of services). These industries have also been noted to have strong linkages to either one or more SDGs. For example, Healthcare strongly impacts SDG3 which is to achieve good health and well-being for all, while Oil and gas are strongly linked to SDG 7 (affordable and clean energy). Consumer goods, industrials and consumer services impact across SDG12 (responsible consumption and production), SDG2 (achieving Zero hunger) and SG14 (on the protection of the marine environment). Furthermore, the Utilities sector is known for infrastructure provision impacts across SDG 6 (clean water and sanitation), SDG7 and SDG9 (decent work and employment). Others such as SDG 1 (poverty), SDG4 (education), and SDG 5(gender equality) have been known to be of low impact on a particular sector and receive supporting actions from the earlier discussed industrial /business sectors.

As several clustering techniques have been extensively reported in the literature, chances also exist of a corresponding application of clustering techniques in several identified industries/sectors. Using the SDG classifications indicated above, we select sectors important in driving most of the SDGs. These sectors are mostly grouped under Transportation and logistics (such as consumer services), Manufacturing (such as Industrials, basic materials, consumer goods), Energy (such as Oil and Gas, Utilities) and Healthcare. In addition, the selected industries positively impact or stimulate economic growth, innovation, development gaps and well being for a typical economy (Nhamo et al., 2020), (Shi, 2020), (Abbaspour and Abbasizade, 2020).

3.2.1 Transportation and Logistics

The application of clustering in the transportation industry has been generally noted to be in the identification of similar patterns in various modes of transportation (Almannaa et al., 2020). Some fields under the transportation sector, where clustering application has been applied have been hazardous transportation, road transportation urban/public transportation (De Luca et al., 2011),(Lu et al., 2013), (Rabbani et al., 2017) (Sfyridis and

Agnolucci, 2020), (Almannaa et al., 2020). Recently, Wang and Wang (2020) discussed the application of genetic fuzzy C-means algorithm and factor analysis to identify the causes and control high-risk drivers. (de Armiño et al., 2020) combined the hierarchical clustering and neural networks to develop a linkage between road transportation data and macroeconomic indicators. Almannaa et al. (2020) developed a multi-objective clustering that can maximize purity and similarity in each cluster formed simultaneously. They also noted that the convergence speed of the multi-objective clustering method was fast, and the number of clusters obtained was stable to determine traffic and bike pattern change within clusters.

3.2.2 Manufacturing

Similarly to clustering applications in the transportation sector, the manufacturing sector and systems such as discussed by (Delgoshaei and Gomes, 2016), (Delgoshaei et al., 2021) also possess a wide application of clustering techniques. The applications are mostly a hybrid method with other analytical methods. Using the case study of the textile manufacturing business, Li et al. (2011) used clustering analysis to classify customers based on selected customer characteristics and further used some cross-analysis for customer behavior tendencies. P CHANDRASEKHARAN and Rajagopalan (1986) adopted k-means in a group technology problem following which the initial groupings obtained were improved. There has been a recent trend of application of clustering techniques in cloud manufacturing, cyber manufacturing, smart manufacturing, manufacturing systems and cellular manufacturing. Delgoshaei and Ali (2019) reviewed hybrid clustering methods and search algorithms such as metaheuristics in the designing of cellular manufacturing systems. Liu and Chen (2019) used the k-medoids clustering-based algorithm and trust-aware approach to predict the quality-of-service records which might become intractable under cloud manufacturing. An improved k-mean clustering technique was compared to a k-means random by Yin (2020). The comparison was done to determine which method could provide an optimal number of edge computing nodes in a smart manufacturing setup. Sabbagh and Ameri (2020) demonstrated the application of unsupervised learning in text analytics. They used the k-means clustering and topic modelling techniques to build a cluster of supplier capabilities topics. Subramaniyan et al. (2020) clustered time-series bottle-neck data using dynamic time wrapping and complete-linkage agglomerative hierarchical clustering technique for determining bottlenecks in manufacturing systems. Ahn and Chang (2019) discussed business process management for manufacturing models. They used agglomerative hierarchical clustering in the design and management of manufacturing processes. A hybrid dynamic clustering and other techniques for establishing similarities in 3D geometry of parts and printing processes were investigated by Chan et al. (2018).

3.2.3 Energy

Clustering techniques have also been widely used in the field of energy both in isolation and in combination with other analytical techniques. Some fields under energy where clustering applications that have been used include energy efficiency, renewable energy, electricity consumption, heating and cooling, nuclear energy, and smart metering. The k-means clustering technique and its variants have mostly been used in the energy sector clustering. Vialetto and Noro (2020) used the k-means clustering, silhouette method to define the number of clusters while clustering energy demand data. They used clustering in the design of cogeneration systems to allow energy-cost savings. Wang and Yang (2020) used fuzzy

clustering and an accelerated genetic algorithm to analyze and assess sustainable and influencing factors for 27 European Union countries' renewable energy. Fuzzy C means and multi-criteria decision-making process were applied by (Tran, 2020) to design the optimal loading of ships and diesel fuel consumption of marine ships. Tanoto et al. (2020) applied a hybrid of k-means clustering, neural network based-self organizing map to group technology mixes with similar patterns. Their method was designed for the energy modelling community for the understanding of complex design choices for electricity industry planning. Suh et al. (2020) applied text mining in nuclear energy. Clustering analysis and technology network analysis were used to identify topics in nuclear waste management over time. Shamim and Rihan (2020) compared using k-means clustering and k-means clustering with feature extraction in smart metering electricity. Results of their experiments showed that clustering using features from raw data obtained better performance than direct raw data.

3.2.4 Healthcare

The healthcare industry has been described as one that can generate a vast amount of data from diverse clinical procedures and sources in which clustering techniques are found useful (Palanisamy and Thirunavukarasu, 2019), (Ambigavathi and Sridharan, 2020). According to Jothi and Nur'Aini Abdul Rashidb (2015), Manogaran and Lopez (2017), Palanisamy and Thirunavukarasu (2019) and Shafqat et al. (2020) some heterogeneous data sources in the healthcare industry include electronic health records, medical imaging, genetic data, clinical diagnosis, metabolomics, proteomics and long-term psychological sensing of an individual.

Clustering techniques have been useful in the healthcare industry as part of data mining techniques for the identification of patterns in healthcare data sets (Jothi and Nur'Aini Abdul Rashidb, 2015), (Ahmad et al., 2015), (Ogundele et al., 2018). As described by Ogundele et al. (2018) data mining is the field of study that seeks to find useful and meaningful information from large data. This definition makes data mining techniques such as clustering relevant in the health care industry. Ahmad et al. (2015) showed with examples that clustering algorithms could be used as a stand-alone technique or as a hybrid with other analytical techniques in understanding healthcare datasets. The use of clustering algorithms such as k-means, k-medoids, and x-means has been used to diagnose several diseases such as breast cancers, heart problems, diabetes, and seizures (Ahmad et al., 2015) (Alsayat and El-Sayed, 2016), (Kao et al., 2017), (Ogundele et al., 2018), (Shafqat et al., 2020). To understand patterns in the automatically-collected event in healthcare settings, patient flow and clinical setting conformance, Johns et al. (2020) discussed the use of trace clustering. Density-based clustering has also been applied to obtain useful patterns from biomedical data (Ahmad et al., 2015). Hybrid techniques for analyzing and predicting health issues such as the use of clustering algorithms and classification trees, the use of k-means and statistical analysis and hybrid hierarchical clustering were discussed by (Ahmad et al., 2015).

