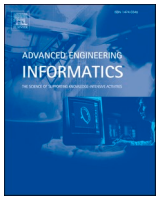
Advanced Engineering Informatics 62 (2024) 102799

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Review article

A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects

Jaswinder Singh *, Damanpreet Singh

Department of Computer Science & Engineering, SLIET Longowal, Sangrur, India

ARTICLE INFO

Keywords:
Machine learning
Unsupervised learning
Data clustering
Pattern recognition
Automatic data clustering
Hybrid clustering

ABSTRACT

Clustering is a set of essential mathematical techniques in artificial intelligence and machine learning for analyzing massive amounts of data generated by applications. Clustering uses data mining approaches to group data points based on their intrinsic characteristics or similarity measures. These measures are essential for data mining tasks such as information retrieval, pattern recognition, classifications, and clustering from the raw data. However, significant efforts are needed to select the appropriate similarity metrics based on the distribution of data points and problem domains. Various clustering approaches and techniques are actively utilized in a variety of disciplines of research, including data science, machine learning, computer science, pattern recognition, computer vision, etc. Some partitional clustering techniques, such as K-Means and density-based clustering, etc., take some sensitive parameters from the user prior to grouping the objects, and variations on these parameters may lead to different results for the same dataset. Some traditional statistical techniques for clustering are unable to find optimal clusters or handle high-dimensional data effectively. The need for a priori requirements of elements, for instance, the number of groups, termination at the local optimum, and expensive computations are some of its limitations that must be solved. To get beyond the aforementioned restrictions, innovative, adaptable, efficient, and hybrid clustering algorithms must be created. This study provides a comprehensive review of the literature on traditional and novel clustering techniques in a cohesive manner, their trending applications in various domains, their summarization, challenges, and future scope. In addition, data clustering embraces various scientific disciplines. Thus, this study will be beneficial and will provide an effective reference point for progressive researchers, analysts, and artificial intelligence professionals to develop novel, flexible, and efficient state-of-the-art clustering techniques.

1. Introduction

Classification and clustering are the fundamental data analysis tasks in data mining and knowledge discovery. Classification is labeled as a supervised machine learning technique that allows a model to learn from labeled datasets and predict the label of unseen data [1], whereas clustering is labeled as an unsupervised machine learning (UML) technique that allows a model to learn from unlabeled datasets and form homogeneous groups of data points called clusters. Data clustering is a widely used data processing technique that groups items into distinct clusters based on some measure of similarity between data points [2]. In clustering, organizing data points into homogeneous groups based on [3,4] similarity and feature values is a fundamental task of research for discovering hidden patterns in data. The three primary uses of data clustering are as follows [5,6]:

- Underline the pattern and insight into the data.
- Identify the degree of similarity between data points.
- Data organization and summarization through cluster prototypes.

The commonly used key words in this study and their definitions are

* Corresponding author.
E-mail addresses: pcs2219_jaswinder@sliet.ac.in (J. Singh), damanpreets@sliet.ac.in (D. Singh).

https://doi.org/10.1016/j.aei.2024.102799
Received 28 March 2024; Received in revised form 21 August 2024; Accepted 29 August 2024
Available online 11 September 2024

listed below for the common understanding of the readers.

Key Terms	Definition
Dataset	A dataset is a collection of data values in the form of 2-dimensional matrix *NXD* where *N* represents individual instance and *D* represents dimensions.
Instance	An instance is a specific row or observation describing a physical object in a dataset. Some other common terminologies are feature vector, data point and pattern.
Dimension	Dimension of a dataset refers to the number of features of the dataset.
Data Mining techniques	Discover meaningful information from a huge volume of data. Data mining techniques (DMT) can use several types of data modeling, such as classification and data clustering.
Machine Learning	It is a subdomain of artificial intelligence with an emphasis on models that can learn from input data. There are several machine learning algorithms for different data mining models.
Supervised Machine Learning	A machine learning model tries to build a relationship between inputs and outputs and predicts output based on similar unseen data.
Unsupervised Machine Learning	A machine learning model that groups similar data points based on their intrinsic characteristics or similarity measures.
Classification	A type of supervised machine learning that predicts the class label of a given data points.
Clustering	A type of unsupervised machine learning that partitions a dataset into homogeneous groups based on similarity measures.

Clustering techniques split a dataset $D = \{X_i\}_{i=1}^{n}$ comprises n-data points of a d-dimensional feature space ($\mathbb{R}^d$) into subsets s_1, s_2, ... s_k in such a way that every data point belongs to only one set as shown in Eq. (1) [1]. The union of all sets creates the original dataset presented in Eq. (2) and data points of different sets constitute clusters [5,7].

$$\bigcap_{i=1}^{k} s_k = \emptyset \tag{1}$$

and

$$\bigcup_{i=1}^{k} s_k = D \tag{2}$$

The main purpose of clustering is to predict groups by splitting the dataset into subsets with similar features. The idea behind data clustering is to maximize intra-cluster similarity while minimizing inter-cluster similarity [8]. A good clustering algorithm will create well-separated clusters of data points with higher intra-cluster similarity but lower inter-cluster similarity [9,10] shown in Fig. 1.

The distance among the data points is the most widely used method for determining how similar the data points are [11]. A variety of data clustering techniques have recently been developed and applied to address this task. Various different criteria are used in the literature to define the similarity metrics among the data points. With the exponential growth of the internet, mobile devices, security cameras and sensing technology, a variety of data is created in high dimensions and volumes on the internet in the form of numbers, texts, audio files, videos, images, time series and multi –view data [12,13]. Now, it is becoming more challenging for researchers to analyse high-dimensional datasets and retrieve hidden information from it [1,7]. As unstructured data is so common nowadays, efficient clustering techniques to extract hidden structure and pattern from large amounts of high-dimensional data are required [3 14]. In order to handle high-dimensional data, subspace clustering approaches have been developed that identify clusters in subspaces (a subset of attributes), unlike traditional algorithms [14]. Cai et al. [15] surveyed an efficient set of feature selection methods for handling redundant and irrelevant features for better clustering results. Clustering is an important data analysis method that is active in various fields [4]. Clustering of objects is widely useful in several disciplines of research such as engineering, technology, medical, marketing analysis, banking, and many day-to-day applications [11 16]. In general, the literature divides cluster analysis approaches into two major categories: hierarchical and partitional. Additionally, the researchers have proposed a number of techniques for each sub-category of hierarchical and partitioning techniques [14]. In these categories, clustering algorithms have sensitivity to the tuning of the starting parameters chosen by the user and produce different results for the same dataset [4,11]. Without good domain knowledge of the dataset, it is challenging to estimate the exact number of clusters (NC) apriori from high-dimensional datasets [17]. Therefore, in the late 1990s, some approaches were developed to automatically estimate the optimal number of partitions from datasets which includes: trial-and-error and artificial intelligence (AI) based clustering methods such as evolutionary and swarm intelligence (SI) [1,4,17]. The advancement of automatic clustering has accelerated in the recent past. In the earliest stage, researchers explored stochastic optimization techniques including genetic algorithms (GAs), tabu search and simulated annealing to address the automatic data clustering. Liu et al. [18] applied GA based approach called automatic genetic clustering (AGCUK) to find unknown NC. In later stages, metaheuristic

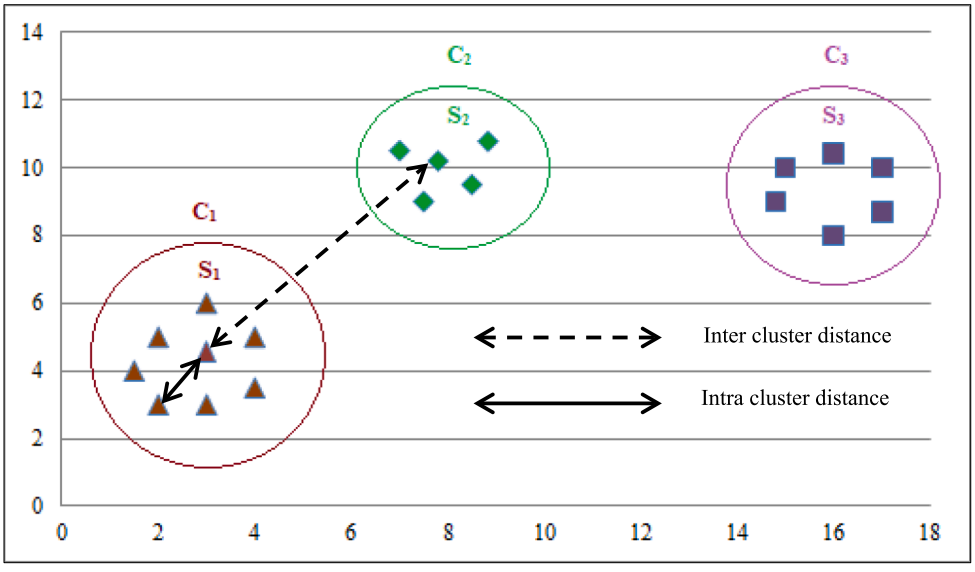


Fig. 1. Euclidean distance-based partition of the dataset into three disjoint sets s_1, s_2 and s_3 represents three clusters c_1, c_2 and c_3 respectively.

algorithms inspired by nature are applied to address automatic data clustering problems [11,17]. These are successfully applied in different fields of science and engineering. Javidan et al. [19] proposed an automatic K-Means (KM) clustering along with multi-class support vector machines (SVM) for the diagnosis of diseased grape leaves and are classified into three classes, namely leaf blight, black rot, and black measles. Some researchers combined traditional clustering algorithms, such as KM with nature inspired metaheuristic techniques for automatic clustering [20].

Undoubtedly, clustering has gathered much attention over the years [21]. Much has been accomplished with regard to clustering, including new upcoming research paths in automatic cluster analysis methods, as can be seen from the literature, which shows a significant increase in multidisciplinary interests and activities in the application of cluster analysis to various research disciplines. Various aspects of clustering algorithms have been covered in the existing surveys [1,4,22,23].

The first and most important stage in conducting a systematic review is to compile a list of research questions. The following is a list of research questions pertaining to clustering methods. As the main contributions of this review, the following findings will be addressed in the subsequent sections:

(i) Broad categories of clustering techniques and different proposed algorithms in those categories.
(ii) Finding out the various possible combinations of hybrid clustering techniques.
(iii) Various techniques of performing clustering without knowing the domain knowledge of datasets.
(iv) Different types of data suitable for clustering.
(v) Challenges involved in clustering and recent developments in clustering.

Reviewing existing clustering techniques is essential for identifying research gaps and limitations, as well as for assessing the strengths and weaknesses of each technique under various conditions and datasets. Clustering techniques are used across various disciplines, such as computer science, biology, finance, marketing, etc. Reviewing techniques from multiple domains facilitates cross-disciplinary insights and encourages the transfer of knowledge and methods between different fields. Also, existing clustering algorithms establish the foundation for combining conventional and AI-based approaches to produce more extensive and refined analysis. Therefore, this study will provide researchers carrying out work in UML field with a systematic comprehensive understanding of the traditional and current state-of-the-art methods as well as an understanding of the vast study topics that can still be explored. This survey aims to provide new researchers with a simple research route by offering a complete review of various data clustering approaches as well as the evolution of clustering techniques throughout time. Also, this survey will assist research community in creating novel algorithms to address new problems in the emerging field of study. The primary contributions of this study are summarized as follows:

- Firstly, it presents an in-depth review of both traditional and cutting-edge clustering techniques.
- Secondly, the concepts of clustering algorithms, architecture, and taxonomy succinctly are covered.
- Thirdly, hybrid clustering approaches have been explored.
- Fourthly, recently developed clustering techniques applied in various disciplines are reviewed.
- Fifthly, a discussion on research challenges in clustering problems is presented.
- Lastly, it enlightens possible future research directions within the scope of our study and trends with a reference to the applicability and use of clustering algorithms in many real world research fields.

The rest of the paper is organized as follows: Section 2 outlines the methodology as well as keywords, academic databases and article inclusion and exclusion. Section 3 compares the previously published review papers. Section 4 discusses the taxonomy of clustering algorithms. Section 5 presents the recent work on clustering. Section 6 describes challenges in clustering. Section 7 represents some useful real world applications of clustering and Section 8 provides a conclusion and future scope.

2. Methodology

This section will include the methodology adopted in conducting this systematic review. This review paper deals with the study of AI based clustering techniques in data mining tasks. In order to find hidden information in a dataset, clustering is an essential technique and has been widely used for so long that different groups of researchers have proposed various clustering algorithms. Finding hidden and meaningful patterns in datasets is undoubtedly an NP-hard problem. Perhaps the challenges inspired the researcher to pursue this objective. An appreciable amount of work has been done and published in various research papers to date. Hence, a significant effort has been made to conduct this comprehensive survey. Also, a substantial amount of review papers, research papers, books and articles have been used.

The search was carried out twice, in distinct phases. During the initial phase of the search total 282 papers were selected whereas the second phase mainly focused on the applications part of the clustering and a total of 84 papers were selected using specific keywords. During the search, we pursued almost every avenue in the search for an answer of the research questions after reading the complete papers. In total, 366 papers were selected to write this comprehensive review paper. The complete flow chart of the methodology used in selecting articles is shown in Fig. 2. Moreover, the distribution of the selected research papers (year-wise) is shown in Fig. 3. In the subsequent sub-sections, the different keywords, academic databases, and criteria of inclusion and exclusion employed in this survey are discussed in the following sequence.

2.1. Keywords for obtaining relevant literature

Keywords are essential for collecting relevant literature for review. A specific set of search keywords is useful to find articles from academic databases that address particular aspects such as research questions or objectives, methodologies, and the scope of a literature study. Therefore, a set of keywords have been used to obtain exhaustive list of research papers, articles, books and surveys as to fulfill the goal of this review. These keywords can be classified into two classes: generic keywords and specific keywords. The generic search keywords include “Unsupervised Learning”, “Clustering”, “Clustering in AI”, “Distance Metrics in clustering”, “Clustering in Machine Learning (ML)”, “Big Data clustering”, “Clustering Survey”, “Approximations in Clustering”, “Variations in Clustering Algorithms”, “Validation Measures”, “Hybrid Clustering” and “Clustering Methods”. The generic search keywords and reasoning for choosing them are summarized below in the form of a list.

- **Generic keywords:**
- Unsupervised Learning: It is one of the broad categories under the umbrella of AI.
- Clustering: A fundamental unsupervised learning technique for knowledge discovery.
- Clustering in AI: It facilitates a comprehensive understanding of the scope of literature review by leveraging AI techniques.
- Distance Metrics in clustering: To retrieve various distance metrics and assess their applicability to the nature of the data and the clustering algorithm.
- Clustering in Machine Learning (ML): Advance clustering techniques within the context of machine learning.

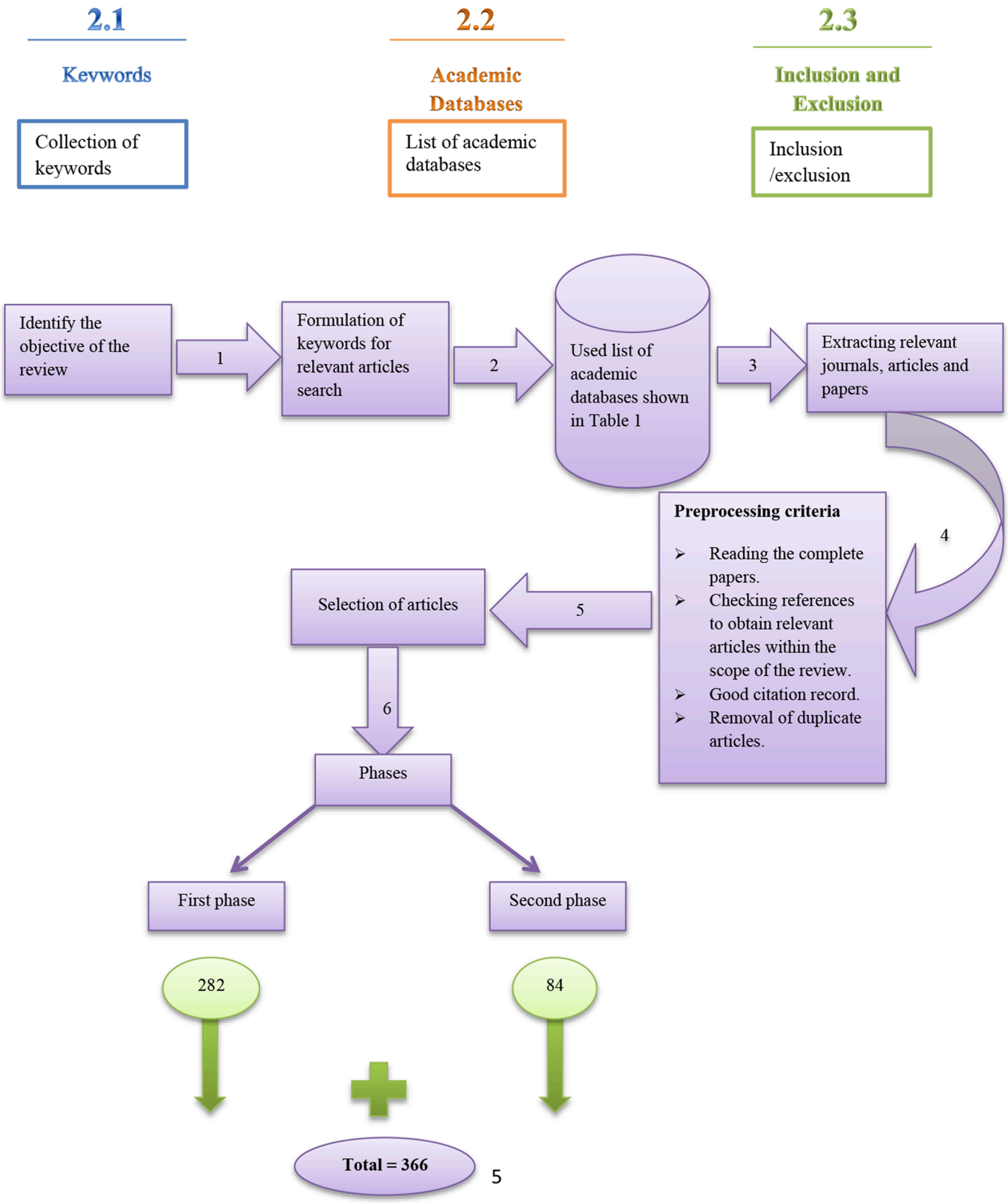


Fig. 2. Flow chart of the methodology used in this review.