Yoo et al. (2012), Jothi and Nur'Aini Abdul Rashidb (2015) and Ogundele et al. (2018) indicated that clustering techniques (unsupervised learning) form the descriptive components of data mining techniques. In addition, Jothi and Nur'Aini Abdul Rashidb (2015), noted that clustering techniques are not as utilized as the prescriptive (Supervised) components of data mining techniques. Ahmad (2015) however pointed out that a combination of different data mining techniques should be used to achieve better disease

prediction, clinical monitoring, and general healthcare improvement in the healthcare industry.
Figure 4 below summarizes the general application of clustering techniques based on the identified industries above.

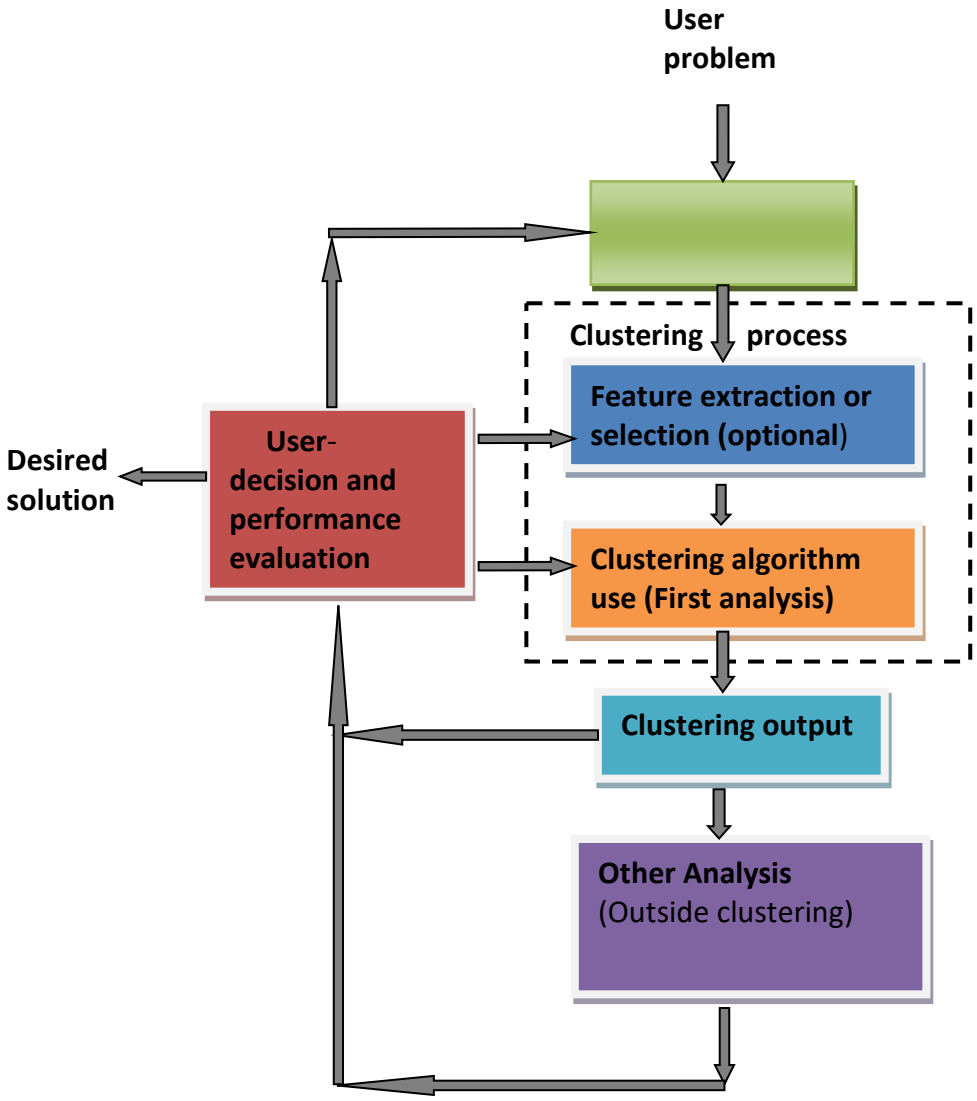

Figure 4 General application of clustering techniques

4 Data Size, Dimensionality, and Data type issues in clustering

One of the approaches earlier listed for classifying clustering algorithms is the type of input data. Liao (2005) observes that the data that can be inputted into any clustering task can be classified as binary, categorical, numerical, interval, ordinal, relational, textual, spatial, temporal, spatio-temporal, image, multimedia, or mixtures of the above data types. This

classification can also be sub-classified. For example, numeric raw data for clustering can either be static, time series or as a data stream. Static data do not change with time while time-series data have their data objects changing with time. Aggarwal et al. (2003) described data stream as large volumes of data arriving at an unlimited growth rate. As noted by Mahdi et al. (2021) data types that are vast and complex to store such as social network data (referred to as big data) and high-speed data (data stream) such as web-click streams, network traffic could be challenging to cluster. In addition, they emphasized that the type of data type considered often influences the type of clustering techniques selected.
The application of some clustering algorithms directly to raw data has been noted to be an issue as the data size becomes larger (Gordon, 1999), (Parsons et al., 2004). Two reasons were given for this observed problem. The first reason indicated was based on the type of clustering algorithm used. This is such that some clustering algorithms fully take into consideration all dimensions of the data during the clustering process. As a result, they conceal potential clusters of outlying data objects. The second was because, as dimensionality increases in the given data, the distance measure for computing similarity or dissimilarity among data objects becomes less effective. Feature extraction and selection were suggested as a generic method to solve this problem by reducing the dimensionality of the data before the clustering algorithms are applied. However, they noted that this feature-based method could omit certain clusters hidden in subspaces of the data sets. Subspace clustering was the method suggested to overcome this.
Research in the field of reducing the dimensionality of the original data through feature extraction and selection methods and variants such as subspace clustering has continued to be investigated by several authors (Huang et al., 2016), (Motlagh et al., 2019), (Solorio-Fernández et al., 2020). Huang et al. (2016) specifically indicated time-series data to be subject to large data sizes, high dimensionality, and progressive updating. They suggested the preference of clustering over time segments of time-series data compared to the whole time-series sequence to ensure all hidden clusters in the time series data are accounted for. Hence data pre-processing techniques such as (Normalization, cumulative clustering etc.) have been suggested. Pereira and Frazzon (2020) utilized data preprocessing to detect and remove outliers followed by normalization before a clustering algorithm was applied. Li et al. (2020a) considered ameliorating datasets to improve clustering accuracy by transforming bad data sets into good data sets using the HIBOG. Solorio-Fernández et al. (2020) presents a comprehensive review of feature selection to highlight the growing advances of unsupervised feature selection methods (filter, wrapper and hybrid) for unlabeled data.
Clustering of data could also become an issue when multi-source and multi-modal data are considered. Multi-source data (originating from several sources) have been observed with characteristics such as complexity, heterogeneity, dynamicity, distribution and largeness (Uselton et al., 1998). As noted by Sprague et al. (2017) and Afyouni et al. (2021) the combination or fusion of data from diverse organizations having different reporting formats, structures and dimensions could present some complexities in multi-source data. Lahat et al. (2015) and Li and Wang (2021) discussed the complementary and diverse attributes of multi-modal data (e.g. the same data from text, image audio, and video) and also provided similar challenges of complexity resulting from the fusion of such data. Adaptation of existing clustering algorithms or development of new clustering algorithms will become useful to analyze such potential big and complex data.
Since clustering results are strongly linked to the type and features of the data being represented, their performance is being improved through current supervised machine