- Big Data clustering: Utilization of clustering techniques in big data analytics.
- Clustering Survey: To retrieve existing clustering reviews that serve as foundational knowledge and for comparisons.
- Approximations in Clustering: To include approximation based methods in data clustering.
- Variations in Clustering Algorithms: To retrieve articles based on improvements and variants.

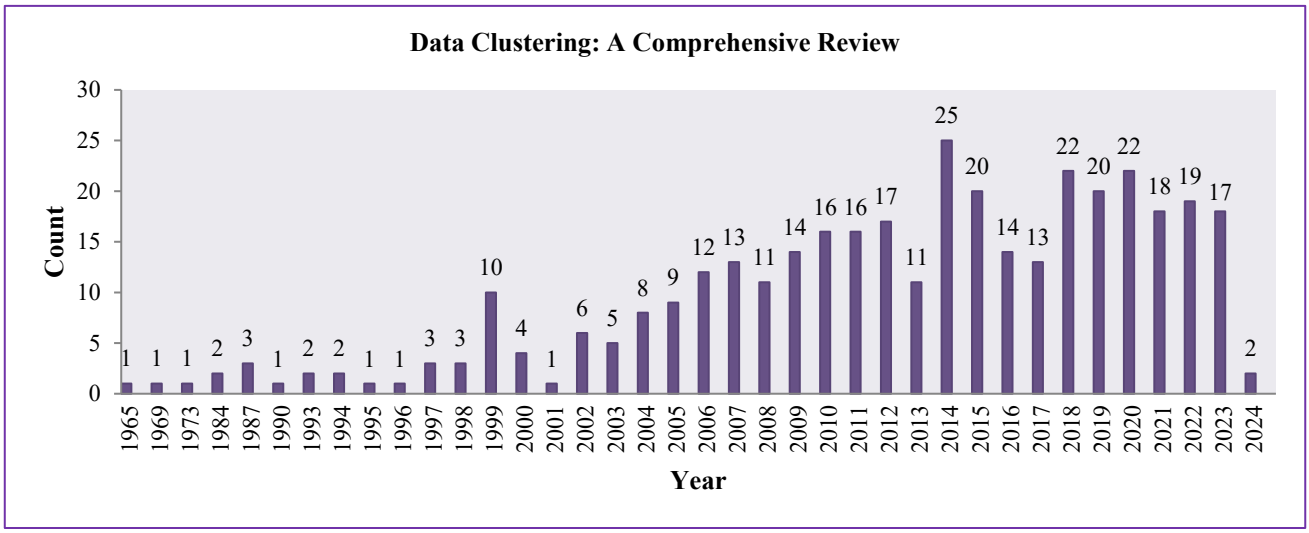


Fig. 3. Distribution of clustering papers over the years.

- Validation Measures: To analyze the output of the clustering algorithms.
- Hybrid Clustering: To address different combinations of various clustering algorithms or techniques in order to improve clustering performance.
- Clustering Methods: To find various methods used in unsupervised learning in order to group similar objects or data points.

Some specific keywords are used to retrieve research papers relevant to various application areas of clustering algorithms.

2.2. Academic databases

The literature was searched on Google Scholar using the developed keywords. This study focused on reliable peer-reviewed journals, conference proceedings and published books that were indexed in ten academic databases. The list of academic databases used throughout the search is shown in Table 1. This review is a collection of publications published from 1999 to the present in different domains of unsupervised learning in clustering data.

2.3. Article *inclusion / exclusion criteria*

To extract the relevant articles, journals, and papers, criteria for inclusion and exclusion have been established. The tabular representation of the parameters for inclusion and exclusion criteria is shown below in Table 2. The selected articles' eligibility was determined by applying each of the standards listed in the same table.

Table 1
The list of academic databases with links to each one.

List of academic databases	Web Link
ACM digital library	https://dl.acm.org/
Applied Sciences	https://www.mdpi.com/
arXiv by Cornell University	https://www.arxiv.org/
IEEE Xplore	https://ieeexplore.ieee.org/
ISI Web of science	https://apps.webofknowledge.com/
Research gate	https://www.researchgate.net/
Science Direct	https://www.sciencedirect.com/
Scopus	https://www.scopus.com/
Springer Link	https://www.link.springer.com/
Wiley online library	https://onlinelibrary.wiley.com/

Table 2
The applied inclusion and exclusion criteria.

Inclusion	Exclusion
The review concentrated on different clustering methods and approaches in data mining.	Publications on general data mining approaches were not taken into consideration.
Articles that studied certain clustering approaches and variants.	Articles on parallel and distributed clustering domains were omitted.
English is the most extensively used language in academic database writing, particularly in topics such as computer science, data science, and related disciplines. Therefore,English-language articles were taken into consideration.	The review did not include any articles written in other languages. Blog posts, PowerPoint presentation slides and keynote addresses were excluded.
Accepted materials, including published papers from reliable journals with peer review, conference proceedings and published books.	

3. Comparison and motivation

This section highlights the significant distinctions between the prior review studies compared to this study. There are a variety of data sources available in the real world, and different structures and patterns exist among them. Clustering is a fairly comprehensive field of study that has found in various applications. As a result, a plethora of clustering methods has been proposed by researchers over time in different domains. There have been numerous attempts in the literature to offer a systematic overview of various clustering techniques and their popular application fields. This study covers a comprehensive list of trending clustering application fields and their references. Therefore, there is still a need to review published articles due to the dynamic nature of the research areas and their contribution to both theoretical and applied fields of study in relation to data mining and knowledge discovery.

The purpose of this study is to present developments in data clustering, including fundamental concepts, existing methodologies, comparisons, and various key applications. Although there are some reviews and surveys that focus on the clustering techniques, their developments, and their applications, none of them tend to be as comprehensive as the one presented in this study. A comparative analysis of existing classification and clustering algorithms is presented in Table 3. Table 4 summarizes a comparative study of this review to other existing reviews, which will aid in evaluating and identifying research challenges based on prior research.

There has been a long and rich history of clustering in many different

Table 3
A comparative analysis of existing classification and clustering algorithms.

S. No.	Publishing date	Remarks	Application problem domains	Dataset type	Advantage	Limitation
1.	This review	This study provides an in-depth literature review on both standard and novel clustering strategies in a cohesive manner, their trending applications in various domains, their summarization, challenges and future scope. In addition to this, this paper presents hybrid clustering strategies, which are novel to the earlier literature reviews.	Data clustering and related areas such asdata mining, web intelligence, intelligent medical science, urban development and privacy protection etc.	Categorical,numerical, text, mixed data types, time series, data stream: web data, sensor data, network data, multiview data and multimedia data.	Covers a broad taxonomical analysis of existing and novel clustering techniques, including a summary, comparison, evaluation, and research progress over time. It also highlights recent developments in cluster analysis, their applications, challenges, and future prospects for creating efficient clustering techniques.	Parallel and distributed clustering techniques are not included.
2.	2024	Presents a systematic and state-of-the-art literature review on hybrid models involving optimization and ML techniques for clustering and classification [24].	Classification,data clustering and optimization.	_	Useful for designing techniques that use optimization and machine learning models.	Discuss simply an overview of hybrid models for classification and clustering.
3.	2023	This study provides an overview and taxonomy of the KM clustering algorithm and its variants. It also covers challenges, limitations, and future perspectives [25].	Image processing and recognition, market analysis, data processing, segmentation of medical images, riskevaluation, tumor detection, medical services etc.	Numerical andcategorical.	Various variants related to design and implementation point of view are presented.	Study presents only KM clustering algorithm and its variants.
4.	2022	Discussed a comprehensive and systematic analysis of the taxonomy, challenges, similarity measures, clustering validation measures, and various application domains for clustering algorithms in AI and ML [4].	Data clustering and related domains such as data mining, pattern recognition, security and information retrieval etc.	Categorical,numerical, mixed data type, time series, streaming and multiview.	• Helpful to design improved and efficient state-of-the-art clustering algorithms. • Presents approaches for automatic data clustering.	Missing hybrid approaches for data clustering.
5.	2021	The thirty-two density-based clustering algorithms that have been presented since 1996 are comprehensively reviewed in this study, along with their four classification categories and five application domains [26].	Earth sciences, molecular biology, astronomy, geography and multimedia.	Spatial, non spatial and multimedia.	Useful to discover clusters in an arbitrary sizes, shaped and densities.	Presents only an empirical study of the density-based clustering methods.
6.	2020	This study presented an updated review of nature-inspired metaheuristic clustering techniques applied to automatic clustering. Also, analyzed the efficiency of traditional and novel metaheuristic algorithms based on various datasets [11].	Marketing, Biology, library, insurance and medicine etc.	Numerical.	Provides computational results on different datasets.	Selected metaheuristic algorithms are reviewed.
7.	2020	Introduced a study on the selection of an optimal subset of features for clustering as well as their summarization in four categories along with ideas, results, and limitations. The authors also reviewed potential future trends and difficulties in this field, including scalability, algorithmic speed, and potent evaluation tools for feature selection for clustering [27]	Clustering, feature selection, Data mining, Evolutionary computation.	_	Feature selection methods for evolutionary clustering.	Noise and outliers detection methods are not highlighted.
8.	2018	This study summarised multiview clustering methods including their taxonomy, the principals involved, and their applications [28].	Useful to extract information from multiview datasets using existing algorithms.	Multiview.	Classifies and summarizes multiview clustering.	Study focuses on multiview clustering.
9.	2017	This paper discussed several clustering techniques, technological advancements over time, and their advantages and limitations. This paper also discusses the fundamentals of clustering, such as similarity	Data clustering and related application domains: Image segmentation, business and object and character recognition etc.	Numerical and categorical.	Provides comparative analysis of clustering in terms of time complexity, scalability and types of datasets.	• Includes the most basic taxonomy of clustering techniques. • Presents a limited number of clustering applications.

(*continued on next page*)

Table 3 (*continued*)

S. No.	Publishing date	Remarks	Application problem domains	Dataset type	Advantage	Limitation
		measures as well as evaluation criteria [1].				
10.	2016	This review offered a three-category knowledge discovery paradigm in data mining approaches for multi-objective optimization problems [29].	Knowledge discovery from multi-objective optimization datasets.	Categorical and numerical.	Useful to extract knowledge from discrete nature of variable.	Approaches are from the domain of exploratory data analysis.
11.	2015	This survey provides an updated overview of all significant nature-inspired metaheuristic algorithms that were previously applied for automatic clustering. Encoding schemes, validity indices, and proximity measurements are just a few of the key elements that go into creating metaheuristics for automatic clustering. 65 automatic clustering methods with 3 %, 69 %, and 28 % employment of single-solution, single-objective, and multi-objective metaheuristics respectively are examined [17].	Automatic data clustering, Single and multi-objective optimization.	Numerical.	Cluster validation and similarity measurements are explored.	• The study covers only nature-inspired metaheuristic-based approaches for data clustering. • Missing comparative analysis to show algorithms superiority.
12.	2015	This comprehensive survey presented the nine categories of classic algorithms, which comprise 26 algorithms, and the ten categories of new approaches, which comprise 45 clustering algorithms, their elements and comparisons were discussed [22].	Clustering algorithm and clustering analysis.	Spatial and data stream.	• Covers traditional clustering categories and their methods. • Presents time complexity of clustering algorithms.	• Presents only fundamental concepts of the frequently used clustering algorithm. • Clustering applications are not addressed.
13.	2006	A novel framework with its future direction for data-dependent geometric regularization has been presented that serves as the foundation for a number of unsupervised, semi-supervised and supervised learning techniques. Conducted studies on a hypothetical dataset and three real-world classification issues involving speech, image, and text categorization [30].	Machine learning including supervised and unsupervised	Text.	Presents the concept of regularization in ML.	Presents basic framework for data-dependent geometric regularization in unsupervised learning.
14.	1999	Pattern representation, similarity measures, clustering stages were all represented by the author along with statistical, fuzzy, evolutionary, neural, and knowledge-based approaches to clustering with brief taxonomy. Some uses of clustering were discussed in the study such as: (i) Image segmentation (ii) Object recognition (iii) Document retrieval and (iv) Data mining [31].	Clustering based image segmentation, object and character recognition, information retrieval and data mining.	Numerical, categorical text and image.	• Addresses different component involved in clustering task. • Study includes statistical, fuzzy, evolutionary, and knowledge-based techniques to cluster analysis.	• Covered clustering methods with regards to statistical perspective. • Clustering validation measures are not addressed.

scientific domains. This research is ongoing to develop novel clustering techniques to facilitate complex clustering procedures [13]. Some of the previously published survey papers cover only limited classifications of clustering methods. Moreover, some review papers cover only specific categories of clustering methods such as KM and density-based clustering approaches. For example A. K. Jain [3] summarized only the most popular clustering algorithms, such as KM, feature selection, and challenges and future perspectives of clustering in the last fifty years. Similarly, Ikotun et al. [25] presented the taxonomy of KM algorithm, its variants and developments with time. Celebi et al. [5] presented a comparative study on initialization techniques for KM with a focus on their computational efficiency. Bhattacharjee et al. [26] presented an extensive review of various density-based clustering methods with their features, properties, and challenges. Fahad et al. [12] presented a review of big data clustering algorithms; taxonomy and empirical analysis; but covered limited taxonomy and validation metrics discussed. Clustering in high-dimensional dataset is a bit challenging tasks. Mittal et al. [14] presented survey on clustering high-dimensional datasets. Aggarwal et al. [32] conducted survey on text clustering. Hruschk et al. [33] presented a review of evolutionary algorithms created for clustering applications. Similarly, Liao [34] and Nanda et al. [35] presented survey on time series and nature inspired metaheuristics clustering techniques respectively. Due to the subjective nature of the clustering methods, a single method is not capable of extracting patterns from all types of

Table 4
Comparative analysis of the current review with other relevant reviews.

Reviews	This	1	2	3	4	5	6	7	8	9	10	11	12	13
Year of Publication	**Review**	**2024**	**2023**	**2022**	**2021**	**2020**	**2020**	**2018**	**2017**	**2016**	**2015**	**2015**	**2006**	**1999**
• Motivation for research	✓	x	x	✓	×	✓	×	×	×	×	×	×	×	✓
• Search Keywords	✓	✓	✓	✓	×	×	×	×	×	×	×	×	×	✓
• Academic Databases	✓	✓	✓	✓	×	×	×	×	×	×	×	×	×	×
• Inclusion/Exclusion	✓	✓	✓	✓	×	×	×	×	×	×	×	×	×	×
• Year-wise Distribution	✓	✓	x	×	×	×	×	×	×	×	×	×	×	×
• Comparative Analysis	✓	✓	✓	✓	✓	✓	×	×	×	×	×	✓	×	✓
• Taxonomy	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	×	×	×	✓
• Hybrid Approach	✓	✓	✓	×	×	✓	✓	×	×	×	×	×	×	×
• Recent Work	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	×	×
• Gaps and Challenges	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	✓
• Evaluation Metrics	✓	x	x	✓	×	×	×	×	✓	✓	×	✓	×	×
• Future Directions	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×

datasets [36]. In addition to evaluating the clustering procedures, this study identifies key parametric characteristics, which benefits the already accomplished work and lays the path for in-depth future research in this area. Furthermore, the clustering method may be hybridized with a number of different clustering algorithms to enhance its overall performance.