learning methods such as Deep Neural Networks (DNN). As noted by James et al. (2015) and Ni et al. (2022), DNN have had more successful performance (e.g. in speech and text modelling, video and image classification) compared to the earlier developed neural networks as seen in (Hastie et al., 2009) due to the less training tinkering required and increasing availability of large training data sets. DNN could be used to obtain improved feature representation useful for clustering before the actual clustering is performed. This has been referred to as deep clustering in the machine learning field (Aljalbout et al., 2018). According to Min et al. (2018) emphasis was placed on prioritizing network architecture over clustering loss in classifying deep clustering due to the basic desire for clustering-oriented representations. They further classified deep clustering based on: (I) the use of Autoencoder (AE) to obtain the feasible feature representation (II) Feedforward networks such as Feedforward convolutional networks which can use specific clustering loss to obtain feasible feature representation (III) Generative Adversarial Network (GAN) and Variational Autoencoder (VAE) which uses effective generative learning frameworks to obtain feature representations.

5 Discussions

In this section, we highlight the major considerations in the earlier sections and project possible application trends in the field of clustering. In section 2, we noted some inconsistencies in terminologies and classification criteria used in grouping clustering algorithms and their variants. Authors in the field of data clustering have suggested different terminologies for group clustering algorithms. The partitioning and hierarchical approaches have primarily been used to group clustering algorithms. Other approaches such as density-based, model-based, and grid-based have been suggested as an extension to the primary approaches. The classification of the five clustering approaches earlier mentioned can be categorized as clustering strategies. Other clustering criteria such as proximity measure, input data, size of input data, membership function style, and generated clusters can further be used to categorize different approaches employed in classifying clustering algorithms. The selection and design of clustering algorithms are observed to be a vital step in the clustering components. We suggest that the clustering component steps tend towards being cyclical with feedback than a straight follow. This relates more with the reality of iteration in obtaining the appropriate cluster results.

The reality is that there is no universally accepted clustering algorithm to solve all clustering problems (Jain et al., 1999), (Rodriguez et al., 2019) and the limitation of clustering algorithms is a strong motivation for the emergence of new clustering algorithms or variants of the traditional clustering algorithms. As new clustering algorithms emerge, it is expected that existing terminologies and classification approaches could become broader with a seeming departure from the traditional approaches. With the growing number of clustering algorithms is also the growing number of clustering validation indices. This perhaps is due to the reason that users of clustering results are more interested in knowing with good confidence that clustering results obtained are well suited for the application. To test the suitability of different clustering algorithms and indices in meeting the users' needs and also due to the increase in computing technological capabilities, clustering algorithms and indices are being combined in computer programs. Rodriguez et al. (2019) presented a

comparative study of 9 clustering algorithms available in the R programming language. Other authors such as Sekula (2015) have indicated some clustering packages in the R-programming language that can be useful for comparison and as a friendly user application. Besides, computer programs are used to suggest a suitable number of clusters for clustering algorithms (e.g k-means) requiring an input of clusters as applied by (Rhodes et al., 2014), (Charrad et al., 2015).

In section 3, we considered that the application of clustering has largely been reported in areas such as image segmentation, object recognition, character recognition, information retrieval, and data mining. We have considered these areas to be specific applications of clustering algorithms. It is expected that more field applications will be reported due to the vast applicability of clustering techniques. Also emphasized is the application of clustering in selected industrial sectors. We specifically noted the diverse classification schemes and groupings of industrial sectors. The numerous clustering algorithms in existence have the corresponding possible applicability in several of these industries. We, however, selected manufacturing, energy, transportation and logistics, and healthcare as examples to illustrate the application of clustering in industries with important links to achieving sustainable development goals. The application of clustering techniques in these industries appears to be a move from a stand-alone analytical technique into hybrid techniques with other analytical processes. This suggests that clustering techniques will continue to be relevant as an integrated analytical technique in different industries and sectors. Besides, the vast application of clustering techniques will imply practitioners or users with a basic understanding of clustering techniques can use the clustering algorithm embedded into the software with little difficulty.

In section 4, we highlighted some data sources used in clustering and discussed some data issues users of clustering techniques are likely to deal with. Clustering raw data inputs are generally observed to be more problematic than refined data inputs. This is attributable to the dimensionality problem. Due to the increase in computing technology for many industrial applications and cloud computing, the use of clustering techniques to analyze high volumes of static, time-series, multi-sources, and multimodal data are trends in the future. For multi-sources and multimodal data, applications or frameworks that can effectively integrate or fuse the complementary attributes of such data are currently observable trends. As such clustering techniques will be more readily deployed in such secondary data-use domain.

As the size of data becomes larger due to modern data mining capabilities and the need to avoid incomplete knowledge extraction from single sources or modes of data, methods that fuse complementary and diverse data with a goal of understanding and identifying hidden clusters are also notable trends. For example, deep learning methods are sometimes merged with traditional clustering methods to further search for underlying clusters and thereby improve clustering performance.

Putting the main observations in this paper together, the emergence of new clustering algorithms is expected due to the subjectivity nature of clustering and its vast applicability in diverse fields and industries. This suggests that emerging scholars can find meaningful research interest in several aspects of data clustering such as the development of new clustering algorithms, validity indices, improving clustering quality and comprehensive field and industry reviews of clustering techniques. Industry Practitioners will also find use in the application of specific clustering algorithms to analyze unlabeled data to extract meaningful information.

6 Conclusion and Future Directions

In this paper, we presented a basic definition and description of clustering components. We extended existing criteria in the literature for classifying clustering algorithms. Both traditional clustering algorithms and emerging variants were discussed. Also emphasized is the reality that clustering algorithms could produce different groupings for a given set of data. Also, as no clustering algorithm can solve all clustering problems, several clustering validation indices are used and have also been developed to gain some confidence in the cluster results obtained from clustering algorithms.

We summarized field applications of clustering algorithms such as in image segmentation, object recognition, character recognition, data mining and social networking that have been pointed out in the literature. Selected applications of clustering techniques and notable trends in industrial sectors with strong links to achieving sustainable development goals were further presented to show the diverse application of clustering techniques. Also suggested are possible application trends in the field of clustering that are observable from both specific and general article reviews in the literature. Some data input concerns in the field of clustering were examined.

This study presents a foundation for other research work that can be projected from it. Firstly, the investigation into feature extraction, selection, alignment, and other methods that could reveal hidden clusters in large volumes, high-frequency data such as data streams, multi-modal and multi-source data obtainable from current data mining capabilities, technologies and computer simulations are current and research interest into the future for the academia and industry.

In addition, the development of new clustering strategies to analyze existing and modern data types (e.g., fused multi-source and multi-modal data) would also be of more interest to researchers. The outputs and knowledge extracted from such data types could be beneficial to policymakers and business practitioners in informed decision making.

Secondly, the use of clustering techniques has a high possibility of finding more applicability in existing fields. Examples are text mining, industrial big data applications, biomedicals, commercial sectors, military applications, space navigation and biological processes. In emerging areas of applications such as Learning management systems and social media that currently churn out huge amounts of data and have recently seen a further increase due to the covid-19 pandemic, the development of effective and efficient clustering algorithms to sufficiently mine the massive amount of data from such fields are currently being projected. Deep clustering will generally find more applications in analysis useful across different business sectors where pure clustering methods have been used. This will be due to observed performance in obtaining better clustering results for example in image classification where the Feedforward convolutional network has been very useful.

Finally, a data clustering trend that summarizes trends from qualitative and quantitative results of the application of diverse variants of clustering strategies will adequately be an improvement on this research efforts.

Data availability Statement
Data sharing not applicable to this article as no datasets were generated or analysed during the current study

7 References

ABBASPOUR, M. & ABBASIZADE, F. 2020. Energy Performance Evaluation Based on SDGs. *In:* LEAL FILHO, W., AZUL, A. M., BRANDLI, L., LANGE SALVIA, A. & WALL, T. (eds.) *Affordable and Clean Energy.* Cham: Springer International Publishing.

AFYOUNI, I., AL AGHBARI, Z. & RAZACK, R. A. 2021. Multi-feature, multi-modal, and multi-source social event detection: A comprehensive survey. *Information Fusion*.

AGGARWAL, C. C., PHILIP, S. Y., HAN, J. & WANG, J. A framework for clustering evolving data streams. Proceedings 2003 VLDB conference, 2003. Elsevier, 81-92.

AHMAD, P., QAMAR, S. & RIZVI, S. Q. A. 2015. Techniques of data mining in healthcare: a review. *International Journal of Computer Applications,* 120.

AHN, H. & CHANG, T.-W. 2019. A similarity-based hierarchical clustering method for manufacturing process models. *Sustainability,* 11**,** 2560.

ALELYANI, S., TANG, J. & LIU, H. 2013. Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications***,** 29.

ALJALBOUT, E., GOLKOV, V., SIDDIQUI, Y., STROBEL, M. & CREMERS, D. 2018. Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv:1801.07648*.

ALMANNAA, M. H., ELHENAWY, M. & RAKHA, H. A. 2020. A Novel Supervised Clustering Algorithm for Transportation System Applications. *IEEE Transactions on Intelligent Transportation Systems,* 21.

ALSAYAT, A. & EL-SAYED, H. Efficient genetic K-means clustering for health care knowledge discovery. 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), 2016. IEEE, 45-52.

AMBIGAVATHI, M. & SRIDHARAN, D. Analysis of Clustering Algorithms in Machine Learning for Healthcare Data. International Conference on Advances in Computing and Data Sciences, 2020. Springer, 117-128.

ANAND, S., PADMANABHAM, P., GOVARDHAN, A. & KULKARNI, R. H. 2018. An extensive review on data mining methods and clustering models for intelligent transportation system. *Journal of Intelligent Systems,* 27**,** 263-273.

ANDREOPOULOS, B., AN, A., WANG, X. & SCHROEDER, M. 2009. A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in bioinformatics,* 10**,** 297-314.

ANSARI, M. Y., AHMAD, A., KHAN, S. S. & BHUSHAN, G. 2019. Spatiotemporal clustering: a review. *Artificial Intelligence Review***,** 1-43.

BAADEL, S., THABTAH, F. A. & LU, J. 2016. Overlapping Clustering: A Review. IEEE.

BAIDARI, I. & PATIL, C. 2020. A Criterion for Deciding the Number of Clusters in a Dataset Based on Data Depth. *Vietnam Journal of Computer Science***,** 1-15.

BAKER, R. 2010. Data mining for education. *International encyclopedia of education,* 7**,** 112-118.

BANDYOPADHYAY, S., SAHA, S., MAULIK, U. & DEB, K. 2008. A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE transactions on evolutionary computation,* 12**,** 269-283.

BANERJEE, A., KRUMPELMAN, C., GHOSH, J., BASU, S. & MOONEY, R. J. Model-based overlapping clustering. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005. 532-537.

BATET, M., VALLS, A. & GIBERT, K. Performance of ontology-based semantic similarities in clustering. International Conference on Artificial Intelligence and Soft Computing, 2010. Springer, 281-288.

BELTRÁN, B. & VILARIÑO, D. 2020. Survey of Overlapping Clustering Algorithms. *Computación y Sistemas,* 24.

BOSE, I. & CHEN, X. 2015. Detecting the migration of mobile service customers using fuzzy clustering. *Information & Management,* 52**,** 227-238.

BOUVEYRON, C. & BRUNET-SAUMARD, C. 2014. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis,* 71**,** 52-78.

BULÒ, S. R. & PELILLO, M. 2017. Dominant-set clustering: A review. *European Journal of Operational Research,* 262**,** 1-13.

CALMON, W. & ALBI, M. 2020. Estimating the number of clusters in a ranking data context. *Information Sciences,* 546**,** 977-995.

CAMPELLO, R. J., KRÖGER, P., SANDER, J. & ZIMEK, A. 2020. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 10**,** e1343.

CELARDO, L. & EVERETT, M. G. 2020. Network text analysis: A two-way classification approach. *International Journal of Information Management,* 51**,** 102009.

CHAN, L. M., INTNER, S. S. & WEIHS, J. 2016. *Guide to the Library of Congress classification*, ABC-CLIO.

CHAN, S. L., LU, Y. & WANG, Y. 2018. Data-driven cost estimation for additive manufacturing in cybermanufacturing. *Journal of manufacturing systems,* 46**,** 115-126.

CHARRAD, M., GHAZZALI, N., BOITEAU, V. & NIKNAFS, A. 2015. Determining the Best Number of Clusters in a Data Set. *Recuperado de: https://cran. rproject. org/web/packages/NbClust/NbClust. pdf*.

CHATTOPADHYAY, A., HASSANZADEH, P. & PASHA, S. 2020. Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports,* 10**,** 1-13.

CHEN, H., YU, Z., YANG, Q. & SHAO, J. 2020. Attributed Graph Clustering with Subspace Stochastic Block Model. *Information Sciences*.

CHENG, H., HONG, S. A. & YE, X. 2020. Clustering users of a social networking system based on user interactions with content items associated with a topic. Google Patents.

CITIZENSHIP, C. 2016. SDGs & Sectors: A review of the business opportunities. *London (UK): Corporate Citizenship*.

CONNELL, S. D. & JAIN, A. K. Learning prototypes for online handwritten digits. Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170), 1998. IEEE, 182-184.

D'HAESELEER, P. 2005. How does gene expression clustering work? *Nature biotechnology,* 23**,** 1499-1501.

DALZIEL, M., YANG, X., BRESLAV, S., KHAN, A. & LUO, J. 2018. Can we design an industry classification system that reflects industry architecture? *Journal of Enterprise Transformation***,** 1-25.

DAS, S., DAS, A., BHATTACHARYA, D. & TIBAREWALA, D. 2020. A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets. *Genomics*.