Regardless of academics' efforts to provide a comprehensive summary of both conventional and novel clustering algorithms, none of the earlier works reviewed hybrid clustering approaches [25]. It is clear from the literature survey that, in order to comprehend the complete

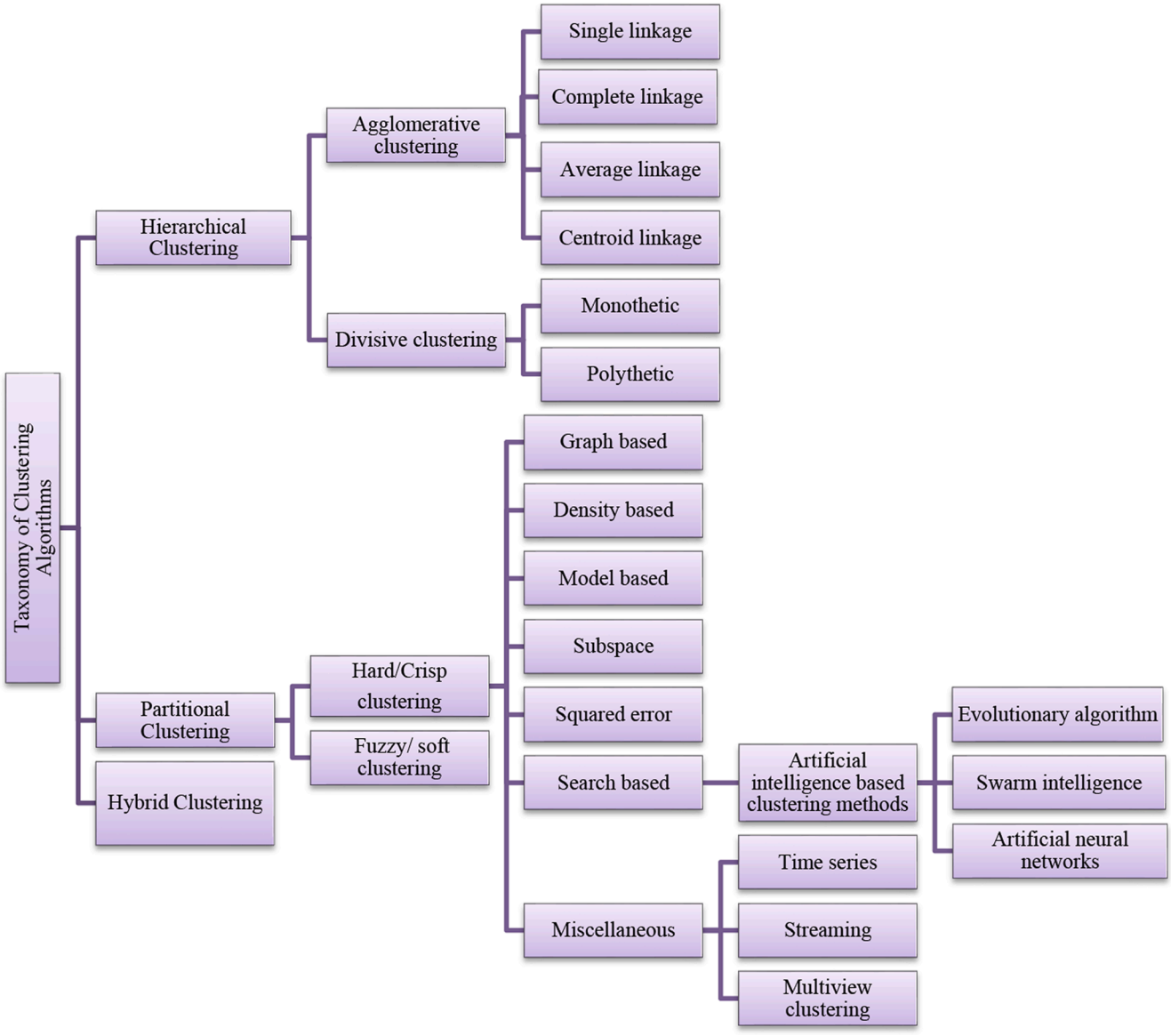


Fig. 4. Taxonomy of clustering algorithms.

taxonomy of clustering, researchers have to refer to different sources. There is a requirement for a comprehensive, systematic review of the literature with regards to traditional and more recently proposed clustering approaches that have been applied in various fields because of the development of technology over time and the intersection of several interdisciplinary domains. Therefore, we were motivated to write a comprehensive and systematic survey on clustering techniques that connect all the categorized reviews for the reader's complete understanding. This study presents the review of innovative, adaptable, efficient, and hybrid clustering algorithms existing in the literature.

4. Classification of clustering algorithms

There are several clustering paradigms that offer diverse cluster models and various algorithmic techniques for cluster discovery [37]. This section presents and discusses the classification of various clustering techniques identified in the literature review. Most of the studies generally classified clustering techniques into two groups: hierarchical and partitional clustering [1–5,11,31,38,39]. There are two types of hierarchical clustering: agglomerative and divisive. Partitional clustering is divided into two categories: hard/crisp clustering and fuzzy/soft clustering. The entire taxonomy of clustering techniques is depicted in Fig. 4.

Some of the popular clustering techniques require human input on the NC, including KM. It is usually challenging for humans to compute an accurate number of partitions in a dataset. In most cases, the genetic algorithms (GAs) based approaches automatically estimate the number of partitions by incorporation good selection criteria [20]. Most of the researches combined standard clustering algorithms with AI based clustering methods for better clustering results termed as hybrid clustering [20,40].

4.1. Hierarchical clustering algorithms

Hierarchical clustering, or hierarchical cluster analysis (HCA), comes under the category of unsupervised learning techniques that are used to better analyze the complex structure of the data. Hierarchical clustering algorithms can provide various consistent partitions of the data at different levels for the same data without running clustering again [2]. To choose the correct threshold for each clustering result, a human expertise is frequently needed because all hierarchical clustering algorithms are quite sensitive to cutoff selection [41].

Hierarchical clustering divides data objects into hierarchical groups. A top-down or bottom-up strategy is used to iteratively build clusters, and the resulting dendrogram shows the hierarchical structure of the formulated clusters [1,4]. Hierarchical clustering is further classified into two classes [42,43] (1) Agglomerative clustering algorithm (Bottom-up approach) and (2) Divisive clustering algorithm (Top down approach). Out of those two, latter is more popular [44,45]. To calculate how similar two points are, a distance (proximity) metric and linkages are required [46]. In hierarchical algorithms, different linkages are used to combine clusters. Single, complete, average and centroid linkages are common ones [1,4,45,47]. Another popular linkage is Ward which examines multivariate Euclidean space for clusters [48].

Single-linkage: The single-linkage clustering can also be referred to as the nearest neighbor, connectedness method, minimum, strongest links first or SLINK [4,43,49]. The proximity can be defined as the smallest dissimilarity between two clusters. According to single linkage, nearest data points of two clusters are merged in each step on the basis of minimum distance. Therefore, the group's regular size and distribution cannot be guaranteed.

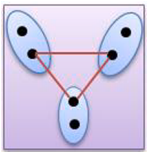
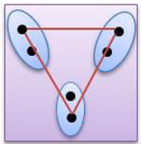
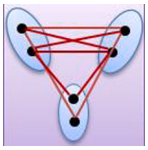
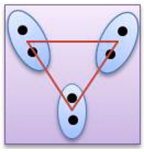
Complete linkage: Complete linkage is the reverse of single linkage. It is often referred to as farthest neighbor or maximum diameter. Unlike single linkage clustering, it computes the largest dissimilarity between two data points of different clusters. According to complete linkage, two clusters with the smallest complete linkage distance are merged in each step. This linkage is less sensitive to noise and produces compact clusters.

Average linkage: Average linkage lies between the maximum and minimum distance. It is commonly referred to as the minimum variance linkage. It minimizes the mean distance between every pair of cluster observations. According to the average linkage, two clusters with smallest average linkage distance are merged in each step. The ability to minimize within-cluster variance and maximize between-cluster variance makes average linkage preferable than single and complete linkage.

Centroid linkage: Centroid linkage refers to the distance between two clusters' centers of gravity (centroids). The centroid is updated when an object is added or removed. According to centroid linkage, two clusters with the smallest centroid linkage distance are merged in each step. When dealing with clusters of various sizes, this linking approach typically outperforms others since it is more resilient to outliers.

When calculating inter-cluster distances, the four previously mentioned proximity metrics consider each point in a pair of clusters. It turns out that the definition of the distance is where these techniques differ. Suppose first cluster prototype $C_a = \{x_1, x_2, \cdots x_m\}$ and second cluster prototype $C_b = \{y_1, y_2, \cdots y_n\}$ are the sets of instances assigned to C_a and C_b respectively. Also, let $d\left(x_i, y_j\right)$ represents the distance between x_i and y_j instance where $i = [1, m]$ and $j = [1, n]$. Accordingly, Table 5 displays the mathematical form of different linkage techniques, their visualization and corresponding mathematical function based on C_a and C_b .

Table 5
Linkage criteria and mathematical distance function.

Linkage Criteria	Mathematical distance function	Description
Single linkage	$D(C_a, C_b) = \min_{ij} d\left(x_i, y_j\right)$ 	This technique joins two clusters with the smallest member distance.
Complete linkage	$D(C_a, C_b) = \max_{ij} d\left(x_i, y_j\right)$ 	This technique joins two clusters with the longest member distance.
Average linkage	$D(C_a, C_b) = \frac{1}{\lvert C_a \rvert \lvert C_b \rvert} \sum_{i=1}^{m} \sum_{j=1}^{n} d\left(x_i, y_j\right)$ 	This technique joins two clusters with the smallest average member distance.
Centroid linkage	$D(C_a, C_b) = d\left(\frac{\sum_{x \varepsilon C_a} x}{m}, \frac{\sum_{y \varepsilon C_b} y}{n}\right)$ 	This technique joins two clusters with the smallest centroid member distance.

4.1.1. Agglomerative hierarchical clustering

Agglomerative hierarchical clustering (AHC) is also referred to as the bottom-up approach. In this approach, every single data point is first assigned to a separate cluster and form a singleton set. The distance between each cluster is calculated using specific distance metric. Next, using a specific linkage, we start merging the clusters that are the most similar (nearest) until all the data points are joined into a single group. The single, complete, average and centroid linkage techniques of the AHC algorithm have been shown to have computational complexity of $O(N^2)$ where N represents number of objects [23,45,49].

The most crucial problem with agglomerative hierarchical clustering is determining how similar the nodes that will be merged are to one another. By creating several types of similarity measures, such as distance-based, closest neighbor-based, quality-based on the partition after merging, and other similarity measures, a variety of agglomerative hierarchical approaches have been established.

The first version of this technique was proposed in 1948 by the Danish botanist Srensen. A mature hierarchical clustering technique has been established after years of exploration and development by several researchers. Users are not required to re-strict a priori properties, such as the sample's distribution functions and NC, in advance [44]. The unique benefit of hierarchical clustering is that, unlike other methods, it arranges all the data into a dendrogram that may maintain the whole hierarchy and clustering relationships between the data in addition to providing the final grouping results. Users are given the option to choose the relevant groupings of characteristics from the dendrogram based on various criteria, and the relationships between the groups are preserved. Several splitting techniques have been developed in the past such as monothetic [47]. However, there are no specific rules for trimming dendrogram. They must be created in accordance with certain specifications [44].

In some scenarios, different linking techniques can produce very different results even with the same metric. Therefore, the characteristics of the study object and the metric design must be taken into consideration while choosing the specific linkage technique [45]. As an object is no longer taken into account after being assigned to a cluster, hierarchical clustering algorithms are unable to rectify any potential prior misclassification. Most hierarchical clustering algorithms have a computational overhead of at least $O(N^2)$, which prevents them from being applied to large datasets.

Hierarchical clustering is less sensitive to noise and outliers in the datasets. Along with some advantages, hierarchical clustering has some limitations associated with it. In terms of time and memory, it is computationally expensive, especially in larger datasets. In order to maintain the benefits of the above or lessen the impact of the aforementioned drawbacks, numerous hierarchical clustering methods have been developed. Many novel HCA techniques have emerged in recent years as a result of the need to handle enormous datasets in data mining and other domains, considerably enhancing clustering performance. Clustering Large Applications based on RANdomized Search (CLARANS), Clustering Using Representatives (CURE), RObust Clustering using linKs (ROCK), Chameleon, and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) are the common examples of AHC methods.

Ng et al. [50] proposed a novel clustering method called CLARANS based on randomized search whose objective is to discover spatial structure present in the data. CLARANS is more effective than Partition around medoids (PAM) and Clustering large applications (CLARA), and can handle large datasets [51]. The time complexity of CLARANS is quadratic in total performance.

Guha et al. [52] proposed a novel clustering technique known as CURE that combines random sampling with partitioning clustering to handle large databases. The proposed method is more robust to outliers, scalable and identifies clusters with various shapes and sizes. To do this, CURE generates a defined number of well-scattered points from each cluster then shrinks those points towards the cluster's center by a specified fraction. These points are then used to represent each cluster. The time complexity of CURE is $O\left(N_{sample}^2 \log\left(N_{sample}\right)\right)$ and space complexity is $O\left(N_{sample}\right)$ where N_{sample} is the number of samples used [53].

Guha et al. [54] proposed a novel clustering algorithm called ROCK which produces good quality clusters with scalability on categorical attributes. It is a type of AHC [55] that relies on the number of links between every pair of points. An enhancement of CURE for handling enumeration-type data, ROCK takes into account the influence of the data around the cluster on the similarity of the data. The time complexity of ROCK is $O\left(n^2 + nm_m m_a + n^2 \log n\right)$, where n denotes the total number of objects, m_a represents average number of neighbors and m_m is the maximum number of neighbors.

Karypis et al. [56] proposed an algorithm for clustering known as Chameleon. The proposed algorithm is based on AHC which focuses on degree of interconnectivity and closeness to merge the data points unlike other methods. The algorithm finds clusters in two phases. Chameleon separates the original data into smaller clusters based on the nearest neighbor graph at first, merging the smaller clusters into larger clusters based on an agglomerative algorithm until satisfied. The time complexity of the proposed algorithm is $O\left(N^2\right)$.

Zhang et al. [57] developed BIRCH, a clustering algorithm suitable for large datasets. The proposed method is relied on clustering features (CF) and CF tree. CF is a tuple of (*N, LS, SS*), where *N* is the number of data points in the cluster, *LS* is the linear sum of *N* data points and *SS* is the square sum of *N* data points. By building the feature tree of clustering, or CF tree, in which one node represents a subcluster, BIRCH realizes the clustering result. The CF tree expands dynamically when a new data point is added to it. Experimental results indicate that BIRCH outperforms CLARANS in terms of quality, speed, and order-sensitivity on a number of big databases. The capacity to handle big datasets and resilience to outliers are the two key motivations behind BIRCH. The computational overhead of BIRCH is $O(N)$ [23]. However, the effectiveness of BIRCH depends on proper parameter setting. The BIRCH clustering method can provide high-quality clustering at a lesser computational cost than CURE, ROCK and Chameleon. The BIRCH algorithm includes several advantages such as: (i) using an incremental clustering technique, creates a tree known as a CF tree from a single scan of the dataset. (ii) Able to efficiently manage noise. (iii) BIRCH is memory-efficient because it saves only a limited number of abstracted data points rather than the entire dataset [58].

4.1.2. Divisive hierarchical clustering

Divisive hierarchical clustering (DHC) follows top-down approach. It is the reverse operation of agglomerative clustering. In aggregative clustering, each point is first assigned to its own cluster, but in divisive clustering, the whole dataset is initially assigned to one cluster, which is then divided into further clusters until each point is assigned to its own cluster. There are two primary division of DHC techniques: monothetic and polythetic [59,60]. When a group of logical qualities, each of which relates to a single variable, are both necessary and sufficient for membership in the group, the cluster is said to be monothetic. Typically, monothetic divisive clustering techniques are variations of the association analysis technique. Divisive approaches need more computational resources than agglomerative methods do. In DHC, global optimal result can be obtained after considering $2^{n-1} - 1$ bipartition where n is the number of objects [61]. Some implementations of DHC are MONA and DIANA [23].

4.1.3. A few approaches to enhance efficiency of hierarchical clustering

Hierarchical clustering requires more time and storage than partitioned clustering, which restricts its use in some specific situations, particularly for large-scale datasets. Thus, several researches have been carried out to increase the efficiency of HCA [2]. The methods for

increasing the effectiveness of hierarchical clustering are described in this section; some of them lower the memory needs while other focuses on both time and space complexity.

Franti et al. [62] proposed a fast agglomerative clustering method that utilizes approximate K-nearest neighbor graph to improve the time complexity. The time complexity improved from $O(\tau N^2)$ to $O(\tau N^* \log N)$ where τ denotes the number of nearest neighbors. Data must be transferred from primary storage to secondary storage as and when required since data mining applications often contain a high number of patterns and attributes that cannot be retained in main memory. Therefore, there is a need to develop efficient clustering algorithm for large datasets. Vijaya et al. [53] developed a Leaders-Subleaders algorithm for handling large datasets of numerical, sequence, text and web-documents types. It is an extension of a previously proposed leader algorithm. Leaders-Subleaders algorithm uses a two-level clustering algorithm that required just two database scans to locate the subgroups or subclusters inside each cluster and can provide a hierarchical structure with the necessary number of levels at a low computation complexity needed in some applications. Jeon et al. [45] proposed a novel linkage method labeled as NC-link for hierarchical clustering. The proposed method displays just linear space complexity while maintaining a quadratic nature in terms of time complexity. Compared to average and complete linkages, NC-link required just 0.7 %-1.75 % of the RAM, and the NC-link-based solution outperformed centroid and Ward's linkages by almost 3.5 times. This allowed us to perform hierarchical clustering much more efficiently and fast than prior approaches. The suggested technique was able to extract hierarchical structures from input data as accurately as the commonly used average and centroid linkage methods in terms of clustering quality.

For analyzing numerical data, a variety of clustering algorithms were developed. However, categorical data is used in a variety of practical applications, including market basket data analysis, DNA or protein sequence analysis, and text mining. Clustering of numerical data is easier than clustering of categorical data [63]. The amount of categorical data produced by practical applications is enormous. As a result, grouping categorical data is challenging since there is no intrinsic similarity metric between category objects [64]. To handle the clustering of categorical data, Xiong et al. [55] proposed a novel algorithm named Divisive hierarchical clustering of categorical data (DHCC) which is effective, efficient and systematic algorithm. It contains two independent phases: preliminary splitting and refinement. Multiple correspondence analysis (MCA) is used for primarily splitting and refinement phase improves the quality of bisection by relocating the objects. Clustering algorithm should be scalable and has ability to handle variety of data. Therefore, another type of a divisive monothetic hierarchical clustering method developed by Chavent et al. [60] called DIVCLUS-T for either numeric or categorical data types. A DIVisive CLUStering-T (DIVCLUS-T) produces CLUStering-Tree as an output. The automatic divisive hierarchical clustering approach (DIVFRP) created by Zhong et al. [61] based on the furthest reference points, does not require any user-defined parameters, not even a cluster validation index. The suggested approach has a lower computing cost than conventional divisive algorithms. Additionally, the existence of outliers does not reduce the quality of the findings from clustering since DIVFRP may remove the outliers bit by bit.

In a nutshell, hierarchical clustering is a cluster analysis approach that group similar data points based on some measure of similarity or distance. It generates a tree-like structure known as a "Dendogram" that reflects the sequence of data point merging. The two primary techniques for hierarchical clustering are agglomerative and divisive, which follow bottom-up and top-down approaches, respectively. These techniques have been shown to have various limitations, including high computation requirements for large datasets. Over time, several novel approaches have been proposed to address these limitations.

4.2. Partition-based clustering

Partitional clustering is also known as iterative relocation algorithm [65]. KM method, which was published in 1957, is where the idea of partitional clustering first emerged [35]. Since then, several studies of conventional partitional clustering techniques have been made [35]. The objective of partitional clustering is to concurrently divide the dataset into homogeneous groups without creating a hierarchical structure [66]. Partitional clustering attempts to directly partition the dataset into a number of distinct groups that optimize a specified criteria function [65,67]. The criterion function is optimized to iteratively partition the entire dataset into a predetermined k number of groups. The most popular partitional clustering approach is based on a criterion function called squared error. The main purpose is to determine the division with the minimum square error for a certain NC [4].

In partitional clustering methods, for instance Fuzzy C-Means (FCM), a datum is allocated a membership value ranges from 0 to 1 to reflect its membership value to a cluster rather than being assigned to a single cluster exclusively [68]. Partitional clustering techniques, also known as prototype-based clustering methods, often assume that the dataset may be represented by a collection of prototypes. Prototype-based clustering's primary objective is to compress and summarize data using cluster prototypes, producing a clear description of the original dataset and a useful division of the dataset. The main differences between partitional clustering algorithms are how prototypes form and how they search for the best clusters and prototypes based on specific criteria.

The prototype-based clustering methods can be broadly divided into two groups based on various definitions of prototypes: point-prototype-based and prototype-based clustering algorithms using non-point prototypes, such as line, hyperplane and hypersphere, which are commonly referred to as non-point-prototype-based clustering algorithms. The most often researched clustering technique in the literature is point-prototype-based clustering, which operates on the presumption that each cluster may be represented by a point in the feature space. The well-known KM clustering, commonly referred to as the extended Lloyd method, is a common example of point-prototype-based clustering. The prototype-based clustering algorithm's key advantages are that it is easy to build and that it converges quickly. Nevertheless, sometimes it's challenging to create a cluster prototype that truly depicts the core structure of the dataset [66].

The partitional clustering algorithm can be classified according to the strategies used to generate the clusters, and the characteristics of the resulting clusters. This includes Hard/Crisp clustering and Fuzzy/Soft clustering. The next subsections represent the different techniques used in each category.

4.2.1. Hard/crisp clustering

In hard/crisp clustering methods, each data point is exactly allocated to a single cluster. Therefore, hard clustering is also known as exclusive clustering [69,70]. Graph-theoretic clustering, Density-based clustering, Subspace clustering, Model-based clustering, and miscellaneous clustering are some examples of the clustering approaches included in this category.

4.2.1.1. Graph theoretic clustering. A graph G: [V, E] is a non-linear organization of data consists of vertices and edges sets. The graph theoretic clustering approach uses graphs to represent clusters. A dataset is translated into a graph in such a way that each object is expressed by a vertex, and edge weight represents the similarity between the vertices [71]. A graph-based clustering approach that is effective when dealing with data that violates a Gaussian or spherical distribution. Without requiring any additional parameters or a certain NC, it may be used to find clusters of any shape or size [72]. A family of graph-theoretical methods based on the Minimal spanning tree (MST) is able to identify various types of cluster structure in arbitrary point sets. Foggia et al.

[72] proposed a graph theoretical clustering method known as (FCM MST clustering algorithm – FMC) for capable of detecting well separated variable shapes clusters.

4.2.1.2. ***Density****-based* ***clustering****.* Density-based clustering is a non-parametric approach that uses the idea of density to discover clusters of various forms, sizes and densities [73]. Density-based clustering methods such as DBSCAN are traditional and popular clustering method [73–75] to extract hidden patterns from datasets by separate high and low density regions based on the neighborhood information [76]. Usually, data objects found in low-density regions are regarded as noise or outliers [74,77]. There are typically two key phases in density-based approaches. In the first step, based on the local neighborhood information the density of each data point is calculated. After that, similar data points in denser regions are identified and combined them to build clusters. DBSCAN clustering algorithm has $O(n(\log n)d)$ time complexity where n is the total number of patterns and d is the dimensionality of the patterns [53]. The DBSCAN algorithm accepts two user input parameters:

(i) ***Epsilon (Eps)*****:** Epsilon represents the maximum radius of a point in a dataset to measure the density.
(ii) ***Minpts*****:** *Minpts* represents the minimum number of points needed inside a circle of radius *Eps* distance for that data point to be categorized as a core point [78] as shown in Fig. 5.

The DBSCAN algorithm divides all points in a dataset into core or border points, noise, or outliers based on the input parameters, which are discussed below.

- Core point: The point at which the number of neighbors must be more than or equal to *Minpts*.
- Boarder point: The point at which the number of neighbors less than *Minpts*.
- Noise/outlier point: A noise or outlier point is a data point that fails to meet the criteria to be a core or border point.

DBSCAN successfully handles outliers and can identify clusters of any shape. However, it is reported to have a number of issues, such as: (i) User inputs (ii) Computational complexity [79]. (iii) Unable to detect density variation [73]. Due to the significant parameter sensitivity of clustering algorithms such as DBSCAN, meticulous manual parameter adjustment is required to provide a useful result [41].

Numerous researchers have attempted to enhance the basic DBSCAN algorithm in an attempt at resolving these issues. Therefore, different variations of DBSCAN algorithm exist in the literature. DENCLUE [78,80] and OPTICS (Ordering Points to Identify the Clustering Structure) [81] are some of the commonly used density-based clustering techniques. An approach called OPTICS is an improvement on DBSCAN. The fundamental benefit of OPTICS is that, unlike conventional DBSCAN, it does not restrict itself to a single holistic parameter setup [82]. Khan et al. [74] explored different enhancement in DBSCAN method such as VDBSCAN [74], FDBSCAN [83], GRIDBSCAN [84], IDBSCAN [79], EDBSCAN [73] etc. to obtain an efficient clustering result. Degirmenci et al. [85] proposed a novel incremental outlier detection technique (iLDCBOF) based on clustering and density-based methods. The iLDCBOF approach gets rid of the three main drawbacks of iLOF such as: 1) doesn't need any user-defined parameters 2) performs brilliantly even on high-dimensional datasets and 3) can identify small outlier clusters. The intrinsic capacity of density-based algorithms to recognize variable shapes and varied densities of clusters in the presence of outliers in the data has led to their increased prominence in recent years. A novel approach based on density-based cluster analysis is called the Density peaks clustering (DPC) algorithm. A DPC may quickly identify the cluster centroids by creating a decision diagram and using the calculation of local density and relative distance [86]. Semi-supervised learning has also benefited from the use of density-based clustering. Additionally, cluster analysis using density-based clustering has been successfully implemented in a variety of contexts [77].

4.2.1.3. ***Model*** *based* ***clustering****.* Model based clustering is a popular technique known for its probabilistic foundations and its flexibility [87]. It assumes that data is generated by a model or an underlying probability distribution [88]. The resultant partition may be understood from a statistical point of view, which is one of the key benefits of this probabilistic technique. Based on standard statistics, model-based approaches estimate the NC [89]. These model-based approaches are divided into statistical and neural network approach methods depending on how

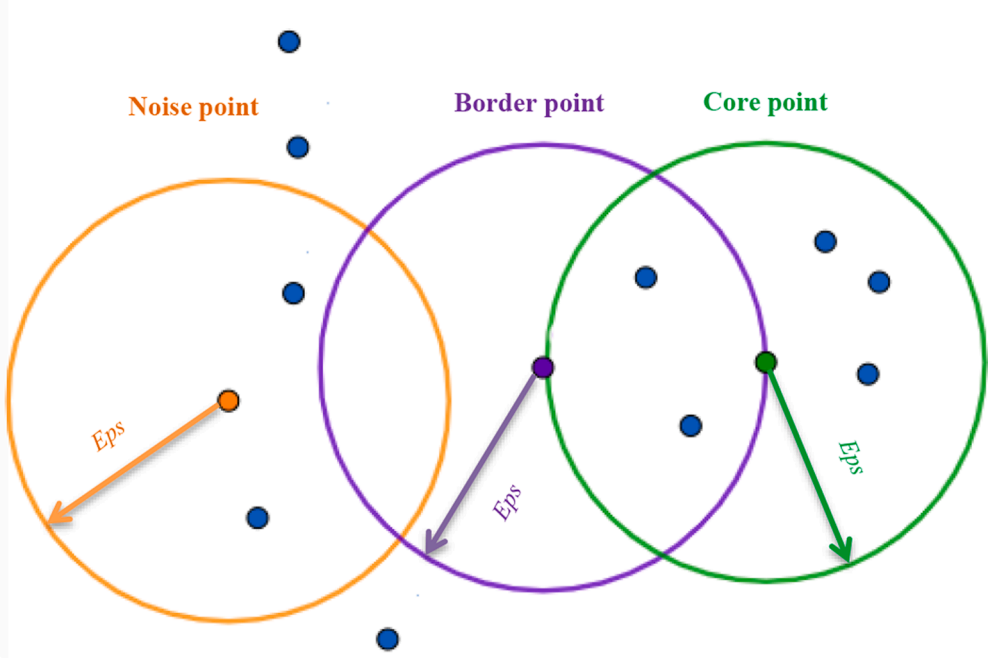


Fig. 5. DBSCAN clustering with *Minpts* = 4.

clusters were generated. “Model-based clustering” is the term used to describe the clustering process that uses (finite) mixture models. A random vector X arises from a parametric finite mixture distribution if, for all $x \subset X$ its density can be written as Eq. (3).

$$f(x|v) = \sum_{g=1}^{G} \pi_g f_g\left(\left(X|\ddot{I}_g\right)\right) \tag{3}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are called mixing proportions, $f_g\left(\left(X|\ddot{I}_g\right)\right)$ is the g^{th} component density, and $v = (\pi, \ddot{I}_1, \cdots \ddot{I}_G)$, with $\pi = (\pi_1, \cdots \pi_G)$, is the vector of parameters [90]. Model-based techniques often employ either statistical methods, such as COBWEB [91], MCLUST [92] and EM [93] or neural network methods, such as ART [94] or Self-organization map [95].

Model-based clustering usually comes with two drawbacks: first, it needs specifying parameters and is reliant on user assumptions that may be inaccurate, resulting in erroneous clusters. Second, dealing with large datasets (especially neural networks) is time-consuming [96].

*4.2.1.4. **Subspace** clustering.* Traditional clustering approaches, such as hierarchical and partitioning perform clustering in entire spaces. Clustering suffers from “the curse of dimensionality” when all features are taken into account. In order to handle complex data subspace clustering (SSC) has been introduced [97]. A technique for finding clusters within various subspaces of high-dimensional datasets is called SSC [97,98]. SSC is an extension of traditional clustering that finds for clusters in various subspaces within a dataset. SSC eliminates irrelevant and redundant features through feature selection, leaving only relevant features for the clustering algorithm to employ when discovering clusters in the dataset. Prior to applying a clustering method in the identified subspace, the clustering process first determines the projections in which clusters may reside [98].

Based on the measure of locality, SSC is classified into two classes: top down and bottom up. Using an APRIORI-style methodology, the bottom-up search strategy leverages the downward closure feature of density to reduce the search space. The different methods of bottom up SSC are Clustering in Quest (CLIQUE) [99], Ensemble clustering (ENCLUS), Merging of Adaptive Intervals Approach to Spatial Data Mining (MAFIA), Cell-based clustering (CBF), Clustering based on decision trees (CLTree) and Density-based optimal projective clustering (DOC). MAFIA and CBF use histogram. CLTree based on decision tree and DOC is based on random search. Examples of top-down SSC techniques are Projected clustering (PROCLUS), Arbitrarily Oriented projected cluster generation (ORCLUS), a fast and intelligent subspace clustering algorithm using dimension voting (FINDIT), Clustering on subsets of attributes (COSA) and δ- Clusters [100]. The grid-based, window-based, and density-based subspace clustering methods are the three main variants of SSC [98].

SSC is applied on different problems of ML and data mining such as classification of diseases, security and privacy in recommender system, computer vision and music analysis [97]. An addition of subspace clustering is known as enhanced subspace clustering algorithms is introduced to overcome the limitations of clustering results. Peng et al. [101] proposed extension of sub space clustering named deep subspace clustering with L1-norm (DSC-L_1) for simultaneous data representation learning and subspace clustering. Agrawal et al. [99] introduced automatic subspace clustering using CLIQUE. CLIQUE automatically identifies subspaces that include high density cluster [102].

4.2.1.5. Squared error clustering. Partitional methods try to divide the datasets into number of separate clusters which are disjoint in nature unlike hierarchical structure. Partition-based clustering methods, such as square-error, aim to minimize square-error, which is the sum of the squared Euclidean distances between every pattern and its cluster center. The prototype-based clustering techniques, which utilize a square-error objective function to measure the distance from each cluster’s center to the pattern, are the most often used partitional clustering algorithms [103]. Early research on minimal sum-of-squared error clustering, or MSSC for short, mostly utilized the well-known KM algorithm [104].

(a) **KM clustering algorithm**

KM clustering algorithm (Hard C-means) [105] is one of the partition based, simplest, and widely used UML algorithms in the literature [1,4,5,106,107]. It is a centroid oriented partition scheme that groups data points into predetermine NC from the given data points [108]. Initially algorithm takes value of K from user and select K number of points as centroids. After locating K centroids, algorithm will calculate the distance between each data point and the centroids and assign every data point to the nearest centroids. Partitional algorithm (KM) usually uses Euclidean distance to calculate the distance between each pair of points and centroids. Again calculate the real centroids from the previously formed clusters and reassign each data point to the nearest centroids. K-mean clustering is an iterative algorithm it will re-compute the cluster till centroids get fixed or terminating condition not met. The standard KM clustering algorithm takes time $O(nkl)$, where n denotes total instances, k represents total NC and l is iteration-number [109]. Let $D = \{X_i\}_{i=1}^{n}$ is a dataset, contains n data points to be partitioned into k clusters. The flow chart illustrating the KM algorithm is shown below in Fig. 6.

KM clustering algorithm usually generates clusters by optimizing an objective function. The objective function (J) [3–5] used in KM algorithm is given in Eq. (4).

$$J = min \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{4}$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a squared distance similarity measure between a data point $x_i^{(j)}$ and the centroid c_j. For each cluster, the squared distance between centroid and its data points should be minimize. The objective function is useful to find the intra cluster similarity between the data points but Aloise et al. [110] proved that the minimization of above objective function is an NP-hard problem and difficult to compute for a smaller value of $k \geq 2$ [3,5,110].

Many researchers, including Steinhaus, Lloyd [111], MacQueen and Jancey from other domains, independently proposed the KM clustering technique between the 1950s and 1960s [25]. Therefore, different variants of this algorithm exists in literature survey [112]. Kanungo et al. [112] implemented simple and practical oriented local search approximation algorithm for clustering known as Lloyd’s algorithm. The steps of Lloyd’s algorithm for KM clustering includes following steps as follows:

KM clustering algorithm consists of following steps:

Input: $D = \{X_i\}_{i=1}^{n}$ // D is a dataset.

Output: Group data points into k–NC.