DE ARMIÑO, C. A., MANZANEDO, M. Á. & HERRERO, Á. 2020. Analysing the intermeshed patterns of road transportation and macroeconomic indicators through neural and clustering techniques. *Pattern Analysis and Applications***,** 1-12.

DE LUCA, M., MAURO, R., RUSSO, F. & DELL'ACQUA, G. 2011. Before-after freeway accident analysis using Cluster algorithms. *Procedia-social and behavioral sciences,* 20**,** 723-731.

DELGOSHAEI, A. & ALI, A. 2019. Evolution of clustering techniques in designing cellular manufacturing systems: A state-of-art review. *International Journal of Industrial Engineering Computations,* 10**,** 177-198.

DELGOSHAEI, A., ARAM, A. K., EHSANI, S., REZANOORI, A., HANJANI, S. E., PAKDEL, G. H. & SHIRMOHAMDI, F. 2021. A supervised method for scheduling multi-objective job shop systems in the presence of market uncertainties. *RAIRO-Operations Research,* 55**,** S1165-S1193.

DELGOSHAEI, A. & GOMES, C. 2016. A multi-layer perceptron for scheduling cellular manufacturing systems in the presence of unreliable machines and uncertain cost. *Applied Soft Computing,* 49**,** 27-55.

DENOEUX, T. 2020. Calibrated model-based evidential clustering using bootstrapping. *Information Sciences*.

DENOEUX, T. & KANJANATARAKUL, O. Evidential Clustering: A Review. 2016.

DEVOLDER, P., PYNOO, B., SIJNAVE, B., VOET, T. & DUYCK, P. 2012. Framework for user acceptance: Clustering for fine-grained results. *Information & Management,* 49**,** 233-239.

DORAI, C. & JAIN, A. K. Shape spectra based view grouping for free-form objects. Proceedings., International Conference on Image Processing, 1995. IEEE, 340-343.

DU, T., WEN, G., CAI, Z., ZHENG, W., TAN, M. & LI, Y. 2020. Spectral clustering algorithm combining local covariance matrix with normalization. *Neural Computing and Applications,* 32**,** 6611-6618.

DURBIN, R., EDDY, S. R., KROGH, A. & MITCHISON, G. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge university press.

EZUGWU, A. E., IKOTUN, A. M., OYELADE, O. O., ABUALIGAH, L., AGUSHAKA, J. O., EKE, C. I. & AKINYELU, A. A. 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence,* 110**,** 104743.

FENNELL, G., ALLENBY, G. M., YANG, S. & EDWARDS, Y. 2003. The effectiveness of demographic and psychographic variables for explaining brand and product category use. *Quantitative Marketing and Economics,* 1**,** 223-244.

FORSYTH, D. A. & PONCE, J. 2002. *Computer vision: a modern approach*, Prentice Hall Professional Technical Reference.

FU, W. & PERRY, P. O. 2020. Estimating the number of clusters using cross-validation. *Journal of Computational and Graphical Statistics,* 29**,** 162-173.

GORDON, A. D. 1999. *Classification*, CRC Press.

GOVENDER, P. & SIVAKUMAR, V. 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research,* 11**,** 40-56.

GRANT, D. & YEO, B. 2018. A global perspective on tech investment, financing, and ICT on manufacturing and service industry performance. *International Journal of Information Management,* 43**,** 130-145.

GULERIA, P. & SOOD, M. 2020. Intelligent Data Analysis Using Hadoop Cluster–Inspired MapReduce Framework and Association Rule Mining on Educational Domain. *Intelligent Data Analysis: From Data Gathering to Data Comprehension*.

HAN, J., KAMBER, M. & PEI, J. 10-cluster analysis: Basic concepts and methods. Data mining, 2012. Morgan Kaufmann, 443-495.

HAN, J., PEI, J. & KAMBER, M. 2011. *Data mining: concepts and techniques*, Elsevier.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H. & FRIEDMAN, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, Springer.

HE, Y., WU, Y., QIN, H., HUANG, J. Z. & JIN, Y. 2020. Improved I-nice clustering algorithm based on density peaks mechanism. *Information Sciences,* 548**,** 177-190.

HEDBERG, S. R. 1996. Searching for the mother lode: Tales of the first data miners. *IEEE Expert,* 11**,** 4-7.

HIRECHE, C., DRIAS, H. & MOULAI, H. 2020. Grid based clustering for satisfiability solving. *Applied Soft Computing,* 88**,** 106069.

HU, J., PAN, Y., LI, T. & YANG, Y. 2020. TW-Co-MFC: Two-level weighted collaborative fuzzy clustering based on maximum entropy for multi-view data. *Tsinghua Science and Technology,* 26**,** 185-198.

HU, W., HU, W., XIE, N. & MAYBANK, S. 2009. Unsupervised active learning based on hierarchical graph-theoretic clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* 39**,** 1147-1161.

HUANG, X., YE, Y., XIONG, L., LAU, R. Y., JIANG, N. & WANG, S. 2016. Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences,* 367**,** 1-13.

HUANG, Z. 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining. *DMKD,* 3**,** 34-39.

HUDSON, I. L., KEATLEY, M. R. & LEE, S. Y. 2011. Using Self-Organising Maps (SOMs) to assess synchronies: an application to historical eucalypt flowering records. *International journal of biometeorology,* 55**,** 879-904.

IZADKHAH, H. & TAJGARDAN, M. Information Theoretic Objective Function for Genetic Software Clustering. Multidisciplinary Digital Publishing Institute Proceedings, 2019. 18.

JAIN, A. K., MURTY, M. N. & FLYNN, P. J. 1999. Data clustering: a review. *ACM computing surveys (CSUR),* 31**,** 264-323.

JAKUPOVIĆ, A., PAVLIĆ, M. & POŠČIĆ, P. Business sectors and ERP solutions. Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, 2010. IEEE, 477-482.

JAMALI-DINAN, S.-S., SOLTANIAN-ZADEH, H., BOWYER, S. M., ALMOHRI, H., DEHGHANI, H., ELISEVICH, K. & NAZEM-ZADEH, M.-R. 2020. A Combination of Particle Swarm Optimization and Minkowski Weighted K-Means Clustering: Application in Lateralization of Temporal Lobe Epilepsy. *Brain topography*.

JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2015. An introduction to statistical learning with applications in R (Springer-Verlag New York).

JIANG, D., WU, S., CHEN, G., OOI, B. C., TAN, K.-L. & XU, J. 2016. epiC: an extensible and scalable system for processing Big Data. *The VLDB Journal,* 25**,** 3-26.

JOHNELL, C. & CHEHREGHANI, M. H. 2020. Frank-Wolfe Optimization for Dominant Set Clustering. *arXiv preprint arXiv:2007.11652*.

JOHNS, H., HEARNE, J., BERNHARDT, J. & CHURILOV, L. 2020. Clustering clinical and health care processes using a novel measure of dissimilarity for variable-length sequences of ordinal states. *Statistical Methods in Medical Research***,** 0962280220917174.

JOTHI, N. & NUR'AINI ABDUL RASHIDB, W. H. 2015. Data Mining in Healthcare–A Review. *Procedia Computer Science,* 72**,** 306-313.

KALGOTRA, P., SHARDA, R. & LUSE, A. 2020. Which similarity measure to use in network analysis: Impact of sample size on phi correlation coefficient and Ochiai index. *International Journal of Information Management,* 55**,** 102229.