(i) ***Initialization**: Initialize number of partitions (k), location of centroids and distance metric.*
(ii) *Select K number of centroids randomly.*
(iii) *Calculate the similarity of each data point to the centroids.*
(iv) *Assign each data point to the nearest centroid.*
(v) *Calculate the k centroids iteratively.*
(vi) *Repeat process (iii) and (iv) until centroids get stable.*

The algorithm depends on the three initial setting of parameters: initialization of clusters k, initial seed points and distance metric. The number of parameters in the model, or NC, is rarely known and must be determined before clustering, as the result of clustering varies as the number of cluster parameters changes. The choice of the NC (k) is the most critical task for research community. Either repeatedly running the

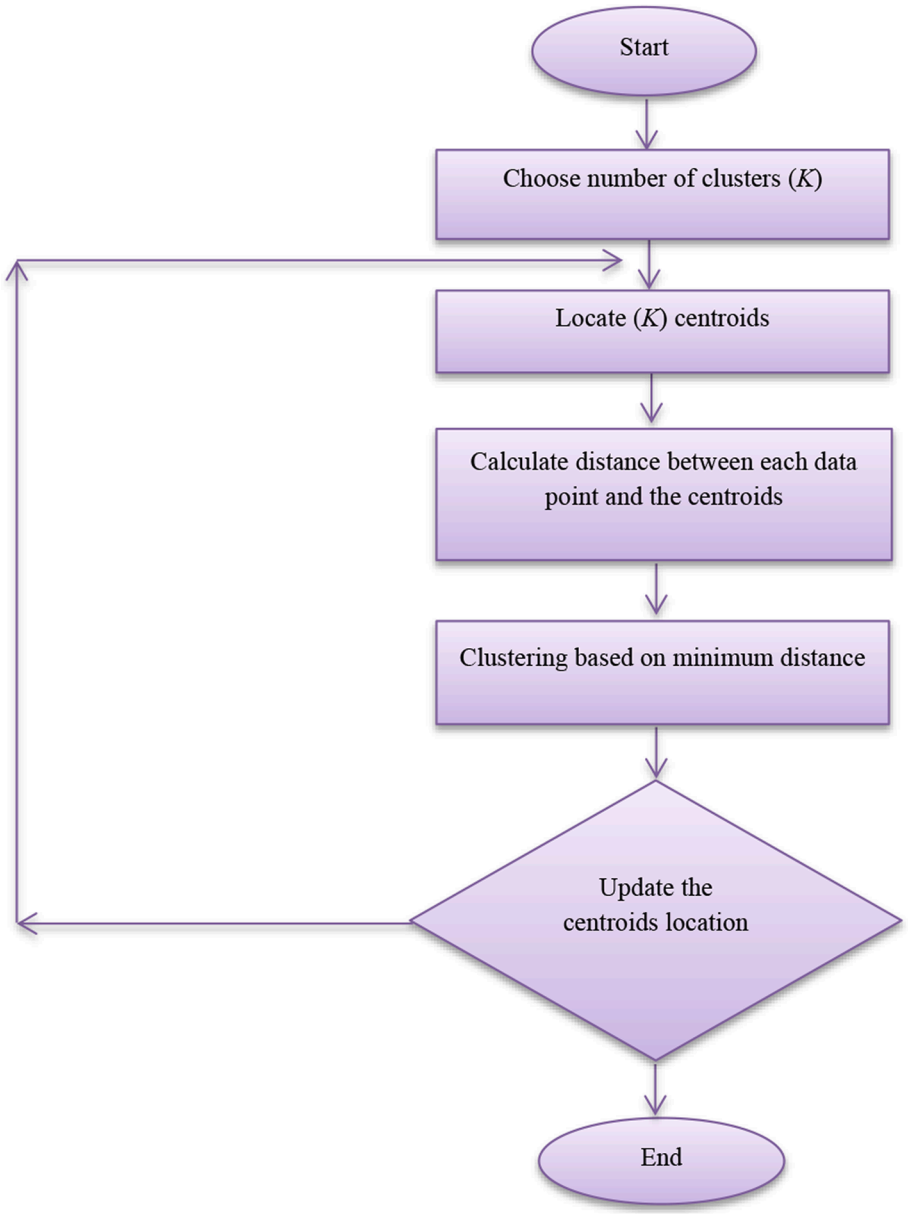


Fig. 6. Flow chart of KM clustering algorithm.

algorithm and selecting the desired NC based on some validity criteria, or automatically determining the NC by some relevant techniques or criteria, is a generic way to find the NC [113]. Kodinariya et al. [107] addressed the problem of determining the NC in KM clustering. The study examined six different methods for choosing the ideal NC for a dataset. Undoubtedly, the most popular partitional clustering algorithm is the KM algorithm [114–116]. There are a few reasons for its popularity. First, it is conceptually straightforward and simple to implement. Second, practically every parameter of the algorithm (initialization, distance function, termination criterion, etc.) may be changed, making it flexible [5]. Third, it has low computational time complexity [25] when distribution of data resembles a Gaussian distribution [117]. Lastly, its convergence is assured.

Although the KM clustering algorithm is extensively used in different fields of studies [3], yet it has some limitations: (a) The algorithm depends on the initial setting of parameters which is not feasible for all types of datasets [112,118]. (b) Sensitive to the initial random selection of k centroids and their locations [116]. (c) Sensitive to noise in the dataset and form cluster with outliers which distort the shape of the cluster and (d) The greedy nature of the KM algorithm may lead it to converge to a local minimum [1,3,4,25]. Thus, numerous researchers have developed heuristic clustering techniques to overcome this limitation [119]. Researchers and practitioners utilize different runs of the KM clustering algorithm to obtain the best NC from the dataset [25]. To address KM's limitations, many variants of the KM clustering algorithm have been developed to improve its performance. To find the NC and the location of initial centroid is the most critical task in KM clustering algorithm. In addressing these challenges, it is important to highlight that the most recent KM expansions and upgrades endeavor to advance the state-of-the-art.

The NC can be determined in one of two ways: either by repeatedly running the algorithm and selecting the desired NC based on some validity criterion, or by determining it automatically using some significant methods or criteria [113]. As the KM algorithm is too sensitive to initial seed points. Therefore, several approaches have been proposed to select the right value of $''k''$ such as the elbow method, an information criterion approach, an information theoretic approach, choosing $''k''$ using the silhouette and cross-validation [106]. Redmond et al. [120]

proposed a technique for initializing the KM clustering algorithm. The proposed approach depends on employing a kd-tree to estimate the data's density across different locations. Finally, select k seeds for the KM algorithm using a variation of Katsavounidis' approach that takes into account this density information. Rahim et al. [121] proposed an efficient method for selecting initial centroids using radial and angular coordinates. The entire dataset is divided into bins, and the distance in Euclidean space is changed to a spherical coordinate system. The initial centroids in the spherical coordinate system are obtained by taking the arithmetic mean of each bin. The improved technique outperforms the conventional algorithm both in terms of time complexity and iterations.

In order to increase the performance and robustness of the KM algorithm, numerous research endeavors have been carried out and reported in the literature. There have been several modified variants based on KM algorithm that have been proposed to address clustering issues for different kinds of datasets [68].

(b). Extension of KM clustering algorithm and its variants

The KM clustering algorithm is extensively acknowledged and applied in the literature. However, the KM clustering algorithm has some drawbacks that need to be addressed in order to solve clustering challenges more effectively. Due of this, various algorithm modifications have been presented in the literature to overcome the acknowledged limitations of the traditional algorithm. Some literature has updated the approach to effectively address the clustering need in a particular domain. Based on this, KM algorithm has different variants that can be classified into four categories:

- **Computational time**

Partitioning a dataset using the KM algorithm is considered an NP-hard problem [1]. Therefore, numerous metaheuristic-based clustering algorithms have been proposed in recent years to address this challenge [122]. Initialization approaches in the KM algorithm have a considerable impact on clustering outcomes. There have been several strategies proposed to initialize the KM algorithm, including linear time complexity, quadratic time complexity, and metaheuristics-based methods [11].

- **Initial centroids selection**

One of the key challenges with the KM algorithm is that clustering results are highly sensitive to the initial centroids. Due to this sensitivity, a variety of initialization techniques for the KM algorithm have been proposed in the literature [11]. Cluster centroids can be chosen at random and optimized by repeatedly executing the process [113].

- **Automatic selection of K**

The cluster number is a user defined parameter needed for the execution of the standard KM clustering algorithm which is a challenging task without having the domain knowledge of datasets. Therefore, several variations have been proposed by researchers to automatically identify the NC (k) for a given dataset. Saha et al. [123] proposed a method called cluster number assisted KM (CNAK). The proposed method can learn the NC during the clustering process. This method is scalable and robust and mainly solves the problem of predetermining the NC and time complexity of Lloyd's KM algorithm in a large-scale dataset. Another approaches of computing NC are U-k-means clustering algorithm [124], MK-means [125], x-means [126] and KM and ANOVA-Based Clustering Approach [127].

- **KM hybridization with metaheuristic algorithms**

The KM algorithm has undergone hybridization with a number of metaheuristic optimization techniques to enhance its overall performance. The hybrid approaches utilized the merits of the respective algorithms [17,128]. Several nature-inspired optimized KM variants exist in the literature, such as KM with the Bat algorithm [129], KM with the Cuckoo algorithm [130], KM with the Wolf Search algorithm [131] and KM with PSO [132]. For better clustering performance, hybrid algorithms combine the metaheuristics methods with the KM algorithm. The performance of the resultant metaheuristics algorithms is impacted by some of the issues with the metaheuristic algorithm, including high computational complexity and time as well as the requirement for parameter tuning for the best outcomes [25].

4.2.1.6. Search based clustering. The KM method has grown in popularity as a center-based clustering technique due to its simplicity, efficacy, and linear time complexity. Nevertheless, KM's drawbacks include its reliance on the initial state and tendency to converge to local minima [133]. Several automatic clustering techniques have been developed to address these issues. Over the last two decades or more, AI based clustering techniques which includes evolutionary algorithm, SI techniques and artificial neural networks (ANNs) are successfully applied to clustering problems [119,134]. Evolutionary algorithms (GAs, Differential evolution (DE)) and SI algorithms (Ant colony optimization (ACO) algorithm, artificial bee colony (ABC) optimization algorithm, particle swarm optimization (PSO), symbiotic organism search, tabu search, firefly algorithm [135], whale optimization algorithm, emperor penguin optimizer, bacterial evolutionary algorithm, invasive weed optimization, etc.) are also known as metaheuristics algorithms. Metaheuristic algorithms have recently been used to address challenging real-world problems in an array of areas, including administration, management, economics, and engineering [136]. The main components of a metaheuristic algorithm are intensification and diversity [137].

Many of metaheuristic algorithms are motivated by biological evolution, swarm behavior, and physics' law. These algorithms may be divided into two types: population-based metaheuristic algorithms and single-solution algorithms. In general, search based clustering techniques are used to determine the optimum value (minimum or maximum value) of the criteria function, also known as the objective function [1]. There are two sub categories of search based techniques: stochastic and deterministic search technique. The most of stochastic techniques are based on evolutionary methods. Avoiding local optima traps is the primary objective of evolutionary algorithms when solving real optimization problems. The remaining search strategies fall under the category of deterministic research techniques. The next sections provide studies on using AI-based techniques to address clustering problems.

(a) Evolutionary algorithm

(i) Genetic algorithm-based clustering

A clustering method based on GA is known as GA-clustering [138]. The GA, which transfers the ideas of evolution in nature to computers and mimics natural evolution, is the most well-known evolutionary algorithm (EA). In general, they solve issues by maintaining a population of potential answers in accordance with the "survival of the fittest" idea. Several non-GA-based clustering methods, including KM [139], FCM, EM, etc., have found wide applications. Yet, in the majority of real-life scenarios, the NC in a dataset is unknown. GAs is used to provide appropriate clustering and to evolve the correct number of cluster. The GA belongs to a class of search algorithms that is particularly suitable to handling difficult optimization problems [140]. As other optimization algorithms, the GA starts by defining the optimization variables, the cost function, and the cost. The search space's parameters are represented as strings (chromosomes) encoded with a combination of cluster centroids. Search direction is controlled by the selection operator, and new search regions are created by the recombination operator [141]. Therefore, selection and recombination operators in GAs play a significant role.

Several research initiatives to create GA-based clustering algorithms have been described in the literature. Sarkar et al. [142] developed a clustering technique inspired by evolutionary programming that splits a

given dataset into the optimal number of groups. This approach avoids locally optimum solutions by choosing the NC and their locations carefully. The selection of the initial cluster centers has little effect on the algorithm's outcome. Ding et al. [143] proposed a GA based algorithm known as GAKFCM clustering technique to enhance the FCM algorithm's clustering performance. The proposed methodology is a combination of GA and kernel technique. First, the initial clustering center is optimized using the improved adaptive GA, and then the KFCM method is utilized to direct categorization in order to enhance the FCM algorithm's clustering performance. Wang et al. [140] presented novel clustering method, Extension genetic algorithm (EGA), is a combination of extension theory and GA. The extension theory uses extension set that extends the fuzzy set range from [0, 1] to [-∞, ∞] which means an element belong to any extension set to a different degree.

(b) Swarm Intelligence Algorithm

SI, a subfield of evolutionary computing (EC), draws inspiration from the natural behavior of social animals and insects. SI is an emerging area that studies the behavior of social insects such as ants, birds, and bees in order to find optimized solutions to search-based problems. The growing field of SI has been modelling the behaviour of social insects, such as birds, ants, or bees, for the purposes of search and problem solving. A paradigm for AI is SI which aims at solving complex computational problems (NP-hard). SI is a growing topic of research that offers a more optimal approach for addressing problems than the conventional approach of problem solving in diverse fields of study [144]. In the recent past, there has been a significant increase in studies on the SI paradigm, particularly with regards to ACO, ABC, and PSO [145,146].

(i) ACO clustering algorithm

The most well-known and popular SI algorithm, ACO, was developed by Dorigo and Di Cario in 1999 [147]. ACO is a modern metaheuristic technique that draws inspiration from the foraging behaviour of actual ant colonies. The ants leave pheromone on surfaces to indicate the route from the nest to the food source, which other ants in the colony should follow. Foraging behaviour, labor division, brood sorting, and cooperative transport are just a few of the several features of real ant colonies' behaviour that have inspired the development of various types of ant algorithms in recent years. An novel heuristic strategy for solving a combinatorial optimization problem was introduced in the early 1990s and is known as the "ant system" algorithm [147].

There are two types of ant-based clustering: the first closely matches the clustering behavior observed in natural ant colonies. The second is less directly influenced by nature [144]. A basic ant-based data clustering method was introduced by Lumer and Faieta, which closely resembles the ant behaviour described in [144]. Later, this approach was extended and modified by a number of researchers in order to enhance performance. This include adaptive setting of the algorithmic parameters [148] and hybridization with KM and FCM [149,150].

Due to the self-organization behaviour of real ant, ACO used in various engineering applications such as digital image processing [151,152], scheduling [153], clustering [154,155], routing algorithms [147,156], text mining [157] and intrusion detection [158]. Azzag et al. [159] presented real ants' self-organization behaviour may be applied to hierarchical tree-structured data partitioning using similarity metrics. The learning process is significantly accelerated by data clustering using tree methods [57]. This allows the use of such algorithms on bigger datasets. Inkaya et al. [160] proposed a novel clustering methodology based on ACO called (ACO-C) for spatial clustering problem. The Proposed methodology can handle variation within cluster size, shape and density without having a prior knowledge of cluster count. This method is useful in image segmentation, computer graphics and geographical information system. Nowadays, ACO is a well-defined and effective metaheuristic that is used more and more frequently to address a wide range of challenging combinatorial optimization issues. Also, there have been several studies in ACO that focus on parallel processing as a means of accelerating the algorithm [148].

(ii) PSO

PSO is an another general purpose metaheuristic-based computational technique proposed by Kennedy and Eberhart [161] in 1995 and inspired by swarm behaviour (cognitive and social behaviour) observed in nature. The study demonstrates that natural behaviour of living things, such as termites, bees, ants, and insects are successfully used to address clustering problems in natural systems. Data clustering has become a significant area of interest for the use of optimization-based approaches due to the growing complexity and size of the datasets. PSO has been implemented as an optimization technique in a variety of fields [162] such as computer graphics, music composition, biological and medical applications [163], PSO based text clustering [164] and sensor networks [165]. A few of the additional PSO-based image processing fields includes image segmentation [166], noise removal, edge detection, image classification [167] and feature selection in medical [167] and image analysis. When applied to an optimization problem, a normal PSO method begins with an initial guess of parameters. Selecting the starting swarm is one of the important initializations. The complexity of the task determines how many particles there are in the swarm. Each individual of the swarm is referred to as a particle in PSO. A potential solution is given in this algorithm as a particle. To reach a global optimum, it moves towards a promising location while using a collection of flying particles (variable solutions) in a search area. There are three types of natural computing paradigms: Epigenesis, phylogeny and ontogeny. The Ontogeny group is connected to the mechanisms through which a specific creature may adapt to its surroundings. In contrast to other categories, this group contains cooperative algorithms such as PSO and GA [163].

Various PSO-based clustering methods have been examined in order to improve their effectiveness and yield better clustering results. The majority of PSO clustering implementations focus on the clustering of text and numeric data. Despite the fact that PSO has been used to address clustering issues, such as complex datasets, overlapping cluster prototypes and high-dimensional feature space, the efficiency is typically not adequate when working with large or complex datasets [168]. The cluster analysis of data that contain a few dozen to thousands of dimensions is known as clustering high-dimensional data. High-dimensional data spaces are common in domains such as health, bioinformatics, biology, recommendation systems, and text document clustering. The issue of clustering high-dimensional data may be successfully solved by using the PSO in combination with other methods, including dimensionality reduction (DR) and subspace clustering [169].

Chen et al. [170] proposed novel clustering technique by combining KM and PSO called RVPSO-K. The proposed technique is used to extract pattern from high-dimension data such as web usage pattern. Chuang et al. [171] proposed an algorithm known as accelerated chaotic particle swarm optimization (ACPSO) which locates effectively and efficiently more suitable cluster centers. In an N-dimensional Euclidean space, ACPSO searches stable data cluster centers using minimal intra-cluster distances as a metric. Kuo et al. [172] proposed a dynamic clustering approach by integrating PSO and GA known as DCPG. The proposed algorithm (DCPG) is used to solve the problem of pre-setting the NC and determining the right NC based on the characteristics of the data. The DCPG algorithm produced improved clustering outcomes with fast convergence. Alswaitti et al. [173] proposed a data clustering strategy based on the PSO clustering algorithm and the kernel density estimation. Significant issues with PSO-based clustering techniques are addressed by the proposed algorithm, such as the challenge of tuning the learning parameters.

The existing PSO-based algorithms must tune some parameters before discovering a better solution, which raises the issue of generalization. As a result, researchers are focused on two major areas: automation of techniques and generalization of PSO-based algorithms [174].