KAO, J.-H., CHAN, T.-C., LAI, F., LIN, B.-C., SUN, W.-Z., CHANG, K.-W., LEU, F.-Y. & LIN, J.-W. 2017. Spatial analysis and data mining techniques for identifying risk factors of Out-of-Hospital Cardiac Arrest. *International Journal of Information Management,* 37**,** 1528-1538.

KAPLAN, J. M. & WINTHER, R. G. 2013. Prisoners of abstraction? The theory and measure of genetic variation, and the very concept of "race". *Biological theory,* 7**,** 401-412.

KESSIRA, D. & KECHADI, M.-T. Multi-objective Clustering Algorithm with Parallel Games. 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA), 2020. IEEE, 1-7.

KHAMPARIA, A., PANDE, S., GUPTA, D., KHANNA, A. & SANGAIAH, A. K. 2020. Multi-level framework for anomaly detection in social networking. *Library Hi Tech*.

KHANMOHAMMADI, S., ADIBEIG, N. & SHANEHBANDY, S. 2017. An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications: An International Journal,* 67**,** 12-18.

KHOUJA, M. & BOOTH, D. E. 1995. Fuzzy clustering procedure for evaluation and selection of industrial robots. *Journal of Manufacturing Systems,* 14**,** 244-251.

KIANG, M. Y., HU, M. Y. & FISHER, D. M. 2007. The effect of sample size on the extended self-organizing map network—A market segmentation application. *Computational Statistics & Data Analysis,* 51**,** 5940-5948.

KOHLI, S. & MEHROTRA, S. 2016. A clustering approach for optimization of search result. *Journal of Images and Graphics,* 4**,** 63-66.

LAHAT, D., ADALI, T. & JUTTEN, C. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE,* 103**,** 1449-1477.

LAM, D. & WUNSCH, D. C. 2014. Clustering. *Academic Press Library in Signal Processing.* Elsevier.

LANDAU, S., LEESE, M., STAHL, D. & EVERITT, B. S. 2011. *Cluster analysis*, John Wiley & Sons.

LEE, Y.-H., HU, P. J.-H., ZHU, H. & CHEN, H.-W. 2020. Discovering event episodes from sequences of online news articles: A time-adjoining frequent itemset-based clustering method. *Information & Management,* 57**,** 103348.

LELIEVELD, S. H., WIEL, L., VENSELAAR, H., PFUNDT, R., VRIEND, G., VELTMAN, J. A., BRUNNER, H. G., VISSERS, L. E. & GILISSEN, C. 2017. Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes. *The American Journal of Human Genetics,* 101**,** 478-484.

LI, D.-C., DAI, W.-L. & TSENG, W.-T. 2011. A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business. *Expert Systems with Applications,* 38**,** 7186-7191.

LI, J. & WANG, Q. 2021. Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene Settings: overview, challenges, and novel orientation. *Information Fusion*.
LI, Q., WANG, S., ZHAO, C., ZHAO, B., YUE, X. & GENG, J. 2020a. HIBOG: Improving the clustering accuracy by ameliorating dataset with gravitation. *Information Sciences*.
LI, W., FU, L., NIU, B., WU, S. & WOOLEY, J. 2012. Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in bioinformatics,* 13**,** 656-668.
LI, X., LIANG, W., ZHANG, X., QING, S. & CHANG, P.-C. 2020b. A cluster validity evaluation method for dynamically determining the near-optimal number of clusters. *Soft Computing,* 24**,** 9227-9241.
LIAO, T. W. 2005. Clustering of time series data—a survey. *Pattern recognition,* 38**,** 1857-1874.
LISMONT, J., VANTHIENEN, J., BAESENS, B. & LEMAHIEU, W. 2017. Defining analytics maturity indicators: A survey approach. *International Journal of Information Management,* 37**,** 114-124.
LIU, J. & CHEN, Y. 2019. A personalized clustering-based and reliable trust-aware QoS prediction approach for cloud service recommendation in cloud manufacturing. *Knowledge-Based Systems,* 174**,** 43-56.
LIU, Y., JIANG, Y., HOU, T. & LIU, F. 2020. A new robust fuzzy clustering validity index for imbalanced data sets. *Information Sciences,* 547**,** 579-591.
LIU, Y., LI, Z., XIONG, H., GAO, X. & WU, J. Understanding of internal clustering validation measures. 2010 IEEE International Conference on Data Mining, 2010. IEEE, 911-916.
LU, J., GAN, A., HALEEM, K. & WU, W. 2013. Clustering-based roadway segment division for the identification of high-crash locations. *Journal of Transportation Safety & Security,* 5**,** 224-239.
MAHDI, M. A., HOSNY, K. M. & ELHENAWY, I. 2021. Scalable clustering algorithms for big data: a review. *IEEE Access*.
MAI, D. S., NGO, L. T. & HAGRAS, H. 2020. A hybrid interval type-2 semi-supervised possibilistic fuzzy c-means clustering and particle swarm optimization for satellite image analysis. *Information Sciences,* 548**,** 398-422.
MANOGARAN, G. & LOPEZ, D. 2017. A survey of big data architectures and machine learning algorithms in healthcare. *International Journal of Biomedical Engineering and Technology,* 25**,** 182-211.
MARBAC, M., SEDKI, M. & PATIN, T. 2019. Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification***,** 1-19.
MASSON, M.-H. & DENOEUX, T. 2008. ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition,* 41**,** 1384-1397.
MATULA, D. W. 1977. Graph theoretic techniques for cluster analysis algorithms. *Classification and clustering.* Elsevier.
MEHROTRA, S. & KOHLI, S. 2016. Application of clustering for improving search result of a website. *Information Systems Design and Intelligent Applications.* Springer.
MIN, E., GUO, X., LIU, Q., ZHANG, G., CUI, J. & LONG, J. 2018. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access,* 6**,** 39501-39514.
MOTIWALLA, L. F., ALBASHRAWI, M. & KARTAL, H. B. 2019. Uncovering unobserved heterogeneity bias: Measuring mobile banking system success. *International Journal of Information Management,* 49**,** 439-451.