(iii) ABC

ABC optimization methods are one of the extensively used heuristic or metaheuristic optimization techniques that are influenced by natural phenomena of real honey bees to solve clustering-related numerical

optimization problems. It is population based metaheuristics algorithm. One of the most extensively studied social insets is honeybees [175]. Recent research on SI has focused on their foraging behaviour, learning, memory, and information exchange features [176]. The foraging behaviour of bees is used as inspiration for the development of an ABC algorithm to address clustering issues. When the number of partitions is known a priori and is crisp in nature, the ABC method for data clustering can be used [119]. ABC blends local search methods with global search methods for finding optimized clusters [177]. Processing high-dimensional data is not suitable for the conventional optimization techniques. The major objective of the ABC technique is to minimize the execution time and to obtain optimum clusters for the different dataset sizes [177]. Numerous applications, including neural networks, protein structure, sensor networks image processing [178], data mining, mechanical engineering, civil engineering, and electrical engineering, have effectively exploited ABC. Intelligent characteristics of bee swarms, such as autonomy, self-organization, distributed functioning, labor division, etc., enable solutions to difficult transportation issues as well as deterministic combinatorial problems in dynamic environments [179]. Karaboga et al. [177] have applied ABC algorithm for grouping homogeneous objects on different datasets available in UCI ML Repository. Experimental results demonstrated that ABC algorithm performs better in problems involving function optimization as compared to GA, DE and PSO. It can also be extended in multivariate data clustering [180]. Zhang et al. [119] implemented ABC clustering algorithm for optimum partition of N objects into k clusters. Deb's rules guided the search direction for each candidate. The results of the computer simulations are quite positive in terms of the quality of the solution and the required processing time. However, as the majority of these heuristic methods are created for numeric data, they cannot effectively handle categorical data. It is essential to create an ABC-based clustering method for categorical data due to the availability of categorical data in real-world applications. Some researchers merged an ABC-based algorithm with traditional partitional algorithms such as K-modes to address the limitation of falling into local optima. Jinchao ji et al. [181] proposed a novel clustering algorithm by combining traditional k-modes with ABC called ABC-k-Modes for handling categorical data. Authors first present a one-step k-modes process and then combine it with the ABC method.

Another two-step ABC algorithm was proposed by Kumar et al. [182] for efficient data clustering. The three critical challenges of the ABC algorithm—the starting positions of food sources, the solution search equation, and the location of abandoned food—are the focus of this research work. In compared to other metaheuristic algorithms, ABC is a simple and versatile approach that requires less parameter tuning. The ability of the real ABC, its modifications, and hybrid algorithms to address a variety of optimization issues, including continuous, combinatorial, constrained, binary, multi-objective, chaotic, etc. Many studies conducted in the literature indicate ABC's applicability, accuracy, and efficiency in handling a range of optimization problems.

(C) Artificial neural networks

The incredible processing potential of ANNs has been released by recent developments in ML. The neural network technique connects input and output using weighted units. Neural network properties such as in-built parallel computing architecture, quantitative characteristics are created by transforming or processing numerical data, and recursively updating the weights to obtain the best fit of the data, making them useful for clustering datasets [89].

(i) Deep learning-based clustering

The ML area, comprising deep learning, has recently gained popularity and has been shown to be a useful model for classification, data clustering, pattern recognition, and forecasting in a wide range of fields [183]. Deep learning-based clustering methods rely on deep neural networks (DNNs) for discovering homogeneous partitions. Due to their inherent ability to alter data in a non-linear way, DNNs can be used to convert the data into more clustering-friendly representations. In deep learning, input data and clusters are represented by neurons. Each link is assigned a weight, which is first chosen at random before being learned adaptively. The self-organizing map (SOM) is a highly common neural technique for clustering. Together with clustering analysis, SOM is frequently applied for data visualization, feature extraction, and vector quantization [1].

Deep learning-based clustering is a collection of methods that use deep DNNs to develop cluster-friendly representations. Unsupervised learning methods can be used to extract important information about the hierarchy and structure of the data. The extracted knowledge representation, which already has a strong understanding of the underlying nature of the data and should only be fine-tuned for the specific task, can be utilized as the foundation for a deep model that needs fewer labeled instances. Deep clustering algorithms generally have two components that make up the loss function (optimizing objective): network loss L_n and clustering loss L_c. Thus, the loss function may be written as follows shown in Eq. (5):

$$L = \lambda L_n + (1-\lambda)L_c \tag{5}$$

where $\lambda \varepsilon [0, 1]$ is a hyper-parameter that tune L_n and L_c. DNNs often need a lot of training data and powerful computational resources, and adjusting their parameters takes a lot of time [184]. However some studies have been proposed to improve the performance of the neural networks [184]. It makes sense to combine clustering approaches with deep learning to improve clustering results, as deep clustering is widely used in a variety of applications due to its powerful feature extraction capabilities.

4.2.1.7. Miscellaneous clustering techniques.

(i) **Time series clustering**

Real-world applications may now store and preserve data for a longer period of time because of advances in data storage and processing technology. Therefore, data is frequently stored in the form of time series data in many different applications, such as sales data, stock prices, exchange rates in finance, weather data, biomedical measurements (such as blood pressure and electrocardiogram measurements), biometrics data (image data for facial recognition), and particle tracking in physics, etc. Due to the increasing use of time series data, there have been several attempts at initiating research and development in the field of time series clustering [185]. Time series clustering, similar to static data clustering, requires a clustering method or process to create clusters from a set of unlabeled data objects [34]. There are differences between time series data that are discrete-valued or real-valued, uniformly or non-uniformly sampled, univariate or multivariate, and equal or unequal length data series. To cluster various types of time series data, a number of techniques have been developed [34]. Time series data is classified as dynamic data [186]. Some research attempts to adapt existing clustering algorithms to handle time-series data or try to transform time-series data to static data format so that the existing algorithms may be utilized directly [34]. There are many different types of time series data research exist, such as finding similar time series [187], subsequence searching in time series [188], DR and segmentation [189].

As per the literature review, time series clustering has primarily three categories: whole time-series clustering, subsequence clustering and time point clustering as shown in Fig. 7. An interesting research problem in time-series clustering has been identified as the high dimensionality, extremely high features correlation, and generally significant level of noise in time-series data. Many studies have focused on the high dimensionality of time-series data and have attempted to propose a method of representing time-series in a reduced dimension that is compatible with traditional clustering techniques. The following areas can be the subject of further research [190]:

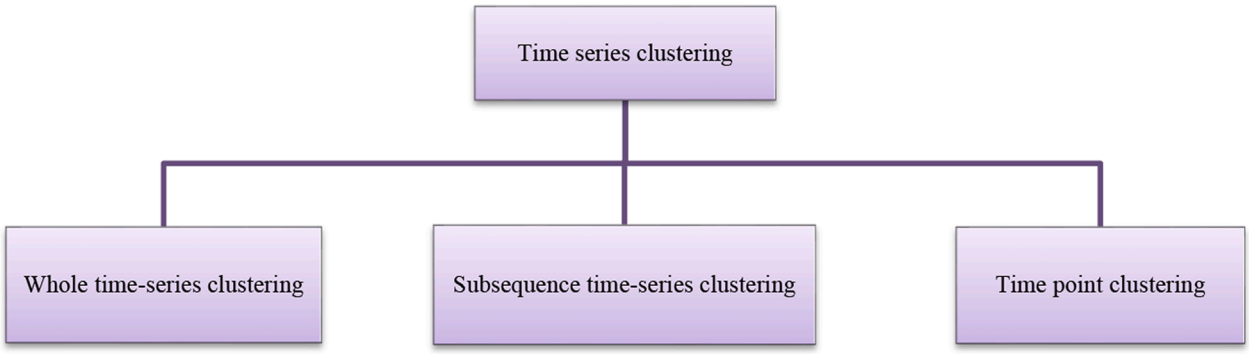


Fig. 7. Time series clustering taxonomy [96].

- Increasing the speed of time series data clustering in high-dimensional data.
- The clipping approach can be used to enhance computation effort in high-dimensional data.
- To forecast future value in time series data, an efficient method may be developed.

(ii) **Streaming**

The rate of data expansion in the modern world is exponential. Several applications, such as sensor networks, smart grids, video surveillance, medical research data, financial systems, network data, web click streams, etc., generate vast volumes of data streams at extremely high rates. [89]. Data stream mining is an ongoing research topic aimed at uncovering knowledge from unbounded large sequences of continuously generated data at very high speeds [191–193]. The massive sequence of continuous data imposes multiple challenges in clustering streaming data such time and memory constraints. Applications for data streams include data mining from sensor networks, weather analysis, and meteorological analysis [192]. Object-based clustering or attribute-based clustering have been the focus of applications of clustering techniques to data streams, former one is more common.

Object-based data stream clustering algorithms consist of two primary phases: the data abstraction phase and the clustering phase. Attribute clustering is also known as variable clustering. This is often thought of as a batch offline method, and the typical approach is to use a conventional clustering algorithm over the transposed data matrix. Data stream mining is a real-time method used to extract intriguing patterns from dynamic and quick-changing data. Finding important information in enormous data streams during a single scan is the challenging problem. Various challenges are involved in real-time data stream clustering such as mixed data types, outliers, noise and dynamic distribution of data stream etc. [89]. Numerous effective data stream clustering methods have been brought forth, and the experimental outcomes are quite good [194].

(iii) **Multiview clustering**

Sensors and the Internet of Things have led to a rise in the use of multiview data in real-world scenarios. This data is semantically richer, more relevant, and more complicated than traditional single view data that represents objects from a single perspective [195]. After years of advancement, the research on traditional single view clustering is almost reached at bottleneck. The primary cause of this scenario is because the datasets only describe things from a single viewpoint, making it difficult for them to fully extract the objects' comprehensive information. Multi-view data, in which the same items are depicted from many angles, has started to become more common. Due to the rapid development in multimedia and big data, multiview data are widely used in everyday applications. Several researchers have attempted to build models for multi-view data based on the idea of KM because of its simplicity [195]. The taxonomy of multiview clustering methods is shown in Fig. 8.

4.2.2. Fuzzy (Soft) clustering

The hard (crisp) clustering algorithms have the restriction that every data point in the dataset may only belong to one specific cluster [197]. The truth value of the hard/crisp set is either 0 or 1, which may be utilized for solving a two-valued problem since each member of the crisp set is part of or is not part of a set [140]. When using hard clustering, this strategy frequently ignores the possibility of overlapping data samples between classes [198] whereas fuzzy clustering, which has its roots in fuzzy logic, illustrates the likelihood or degrees of one data point belonging to many groups or clusters simultaneously [199]. The fuzzy set permits the depiction of notions whose boundaries are not explicit, in contrast to the crisp set. In addition to whether a member belongs to the set, it also matters to what extent it does. The fuzzy set's membership function has a range between 0 and 1 [140]. Fuzzy clustering is a sophisticated approach for processing unlabeled data with outliers and unusual patterns [199]. In 1965, Zadeh [200] proposed fuzzy set theory, which provided an idea of the uncertainty of belonging and was expressed by a membership function. Data points can be a member of several clusters, and each data has a set of membership degrees associated to it. A fuzzy-based cluster analysis model is useful when the dataset and partition may contain ambiguity [201]. With the advent of the fuzzy set concept, a deeper framework for specifying the relationships between points and clusters was made possible [202,203].

As stated before, fuzzy-based cluster analysis methods provide a sense of ambiguity about belonging, as specified by a membership function that is suitable for real-world datasets [204]. Therefore, significant works on fuzzy clustering methods exist in the literature. There are various fuzzy clustering techniques, including the objective function, the equivalence relation, the hierarchical approach, and the graph theory. The approaches based on objective function are the ones that have been investigated the most extensively among these fuzzy clustering techniques. Ruspini introduces the fuzzy concept to hard clustering and suggests fuzzy clustering, such as the FCM clustering method published by Dunn and modified by Bezdek [37,205]. FCM is an implementation of a fuzzy clustering method which uses KM principles to divide a dataset into clusters [12]. The FCM clustering is produced by minimizing the objective function as given in Eq. (6):

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{m} |p_i - v_k| \tag{6}$$

where J is the objective function, n represents number of instances, c is the number of groups, μ_{ik}^{m} is the likelihood value assigning the data point i to the cluster k, m is the fuzziness factor and $|p_i - v_k|$ is the distance (Euclidean distance) between the point p_i and k^{th} cluster center v_k [12].

The FCM algorithm is a distinct and extensively utilized clustering algorithm that has been employed in a variety of commercial and

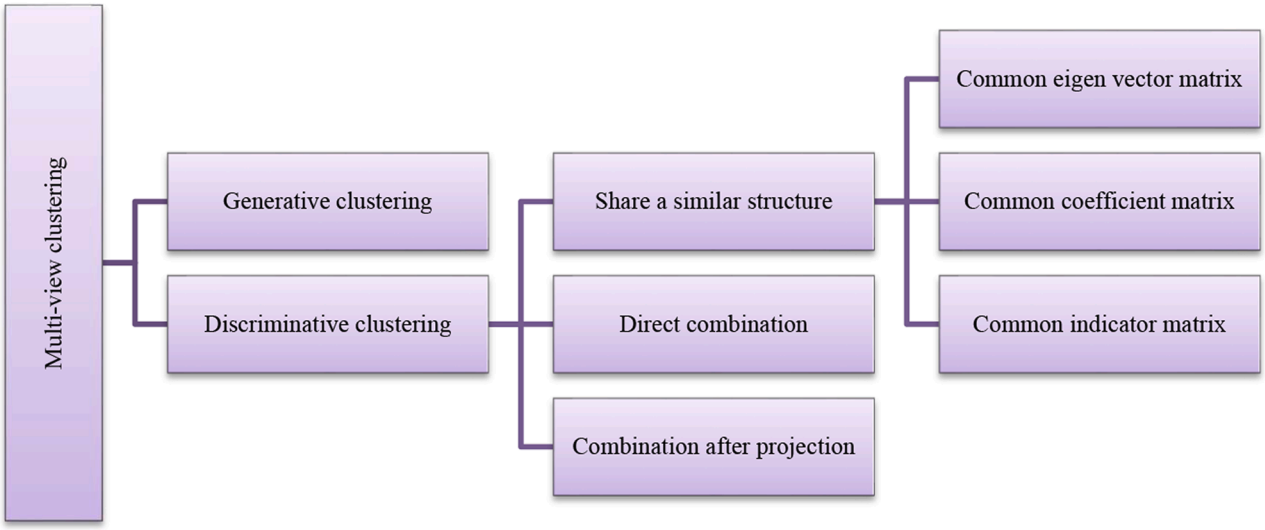


Fig. 8. Taxonomy of multi-view clustering methods [196].

scientific domains, including pattern recognition, data mining, and bioinformatics [204]. Fuzzy clustering can effectively address problems including the ability to recognize patterns, mixed-media information, missing/noisy data, and the provision of faster approximations [1]. However, it has many limitations such as complex calculations, unable to handle high-dimensional datasets [37], cluster validity, a priori knowledge of cluster prototypes [143], shape and density and scalability. Therefore, scholars improved the FCM clustering technique from many perspectives. Gath et al. [205] proposed a novel method by combining the fuzzy KM algorithm and fuzzy maximum likelihood estimation for prototype classification without apriori assumptions. Park et al. [206] presented a novel algorithm called gradient based FCM (GBFCM) aiming at minimize the objective function. The proposed algorithm is a combination of gradient descent with FCM which converges faster than traditional FCM. Wu et al. [207] proposed a Fuzzy Kernel c-means clustering algorithm (FKCM). The novel FKCM technique solves cluster shape problems with fuzzy clustering by integrating FCM with the mercer kernel function. Kuo et al. [208] proposed a kernel intuitionistic FCM (KIFCM) algorithm. The proposed algorithm combines with metaheuristic algorithm to make it more robust and efficient. Zhang et al. [209] proposed a novel cluster validity index weighted global–local based validity index (WGLI) to identify the optimal number of groups. Winkler et al. [210] investigated the challenges in FCM in high- dimensional spaces. Stetco et al. [211] investigated the study of FCM++ method, which improves the efficiency and speed of FCM by employing the seeding process of the KM++ algorithm. The proposed method demonstrated significant improvement in convergence time. A common data clustering technique is the FCM algorithm. However, there is a chance that it will converge to local minima. To escape from the local minima, Kumar et al. [212] proposed a hybrid method based on improved ABC and FCM algorithms labeled as (IABCFCM). The proposed approach exploited the strengths of both algorithms to improve clustering results. The result of the FCM algorithm is employed as a food source, with the rest of the food sources initiated at random within the dataset. The algorithm's performance is tested on six datasets obtained from the UCI Repository, which includes glass, iris, soybean (small), lung cancer, wine, and vowel datasets.

In more recent work, Jayalakshmi et al. [213] proposed a taylor horse herd optimization (THHO)-based deep fuzzy clustering (DFC) for recommendation of web page. The proposed method is useful in intelligent web frameworks for recommending appropriate web pages. Senthilkumar et al. [214] proposed an intelligent segmentation and hybrid ML based classifier for effective rice disease prediction. The proposed method is known as modified feature weighted fuzzy clustering (MFWFC). The SMOTE-based preprocessing is first implemented in order to normalize the data. The MFWFC based segmentation is then presented for effective segmentation. The Principal component analysis (PCA) then goes through the feature extraction procedure to improve the classifier's performance. The selection of features is then carried out using LDA, and HCM is created to increase prediction performance. In this, the SVM and Improved Recurrent Neural Network (RNN) are integrated. Raja et al. [215] proposed a novel method based PSO and FCM for predicting type 2 diabetic mellitus (T2DM) diseases. The method was tested on the Pima Indians Diabetes Database and outperformed other methods by 8.26 %.

4.3. *Hybrid clustering*

Hierarchical and partitioned clustering methodologies have various benefits and drawbacks, giving researchers the insight to integrate the diverse approaches to overcome the limits of clustering methods [216]. A hybrid clustering approach is a novel method that is obtained by cascading the results of different clustering methods. Different learning methods can be incorporated in clustering. Several techniques have embraced the benefits of hierarchical and partitional clustering for hybridization [217].

Traditional supervised classification algorithms frequently focus solely on determining the relationship between input data and class labels; they rarely take into account the potential structural characteristics of the sample space, which frequently yields undesirable classification results. Several researchers have worked to create hybrid models that combine supervised and unsupervised learning [218], unsupervised and unsupervised learning [219], hierarchical and partitioning algorithms and unsupervised learning with metaheuristics optimization techniques [220] in order to enhance the performance of the existing algorithms [221]. The creation of hybrid approaches by integrating concepts of exact algorithms with metaheuristics is a popular current trend [148].