MOTLAGH, O., BERRY, A. & O'NEIL, L. 2019. Clustering of residential electricity customers using load time series. *Applied Energy,* 237**,** 11-24.
MOURER, A., FOREST, F., LEBBAH, M., AZZAG, H. & LACAILLE, J. 2020. Selecting the Number of Clusters $ K $ with a Stability Trade-off: an Internal Validation Criterion. *arXiv preprint arXiv:2006.08530*.
N'CIR, C.-E. B., CLEUZIOU, G. & ESSOUSSI, N. 2015. Overview of overlapping partitional clustering methods. *Partitional Clustering Algorithms.* Springer.
NAGHIEH, E. & PENG, Y. 2009. Microarray Gene Expression Data Mining: Clustering Analysis Review.
NAKAYAMA, H. & KAGAKU, N. 1998. Pattern classification by linear goal programming and its extensions. *Journal of Global Optimization,* 12**,** 111-126.
NEGARA, E. S. & ANDRYANI, R. 2018. A Review on Overlapping and Non-Overlapping Community Detection Algorithms for Social Network Analytics.
NERURKAR, P., SHIRKE, A., CHANDANE, M. & BHIRUD, S. 2018. Empirical analysis of data clustering algorithms. *Procedia Computer Science,* 125**,** 770-779.
NG, A. Y., JORDAN, M. I. & WEISS, Y. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2002. 849-856.
NHAMO, G., NHEMACHENA, C. & NHAMO, S. 2020. Using ICT indicators to measure readiness of countries to implement Industry 4.0 and the SDGs. *Environmental Economics and Policy Studies,* 22**,** 315-337.
NI, J., YOUNG, T., PANDELEA, V., XUE, F. & CAMBRIA, E. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review***,** 1-101.
NIWATTANAKUL, S., SINGTHONGCHAI, J., NAENUDORN, E. & WANAPU, S. Using of Jaccard coefficient for keywords similarity. Proceedings of the international multiconference of engineers and computer scientists, 2013. 380-384.
OGUNDELE, I., POPOOLA, O., OYESOLA, O. & ORIJA, K. 2018. A review on data mining in healthcare. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*.
P CHANDRASEKHARAN, M. & RAJAGOPALAN, R. 1986. An ideal seed non-hierarchical clustering algorithm for cellular manufacturing. *International Journal of Production Research,* 24**,** 451-463.
PAGE, M. J., MCKENZIE, J. E., BOSSUYT, P. M., BOUTRON, I., HOFFMANN, T. C., MULROW, C. D., SHAMSEER, L., TETZLAFF, J. M., AKL, E. A. & BRENNAN, S. E. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery,* 88**,** 105906.
PALANISAMY, V. & THIRUNAVUKARASU, R. 2019. Implications of big data analytics in developing healthcare frameworks–A review. *Journal of King Saud University-Computer and Information Sciences,* 31**,** 415-425.
PANDIT, S. & GUPTA, S. 2011. A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science,* 2**,** 29-31.
PARSONS, L., HAQUE, E. & LIU, H. 2004. Subspace Clustering for High Dimensional Data: A Review. *Dimension,* 1**,** 1.5.
PEDRYCZ, W. 2002. Collaborative fuzzy clustering. *Pattern Recognition Letters,* 23**,** 1675-1686.

PEREIRA, M. M. & FRAZZON, E. M. 2020. A data-driven approach to adaptive synchronization of demand and supply in omni-channel retail supply chains. *International Journal of Information Management*, 102165.
PÉREZ-SUÁREZ, A., MARTÍNEZ-TRINIDAD, J. F. & CARRASCO-OCHOA, J. A. 2019. A review of conceptual clustering algorithms. *Artificial Intelligence Review,* 52, 1267-1296.
PETWAL, S., JOHN, K. S., VIKAS, G. & RAWAT, S. S. 2020. Recommender System for Analyzing Students' Performance Using Data Mining Technique. *Data Science and Security.* Springer.
PIERNIK, M., BRZEZINSKI, D., MORZY, T. & LESNIEWSKA, A. 2015. XML clustering: a review of structural approaches. *The Knowledge Engineering Review,* 30, 297-323.
PIKE, M. & LINTNER, B. R. 2020. Application of clustering algorithms to TRMM precipitation over the tropical and south Pacific Ocean. *Journal of Climate,* 33, 5767-5785.
QIAN, G., SURAL, S., GU, Y. & PRAMANIK, S. Similarity between Euclidean and cosine angle distance for nearest neighbor queries. Proceedings of the 2004 ACM symposium on Applied computing, 2004. 1232-1237.
RABBANI, M., FARROKHI-ASL, H. & ASGARIAN, B. 2017. Solving a bi-objective location routing problem by a NSGA-II combined with clustering approach: application in waste collection problem. *Journal of Industrial Engineering International,* 13, 13-27.
RAI, A., TANG, X., BROWN, P. & KEIL, M. 2006. Assimilation patterns in the use of electronic procurement innovations: A cluster analysis. *Information & Management,* 43, 336-349.
RAMADAN, R. A., ALHAISONI, M. M. & KHEDR, A. Y. 2020. Multiobjective clustering algorithm for complex data in learning management systems. *Complex Adaptive Systems Modeling,* 8, 1-14.
RAPPOPORT, N. & SHAMIR, R. 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *bioRxiv*, 371120.
RASMUSSEN, E. M. 1992. Clustering algorithms. *Information retrieval: data structures & algorithms,* 419, 442.
RATHEE, A. & CHHABRA, J. K. 2018. Clustering for software remodularization by using structural, conceptual and evolutionary features. *J Univers Comput Sci,* 24, 1731-1757.
RAY, S. & TURI, R. H. Determination of number of clusters in k-means clustering and application in colour image segmentation. Proceedings of the 4th international conference on advances in pattern recognition and digital techniques, 1999. Calcutta, India, 137-143.
RHODES, J. D., COLE, W. J., UPSHAW, C. R., EDGAR, T. F. & WEBBER, M. E. 2014. Clustering analysis of residential electricity demand profiles. *Applied Energy,* 135, 461-471.
RODRIGUEZ, M. Z., COMIN, C. H., CASANOVA, D., BRUNO, O. M., AMANCIO, D. R., COSTA, L. D. F. & RODRIGUES, F. A. 2019. Clustering algorithms: A comparative approach. *PLoS ONE,* 14.
RONG, W., ZHUO, E., PENG, H., CHEN, J., WANG, H., HAN, C. & CAI, H. 2020. Learning a consensus affinity matrix for multi-view clustering via subspaces merging on Grassmann manifold. *Information Sciences,* 547, 68-87.
RUSSOM, P. 2011. Big data analytics. *TDWI best practices report, fourth quarter,* 19, 1-34.
SABBAGH, R. & AMERI, F. 2020. A Framework Based on K-Means Clustering and Topic Modeling for Analyzing Unstructured Manufacturing Capability Data. *Journal of Computing and Information Science in Engineering,* 20.