To enhance the classification performance of models, researchers have successively introduced a variety of hybrid models during the past decade years that integrate supervised and unsupervised learning. The standard approach for building a hybrid model is to combine clustering with a decision tree algorithm [184]. Gaddam et al. [222] proposed a novel hybrid method by cascading unsupervised and supervised data partitioning techniques known as "KM+ID3 decision trees" for anomaly detection. The first stage involves performing KM clustering on training

instances to generate k disjoint groups. The next step is to train an ID3 decision tree on each cluster. The results of the KM and ID3 algorithms are merged using the Nearest-neighbor rule and the nearest consensus rule in order to reach a final classification result. Experiments using three datasets for network anomaly detection revealed that the hybrid model performs better than a single ID3 model. In predicting customer churn, Bose et al. [223] suggested a two-stage hybrid model (KM-Boosted C5.0) that combines an unsupervised clustering approach with a boosted C5.0 decision tree. Kaur et al. [224] proposed a hybrid algorithm for intrusion detection system (IDS). The proposed algorithm is a hybridization of simple KM+Firefly algorithm. The training model is constructed using the clustering algorithm KM and classification is carried out on the test set. The limitations of KM algorithm such as random initialization and convergence speed are resolved by firefly algorithm. The authors used NSL-KDD dataset to measure the efficiency of the proposed algorithm. A similar type of study is proposed by Yaseen et al. [225] in IDS in multiagent systems. The proposed hybrid algorithm is a combination of modified KM with C4.5. KDD cup 1999 dataset was used to evaluate the algorithm's performance. An another similar type of study proposed by Horng et al. [58] for IDS by using BIRCH hierarchical clustering and SVM. Instead of using the original large dataset for the SVM training, the BIRCH hierarchical clustering could produce highly qualified, abstracted and reduced datasets. As a result, the training time was significantly reduced, and the resulting SVM classifiers performed better than those that were trained on the redundant dataset from the outset. The suggested method was evaluated using the famous KDD cup 1999 dataset and obtained accuracy of 95.72 %. Subudhi et al. [6] proposed a novel hybrid approach by combining optimized FCM and supervised learning algorithms. The proposed hybrid method used for fraud detection in automobile insurance that involves two phases: training and fraud detection. Based on the test data points' distance from the optimized cluster centers, the GA based FCM (GAFCM) classifies them as belonging to the genuine, malicious or suspicious classifications during the initial phase of fraud detection. Genuine and fraudulent samples are not further processed whereas the suspicious samples are additionally validated in a second phase by four separate methods which include supervised learners-decision tree (DT), SVM, group method of data handling (GMDH) and multi-layer perceptron (MLP) using 10-fold cross validation. Rahman et al. [20] proposed a novel hybrid clustering technique by combining GA with KM for better quality clusters without seeking any user inputs such as NC a priori. Nguyen et al. [40] proposed a hybrid clustering approach by combining GA, KM Logarithmic Regression and Expectation Maximization (EM) labeled as GAKREM that combines the best qualities of the KM and EM algorithms while avoiding their drawbacks, such as the requirement to specify the NC in advance, termination in local optima and time-consuming computations. The proposed study can be used in any type of mixture model that based on parameter estimation. Maksoud et al. [219] proposed a hybrid clustering techniques with the integration of two UML algorithms: KM and FCM. In terms of low calculation time, the suggested method can benefit from KM clustering for image segmentation. Furthermore, it gets benefit from the FCM in terms of accuracy. Ravindra Jain [221] described a hybrid clustering model based on K-mean and K-harmonic mean (KHM) focuses on fast and accurate clustering results. The new algorithm's performance is independent of the dataset's size, scale, and values. Sonawane et al. [220] proposed hybrid heuristic-based automated approach for predicting heart disease. Initially the optimal feature extraction procedure is used to the medical data to extract the important information and remove the redundant data to increase the prediction rate. The hybrid clustering method, which is carried out by optimized KM clustering (KMC) and optimized DBSCAN, provided significant and optimal characteristics for the prediction of heart diseases.

In conclusion, the strengths and weaknesses of hierarchical and partitioned clustering procedures allow researchers the insight to combine the various ways to get around the drawbacks of clustering methods. Clustering algorithms and metaheuristics have recently been hybridized, which has helped in tackling the majority of issues relating to real-life datasets such as ability to do a global search for the ideal number of homogeneous groups and cluster centroids. Furthermore, some metaheuristic- based algorithms have been embedded with the KM algorithm and its variants to enhance their ability to solve automatic clustering issues. Table 6 summarizes clustering techniques, including their advantages and limitations.

5. Recent work on clustering methods

In clustering, every dataset instance will receive a label indicating its participation in a cluster. The base of clustering algorithms is often the same, although the methods used to calculate distance and similarity as well as how labels are chosen frequently vary. Every clustering process is based on an objective function. Therefore, there is no absolute way of choosing the best clustering algorithm [226,227]. In fact, a single clustering approach might not yield the best results across all datasets. Due to the subjective nature of the clustering algorithms, validating a partition obtained using clustering algorithms is a significant issue [42,228–230]. Along with cluster validation, automatic clustering, DR, mixed data type clustering and hybrid clustering are the recent areas of study that have drawn attention to the scientific community [7,37,117] as shown in Fig. 9. The subsequent sub section will discuss recent developments in data clustering.

5.1. Cluster validity

A framework for cluster validity solves the problem of correctly anticipating the NC [231]. Bolshakova et al. [231] studied three validation indices (Silhouette, Dunn and Davies Bouldin indices) were applied to two genome expression datasets (leukaemia and B-cell lymphomadata), using different intracluster and intercluster distances. Li et al. [229] proposed a method for dynamically determining the nearly-optimal NC using the ratio of deviation of sum-of-squares and Euclidean distance (RDSED). Chowdhury et al. [227] introduced maximization of Shannon's entropy for initialization of KM. This method is used for discovering the optimum NC from a dataset. The proposed algorithm's results were compared to those of existing initialization techniques using the parameters initialization time, computational time, and number of iterations. Wang et al. [232] proposed a method to determine optimal cluster centroid based on concurrent multi-objective optimization (FMOEA-K). In contrast to traditional approaches, this approach combines the fuzzy clustering effectiveness index with the multi-objective optimization algorithm (MOEA) which is used to simultaneously search for the right cluster center. Patil et al. [233] proposed an another method known as depth difference (DeD) to find the optimal NC prior to actual clustering is constructed. In order to get the ideal value of k, which serves as an input value for the clustering algorithm, they defined the depth inside clusters, depth between clusters, and depth difference. In a multivariate dataset, data depth measures the median. They calculated the centrality of various data points within a dataset using the Mahalanobis depth function. Each data point in the study is given a number between 0 and 1, and each value represents the centrality or depth of that particular data point. The cluster centroid is corresponding to the maximum depth value. The ideal clusters are found by estimating the average distance between each depth value and the cluster centroid. By increasing the estimated value of the depth difference, the optimal NC is discovered.

5.2. Automatic clustering

There have been several clustering methods created, which may be divided into hierarchical and partitional techniques. Despite the fact that the methods in both of these categories have shown to be quite effective and successful, they usually involve anticipating of the NC [234,235]. In addition, predicting the NC while dealing with real-world

Table 6
Summary of clustering techniques.

Category		Technique	Advantage	Limitation
Hierarchical		• Agglomerative	• Uses bottom-up approach.Independent from NC. • Less sensitive to noise and outliers.	• Sensitive to cut-off selection. • Dependency of output in linkage technique. • Challenging for high dimensional datasets.
		• Divisive	• Uses top-down approach.	• Approaches need more computational resources.
Partitional	**Hard/crisp clustering**	• Graph based	• Useful when dealing with data that violates a Gaussian or spherical distribution.	• Sensitive to outliers.
		• Density based	• Uses the idea of density to discover clusters of various forms, sizes and densities. • DBSCAN successfully handles outliers.	• Sensitive to initial parameters.
		• Model based	• Model-based approaches can estimate the NC.	• Needs specifying parameters and is reliant on user assumptions that may be inaccurate.Dealing with large datasets (especially neural networks) is time-consuming.
		• Subspace	• Finding clusters within various subspaces of high-dimensional datasets.Free from “the curse of dimensionality”.	• Input the NC and subspace sizes as parameters.
		• Squared error	• Based on minimization of objective function	• Sensitive to outliers in the datasets.
		• Search based	• Uses ML model and optimization techniques to find NC automatically.Good exploration ability.	• Require tuning of parameters.
		• Miscellaneous	• Miscellaneous clustering techniques include different types of data such as time series, streaming and multiview.	• Time and memory constraints.
	Soft clustering	• Fuzzy clustering	• Uses fuzzy logic to find the membership. Sophisticated approach for processing unlabeled data with outliers and unusual patterns.	• Complex calculations, unable to handle high-dimensional datasets.
Hybrid		• Clustering and optimization	• Novel approach.Integrate the diverse approaches to overcome the limits of clustering methods.	• Require careful assessment before applying.

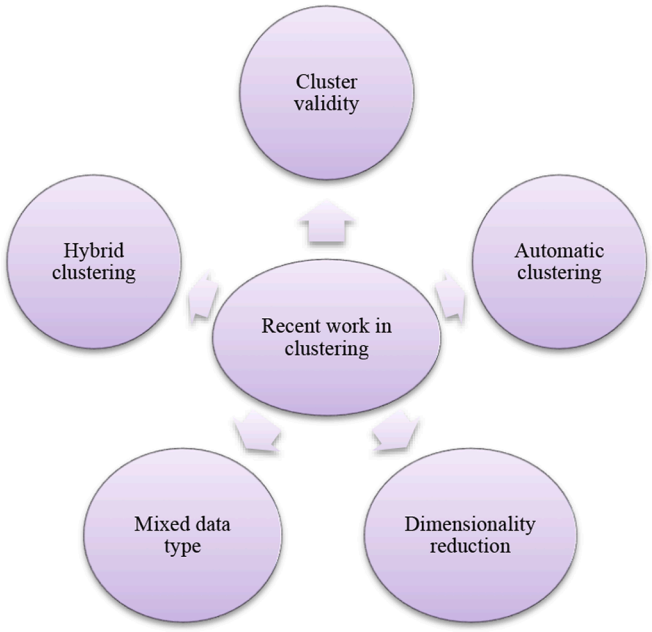


Fig. 9. Recent work in clustering.

datasets might be challenging [4]. For example, traditional KM, k-medoids, and Chameleon algorithms require human intervention to partition the datasets by providing the number of partitions in advance. This may cause problems where the number of partitions is unknown [117,122]. In order to overcome this issue, the idea of automated data clustering methods has been developed.

Duan et al. [8] proposed an enhanced affinity propagation based on optimization of preference for automated clustering to solve two problems in affinity propagation (AP): The first is the challenge of handling high-dimensional, complicated data distribution, and the second is the choice of the right preference. The proposed technique is able to solve difficult task of determining out the preference of affinity propagation and the poor clustering impact for non-convex data distribution on high-dimensional data. To make the original data lower dimensional, t-distributed stochastic neighbor embedding is used. The equilibrium optimizer algorithm and crisscross approach are combined to create a hybrid equilibrium optimizer based on the crisscross strategy (HEOC) optimization method, which improves the capability of global search. Ikotun et al. [7] proposed a novel Firefly KM clustering for enhancing the FA KM algorithm to automatically cluster high-dimensional datasets. The central limit theorem concept was added to the hybrid algorithm's KM phase to reduce the number of distance computations required by the KM algorithm, as well as mutation probability to improve exploitation and exploration for optimal data clustering. The performance of the suggested technique is examined using multiple high-dimensional datasets from the UCI repository. The results indicated that hybrid FA KM clustering method has low computational time compared to (SOS KM, PSODE, FADE and IWODE) advanced hybrid search variants. Tong et al. [117] proposed a method for the automatic determination of the NC using a density-peak-based clustering algorithm. The proposed algorithm consists of four steps: pre-clustering, merging process, acquiring the initial cluster centers and acquiring the final clusters using merging. The authors used Euclidean distance to calculate the distance of each data point, a continuous Gaussian kernel function to design a density metric to overcome the points overlapping of different data points, and finally, a merging scheme from initial clusters to obtain the final cluster automatically using the proposed algorithm.

5.3. Dimensionality reduction

Clustering algorithms typically discover similarity between objects using distance functions. However, as the number of features rises, all pairs of points appear to be equidistant in higher dimensions, making it pointless to distinguish between the closest and farthest point in particular. Therefore, dimension reduction is one of the most intriguing topics among the researchers [236]. Dimensionality reduction techniques (DRTs) are innovative and essential tools in the domains of data analysis, data mining and ML. DR is the technique of minimizing information loss while minimizing the number of characteristics (or dimensions) in a dataset. They provide a technique to comprehend and discover the useful patterns in large and complex datasets by minimizing the number of attributes. This can be done for a number of purposes,

including simplifying a model, improving the efficiency of a learning system or facilitating data visualization. The amount of experimental data available has exploded in recent decades especially in the healthcare, production, IoT devices and experimental life sciences such as biology and chemistry etc. [236]. The complexity of laboratory equipment, which can record hundreds or thousands of measurements for a single experiment, makes it difficult for statistical methods to handle such high-dimensional data. Even so, a large portion of the data is incredibly redundant and may be effectively reduced to a much smaller set of variables without substantially losing information [8]. PCA, factor analysis, linear multidimensional scaling, Fisher's linear discriminant analysis, canonical correlation analysis, maximum autocorrelation factors, slow feature analysis, adequate DR, incomplete independent component analysis, linear regression and distance metric learning are surveyed related to DR [237].

Kabir et al. [238] studied the impact of several DRTs, including PCA, PCA with a kernel and auto encoder, on ML models for reducing the dimensionality of RNA sequencing data. The original, dimensionally reduced, and cancer-relevant data were used to train and evaluate two ML classifiers, the neural network and SVM. The authors' major goal was to investigate the effects of DRTs on ML models for the detection of prostate cancer. The effectiveness of classifiers was evaluated using a variety of measures, including precision, F-Measure, recall, accuracy, ROC curve, and AUC. Mostafa et al. [239] proposed a modified metaheuristic algorithm called mCOOT for DR. The COOT algorithm mimics the movement of American coots in a lake or sea. The proposed algorithm is based on two techniques: opposition-based learning (OBL) & orthogonal learning (OL) to overcome the problem of stuckness at the local level. Reddy et al. [236] studied four well-known ML algorithms—Decision Tree Induction, SVM, Naive Bayes Classifier, and Random Forest Classifier—are examined on the basis of two popular DRTs, Linear Discriminant Analysis (LDA) and PCA, using publicly available Cardiotocography (CTG) dataset from the University of California and Irvine ML Repository. Wang et al. [240] represented the pictorial representation of the four-dimensional characteristics of different Airbnb typologies in a two-dimensional space using PCA.

Cos et al. [241] analyzed the different DRTs such as PCA and its different versions, isometric mapping, locally linear embedding and Laplacian Eigenmaps. Techniques based on dictionaries and projections have developed quite quickly in the last ten years in terms of new methodologies and the application of those methods to real problems [241]. During the late 1990s, a lot of innovative techniques have emerged, and nonlinear DR also known as manifold learning has gained popularity. Examples of recent developments that contribute to this explosive expansion include the usage of novel metrics such as the geodesic distance and graphs to depict the manifold topology. Moreover, new optimization strategies based on spectral decomposition and kernel techniques gave rise to spectral embedding, which incorporates many of the most recently created techniques. For artificial and real-world challenges, the effectiveness of nonlinear approaches is examined and discussed methods for improving the effectiveness of nonlinear DRTs [242]. The efficiency of these DRTs may be tested using high-dimensional data, such as images, text data, and other sorts of data. Moreover, more complicated algorithms such as DNNs, convolutional neural networks (CNNs), RNNs etc. can be employed with these techniques.

5.4. Mixed data type clustering

Conventional data clustering approaches aim to discover groups in numeric data [243]. Despite the fact that many modern applications produce mixed data; the majority of clustering algorithms were originally intended for pure numerical or pure categorical datasets. It raises the issue of how to effectively group objects without losing information by integrating different types of attributes. The analysis of such data has raised new challenges, necessitating the creation of novel statistical methods and procedures.

Behzadi et al. [244] studied on mixed data type using clustering algorithm for mixed-type data including concept trees (ClicoT) algorithm. ClicoT is a top-down clustering technique without parameters. With the use of a minimum description length (MDL) based objective function; ClicoT unifies categorical and numerical characteristics. The authors' tested ClicoT's interpretability and clustering quality using real-world datasets. Several experiments using both synthetic and actual datasets show that ClicoT is noise-resistant and produces findings that are simple to comprehend quickly. Data mining approaches may assist companies in the insurance sector acquire competitive advantage. By using data mining techniques, businesses may fully use data on consumer purchasing trends and behaviour as well as acquire a deeper understanding of their industry to assist reduce fraud and improve risk management. The success of a life insurance company depends on business intelligence methods [245]. The purpose of life insurance is to pay out benefits in the case of any risk, such as a premature death, which is naturally impossible to foresee with certainty. Yin et al. [245] presented a study that used mixed-type variables (numerical, categorical, and spatial attributes) and the k-prototype clustering approach to draw conclusions from a life insurance dataset. The authors used different similarity functions for different types of data such as Euclidean distance for numerical attributes, simple-matching distance for categorical attributes and distance measure for spatial attributes. Szepannek et al. [243] proposed a novel method for clustering mixed-type data on the basis of Huang's k-prototypes technique that can deal missing values. Similar to the KM algorithm, the technique iterates by minimizing within cluster sum of distance for category and numeric variables respectively. In order to use it, two hyperparameters must be specified: the NC and a second parameter which controls how the various data types interact while calculating distance.