SAMOILENKO, S. & OSEI-BRYSON, K.-M. 2019. Representation matters: An exploration of the socio-economic impacts of ICT-enabled public value in the context of sub-Saharan economies. *International Journal of Information Management,* 49**,** 69-85.
SANDROCK, K. 1988. A simple algorithm for solving small, fixed-charge transportation problems. *Journal of the Operational Research Society,* 39**,** 467-475.
SAXENA, A., PRASAD, M., GUPTA, A., BHARILL, N., PATEL, O. P., TIWARI, A., ER, M. J., DING, W. & LIN, C.-T. 2017. A review of clustering techniques and developments. *Neurocomputing,* 267**,** 664-681.
SCHWENKER, F. & TRENTIN, E. 2014. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters,* 37**,** 4-14.
SCOTT, J. & CARRINGTON, P. J. 2011. *The SAGE handbook of social network analysis*, SAGE publications.
SEKULA, M., DATTA, S. & DATTA, S. 2017. optCluster: An R package for determining the optimal clustering algorithm. *Bioinformation,* 13**,** 101.
SEKULA, M. N. 2015. OptCluster: an R package for determining the optimal clustering algorithm and optimal number of clusters.
SFYRIDIS, A. & AGNOLUCCI, P. 2020. Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *Journal of Transport Geography,* 83.
SHAFQAT, S., KISHWER, S., RASOOL, R. U., QADIR, J., AMJAD, T. & AHMAD, H. F. 2020. Big data analytics enhanced healthcare systems: a review. *The Journal of Supercomputing,* 76**,** 1754-1799.
SHAMIM, G. & RIHAN, M. 2020. Multi-Domain Feature Extraction for Improved Clustering of Smart Meter Data. *Technology and Economics of Smart Grids and Sustainable Energy,* 5**,** 1-8.
SHARGHI, E., NOURANI, V., SOLEIMANI, S. & SADIKOGLU, F. 2018. Application of different clustering approaches to hydroclimatological catchment regionalization in mountainous regions, a case study in Utah State. *Journal of Mountain Science,* 15**,** 461-484.
SHARMA, K. K. & SEAL, A. 2020. Multi-view spectral clustering for uncertain objects. *Information Sciences,* 547**,** 723-745.
SHI, L. 2020. Industrial Symbiosis: Context and Relevance to the Sustainable Development Goals (SDGs). *In:* LEAL FILHO, W., AZUL, A. M., BRANDLI, L., ÖZUYAR, P. G. & WALL, T. (eds.) *Responsible Consumption and Production.* Cham: Springer International Publishing.
SHIAU, W.-L., DWIVEDI, Y. K. & YANG, H. S. 2017. Co-citation and cluster analyses of extant literature on social networks. *International Journal of Information Management,* 37**,** 390-399.
SHIAU, W.-L., YAN, C.-M. & LIN, B.-W. 2019. Exploration into the intellectual structure of mobile information systems. *International Journal of Information Management,* 47**,** 241-251.
SHIRKHORSHIDI, A. S., AGHABOZORGI, S., WAH, T. Y. & HERAWAN, T. Big data clustering: a review. International conference on computational science and its applications, 2014. Springer, 707-720.
SOLORIO-FERNÁNDEZ, S., CARRASCO-OCHOA, J. A. & MARTÍNEZ-TRINIDAD, J. F. 2020. A review of unsupervised feature selection methods. *Artificial Intelligence Review,* 53**,** 907-948.

SONG, Z., WANG, C. & BERGMANN, L. 2020. China's prefectural digital divide: Spatial analysis and multivariate determinants of ICT diffusion. *International Journal of Information Management*, 102072.
SPRAGUE, L. A., OELSNER, G. P. & ARGUE, D. M. 2017. Challenges with secondary use of multi-source water-quality data in the United States. *Water research,* 110, 252-261.
SUBRAMANIYAN, M., SKOOGH, A., MUHAMMAD, A. S., BOKRANTZ, J., JOHANSSON, B. & ROSER, C. 2020. A generic hierarchical clustering approach for detecting bottlenecks in manufacturing. *Journal of Manufacturing Systems,* 55, 143-158.
SUH, J. W., SOHN, S. Y. & LEE, B. K. 2020. Patent clustering and network analyses to explore nuclear waste management technologies. *Energy Policy,* 146, 111794.
TAHMASEBI, P., HEZARKHANI, A. & SAHIMI, M. 2012. Multiple-point geostatistical modeling based on the cross-correlation functions. *Computational Geosciences,* 16, 779-797.
TANOTO, Y., HAGHDADI, N., BRUCE, A. & MACGILL, I. 2020. Clustering based assessment of cost, security and environmental tradeoffs with possible future electricity generation portfolios. *Applied Energy,* 270, 115219.
THAKUR, N., MEHROTRA, D., BANSAL, A. & BALA, M. 2020. Implementation of Quasi-Euclidean Distance-Based Similarity Model for Retrieving Information from OHSUMED Dataset. *Soft Computing: Theories and Applications.* Springer.
TRAN, T. A. 2020. Effect of ship loading on marine diesel engine fuel consumption for bulk carriers based on the fuzzy clustering method. *Ocean Engineering,* 207, 107383.
UPTON, G. & FINGLETON, B. 1985. *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*, John Wiley & Sons Ltd.
USELTON, S., AHRENS, J., BETHEL, W. & TREINISH, L. 1998. Multi-source data analysis challenges. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).
USHAKOV, A. V. & VASILYEV, I. 2020. Near-optimal large-scale k-medoids clustering. *Information Sciences,* 545, 344-362.
VALLS, A., GIBERT, K., ORELLANA, A. & ANTÓN-CLAVÉ, S. 2018. Using ontology-based clustering to understand the push and pull factors for British tourists visiting a Mediterranean coastal destination. *Information & Management,* 55, 145-159.
VIALETTO, G. & NORO, M. 2020. An innovative approach to design cogeneration systems based on big data analysis and use of clustering methods. *Energy Conversion and Management,* 214, 112901.
WANG, Q. & YANG, X. 2020. Investigating the sustainability of renewable energy–An empirical analysis of European Union countries using a hybrid of projection pursuit fuzzy clustering model and accelerated genetic algorithm based on real coding. *Journal of Cleaner Production*, 121940.
WANG, W., YANG, J. & MUNTZ, R. STING: A statistical information grid approach to spatial data mining. VLDB, 1997. 186-195.
WANG, X. & WANG, H. 2020. Driving behavior clustering for hazardous material transportation based on genetic fuzzy C-means algorithm. *IEEE Access,* 8, 11289-11296.
XIE, J., KELLEY, S. & SZYMANSKI, B. K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur),* 45, 1-35.
XIE, W.-B., LEE, Y.-L., WANG, C., CHEN, D.-B. & ZHOU, T. 2020. Hierarchical clustering supported by reciprocal nearest neighbors. *Information Sciences*.
XU, R. & WUNSCH, D. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks,* 16, 645-678.

XU, R. & WUNSCH, D. C. 2010. Clustering algorithms in biomedical research: a review. *IEEE Reviews in Biomedical Engineering,* 3**,** 120-154.

XU, X., QIAN, H., GE, C. & LIN, Z. 2020. Industry classification with online resume big data: A design science approach. *Information & Management,* 57**,** 103182.

YE, J. 2011. Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Mathematical and computer modelling,* 53**,** 91-97.

YIN, L. 2020. Intelligent Clustering Evaluation of Marine Equipment Manufacturing based on Network Connection Strength. *Journal of Coastal Research,* 103.

YOO, I., ALAFAIREET, P., MARINOV, M., PENA-HERNANDEZ, K., GOPIDI, R., CHANG, J.-F. & HUA, L. 2012. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems,* 36**,** 2431-2448.

ZHANG, K., COLLINS, E. G. & BARBU, A. 2013. An efficient stochastic clustering auction for heterogeneous robotic collaborative teams. *Journal of Intelligent & Robotic Systems,* 72**,** 541-558.

ZHANG, X., SUN, Y., LIU, H., HOU, Z., ZHAO, F. & ZHANG, C. 2020. Improved Clustering Algorithms for Image Segmentation Based on Non-local Information and Back Projection. *Information Sciences*.

ZHAO, K., JIANG, Y., XIA, K., ZHOU, L., CHEN, Y., XU, K. & QIAN, P. 2020. View-collaborative fuzzy soft subspace clustering for automatic medical image segmentation. *Multimedia Tools and Applications,* 79**,** 9523-9542.

ZHU, Q., ZHANG, F., LIU, S. & LI, Y. 2019. An anticrime information support system design: Application of K-means-VMD-BiGRU in the city of Chicago. *Information & Management***,** 103247.