5.5. Hybrid clustering

The prevalence of imbalanced distribution, a defining feature of high-dimensional data, is becoming more evident in a variety of big data applications. At the same time, the majority of the current feature selections and clustering techniques are built on maximizing clustering accuracy. Also, the hybrid technique successfully addresses the issue of imbalanced data clustering [246]. Nowadays, the growth of numerous social media platforms has established online social networks (OSNs) for comfortable communications and quick information sharing. However, when information is shown to be incorrect, misleading or unsuitable, it poses a major danger to social stability and public safety. Consequently, it is important to accurately and timely identify unauthorized contents in order to effectively and continuously stop rumor propagation. Hu et al. [247] proposed hybrid clustered shuffled frog-leaping and PSO method for quickly and consistently stopping the propagation of rumors in online social networks (OSNs). Clustering of imbalanced datasets is more difficult than clustering of balanced datasets because the model shows bias in the majority of class samples. Mirzaei et al. [248] proposed a hybrid method called clustering and density-based hybrid (CDBH) to handle the imbalanced dataset in clustering problem. The proposed method firstly applied the well-known clustering algorithm i.e. KM on the minority of class samples and obtained the density of each clusters. After that new minority samples are generated by the denser minority samples. Finally, denser majority samples are selected from training set and other samples are removed to balance the datasets. The proposed method can convert two-class imbalanced datasets into balanced datasets and tested on several imbalanced datasets obtained from the UCI repository. Table 7 shows the recent areas in clustering, their applications and the number of citations and Table 8 represents the summary of recent work in clustering.

Table 7
Summary of recent work in clustering.

Recent areas	Proposed works	Years	Application areas	Citation (2023)
Cluster Validity	Li et al. [229]	2019	The authors proposed a cluster validation approach that uses RDSED to dynamically determine the approximately optimal NC.	23
	Chowdhury et al. [227]	2020	The proposed method used in 2D and 3D real-life remote sensing data.	57
	Wang et al. [232]	2023	Concurrently searching for the correct cluster center in four real datasets: Bupa lever dis-order, iris, Wisconsin diagnostic burst cancer, and Wine.	2
	Patil et al. [233]	2019	Includes total 18 real world datasets: Face images, Iris, Wine, Seed, Flame Pathbased, Spiral, Stampout, Jain, R15 Breast Cancer, Pima, Aggregation, Pen Dim032, Shapes, Land Area, Shuttle	81
Automatic Clustering	Duan et al. [8]	2023	Optimize problem of non-convex data distribution. Real world datasets: Wine, Seeds, Glass, Jain, Flame, Aggregation	10
	Ikotun et al. [7]	2022	Bridge, D31, Dim1024,Housec5, Housec8Letter, Yeast	3
	Tong et al. [117]	2021	Flame, Jain, Aggregation, Smile, Double-Circle, Heart, Wine, Soybean, Spiral, Iris.	30
Dimensionality Reduction	Kabir et al. [238]	2023	Healthcare: RNA sequencing data.	25
	Mostafa et al. [239]	2022	Nine datasets from UCI ML repository have been used. CongressEW, IonosphereEW BreastEW Lymphography, SonarEW, PenglungEWVote, WineEW, Zoo.	26
	Reddy et al. [236]	2020	IDS datasets Diabetic Retinopathy (DR)	665
Mixed data type clustering	Behzadi et al. [244]	2020	MPG, Automobile, Adult datasets and Airport dataset.	12
	Yin et al. [245]	2021	Risk Management Life Insurance	7
	Szepannek et al. [243]	2018	Biology seed	80
Hybrid clustering	Zhang et al. [246]	2023	Lenses, Lymphography, Arrhythmia Bach-Chorales, Connect-4,Covertype	2
	Hu et al. [247]	2023	Online social networks	8
	Mirzaei et al. [248]	2020	44 Imbalanced datasets selected. Ecoli034vs5, Glass5, Yeast2vs4, Shuttlec0vsc4, Yeast5, Ecoli067vs35, Ecoli0234vs5, Glass015vs2, Yeast0359vs78, etc.	51

Table 8
Summary of recent work in clustering.

Study	Advantage	Limitation
Cluster validity	Useful in finding compact and well-separated clusters.	• Validation metrics require a minimum of two clusters to calculate inter cluster similarity. • Computationally expensive for a larger number of clusters.
Automatic clustering	Estimating NC automatically.	Challenging for high dimensional. datasets
Dimensionality reduction	• Techniques for compressing datasets into desire dimensions. • Useful to analyze high dimensional datasets.	Results in loss of information.
Mixed data type clustering	Cluster of dataset comprises bothcategorical and numerical value.	Challenging to apply direct mathematic functions and computing similarity measures between mixed features.
Hybrid clustering	Hybrid technique successfully addresses the issue of an imbalanced dataset for better data clustering.	• A hybrid algorithm combines concepts and procedures to balance the strengths and limitations of several approaches. Implementing a hybrid clustering approach requires expertise in multiple clustering techniques.

6. Challenges in clustering

Data clustering is a useful approach in data mining and knowledge discovery that forms a group of unlabeled data points based on their degree of similarity. In other words, the clustering process extracts the cluster characteristics and cluster count from data that are not known a priori. The steps in a typical clustering process include feature selection, measure selection, data grouping and output evaluation. In order to obtain optimal clustering results, cluster analysis faces different challenges at different phases. No one technique can be expected to work effectively for all types of data since clustering techniques are subjective in nature. There are many challenges involved with regards to selection of parametric characteristics when addressing the clustering problems [13]. These parametric characteristics are broadly classified into two categories (1) Choice of datasets (2) Selection of computational methods as shown in Fig. 10 [249].

The first step in pattern extraction and knowledge discovery is data selection, which specifies the scope of the data that will be used to extract information. The data is analyzed and preprocessed to increase

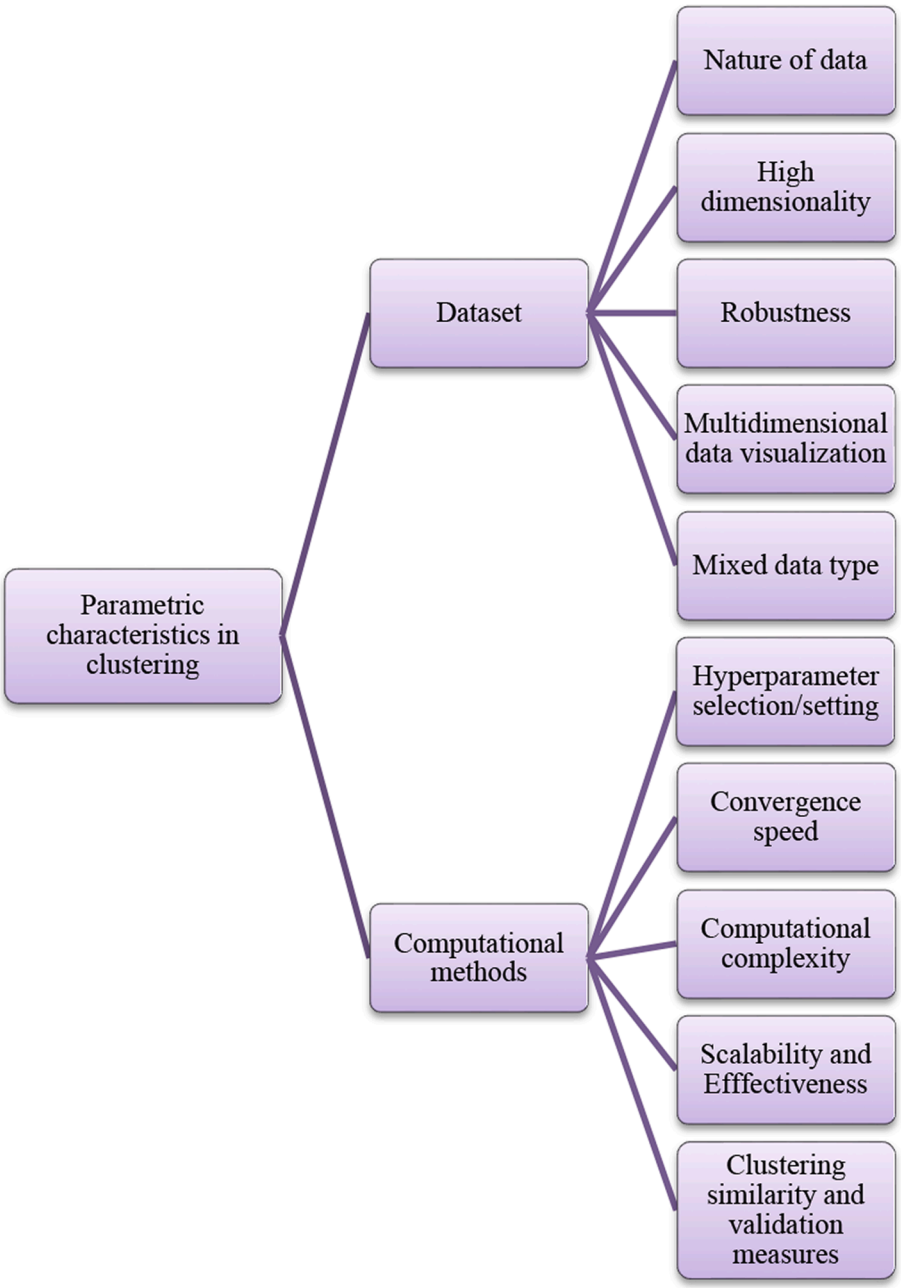


Fig. 10. Classification of parametric characteristics of clustering.

the reliability of pattern extraction, by eliminating unnecessary data, addressing missing values, and removing outliers from the data. The preprocessed data is transformed into a meaningful format, which includes data normalization, sampling, and feature selection for the data mining algorithms to process [250]. Finally, result evaluation and interpretation take place for knowledge extraction. The data mining and knowledge discovery process are shown in Fig. 11 [22,23,174]. Parametric characteristics play a vital role in clustering performance. Many parametric characteristics and challenges associated with the clustering problem are discussed below.

6.1. Dataset

Data preprocessing is an extensive field in data mining that brings together methods from several disciplines to improve the quality of datasets for learning and knowledge extraction tasks. The five Vs-volume, velocity, veracity, variety and value are always used to describe big data. Nearly every data model used to represent large data is dependent on these 5-vs traits. Many studies on velocity and volume have been conducted, but there is still no complete and efficient solution for variety. Today's organizations face substantial hurdles due to data that is increasingly diverse and more complex in form (unstructured/semi-structured), as well as concerns with indexing, sorting, searching, analyzing, and visualizing.

Real-world datasets are not perfect. Common data quality challenges include missing and noisy data, outliers, inconsistent, redundant, or biased data. These are common data quality concerns that need to be addressed [251]. Data preprocessing is an essential step in data clustering which provides final datasets that may be considered as reliable and useful for subsequent data mining algorithms. The goal of data processing is to reduce the data size, find the co-relation between data, normalization of data, removing outliers and extraction relevant

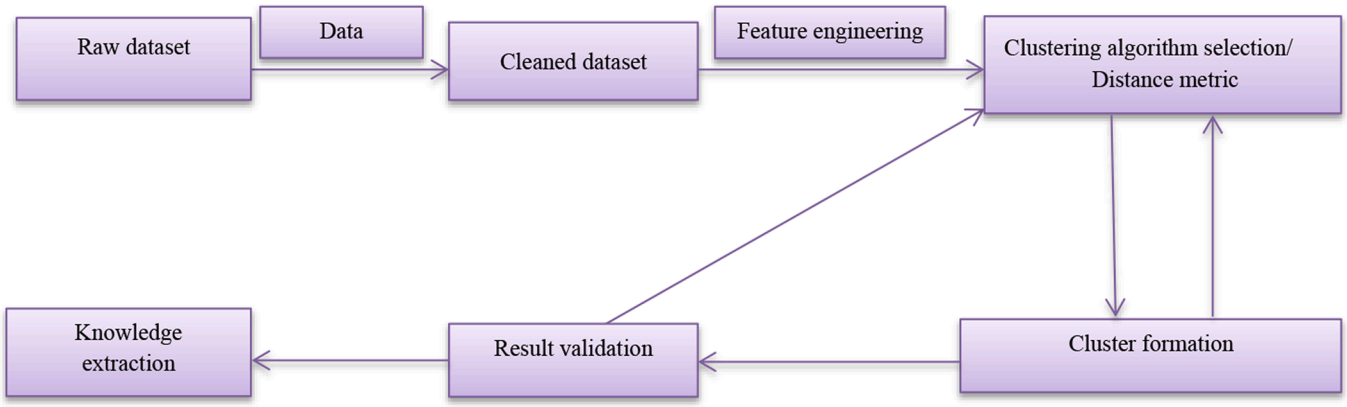


Fig. 11. The data mining and knowledge discovery process.

features for knowledge discovery. Various strategies, such as missing value imputation, transformation, integration, and reduction, have been developed to preprocess data on the dataset before clustering [252]. Datasets differ in a number of ways. The tools and techniques available for data analysis are determined by the type and format of the data [251]. Commonly used different types of data formats in clustering are shown in Fig. 12 [13]. This following sub-section presents various datasets related issues in cluster-analysis.

6.1.1. Nature of the data

Many high-dimensional and high-volume datasets have been created as a result of the quick development of clustering applications such as digital imaging, internet search, video surveillance etc. The nature of the data plays a crucial role in cluster validity, robustness and the selection of algorithms. To make the data comprehensible and helpful for the knowledge extraction process, cleaning and preparation are necessary. Both the standard KM algorithm and the fuzzy KM algorithm can only work with numerical data. Categorical data, however, are often used in real-world applications, particularly in the rapidly developing field of social media analysis [251]. Huang expanded these two existing methods for categorical data clustering and created the well-known fuzzy k-modes and k-modes algorithms [253–255]. There are some challenges involved in a nature of datasets such as missing value imputation, outliers, DR, distribution of data, handling of mixed type data, feature selection and extraction that must be considered before applying clustering algorithms.

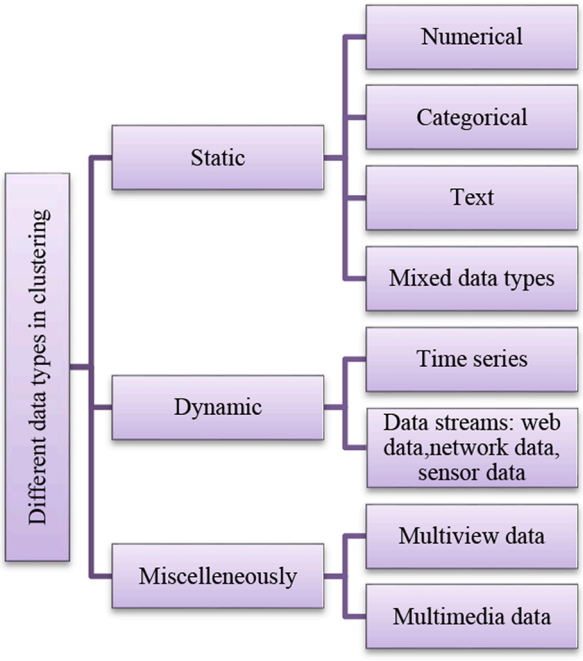


Fig. 12. Different data types in clustering.

6.1.2. High dimensionality

Data objects are currently characterized by several attributes or dimensions in different application domains. Datasets from the real world are extremely sparse and are all almost equally spaced apart in a very high-dimensional feature space [256]. High-dimensional dataset poses challenges in data mining such as cluster analysis, document-clustering and time series analysis [37]. High-dimensional data, i.e., data represented by a large variety of attributes, presents special challenges to clustering i.e. computational problem, risk of overfitting [257], nearest neighbor and density analysis [37,256]. All distance metrics become similar in high-dimensional spaces [37]. Therefore, High-dimensional data clustering cannot be accomplished using conventional distance- and density-based similarity evaluation techniques. The mutual (dis-) similarity is effectively captured by conventional clustering techniques based on distances or density measures in low-dimensional spaces, but they fall short in high-dimensional areas due to diminishing differences. There are dozens to hundreds of dimensions involved in high-dimensional dataset clustering. However, regarding the minimal number of dimensions that makes a dataset high-dimensional is not mentioned clearly [37,169]. There is no agreed-upon standard in the literature regarding high-dimensional datasets.

6.1.3. Robustness

Some clustering algorithms, such as KM and FCM, are extensively used in the literature, but both are sensitive to noise and outliers in the dataset [235]. By robustness mean that noise and outliers shouldn't cause an algorithm's performance to substantially degrade and that it shouldn't be adversely affected by even little deviations from the anticipated model [258]. Since the real world data is certain to have noise and outliers. Therefore, data cleaning is necessary in data mining. These anomalies can be contamination that is introduced at various phases of measurement, storage, and processing. Robust statistics searches for the model that best fits the majority of the data in order to find the outliers. Hence, clustering techniques must be robust enough to identify the noise and outliers in the dataset in order to applicability in real world problems [258].

6.1.4. Multidimensional data visualization

N-dimensional datasets (N>3) are often hard to present on a two-dimensional screen or piece of paper [259]. Data visualization tools transform raw and complex numerical and non-numerical data into visual summary that is easy to understand and comprehend. Unsupervised

learning models are employed in complex techniques across many different fields, including medicine, bioinformatics, and other disciplines. Therefore, data visualization is useful to clean data, explore data structure, find outliers and unusual groupings, identify trends and clusters, notice local patterns, assess model output, and display results using attractive graphics.

Many visualization methods have been created over time in order to depict and analyse massive amounts of data. These techniques include histograms, line charts, tables, pie charts, bar charts, scatter plots, bubble plots, area plots, flow charts, Venn diagrams, time lines, multiple data series, entity relationship diagrams, cone trees, semantic networks, tree maps, and parallel coordinates etc. [260].

6.1.5. Mixed data type

Mixed data comprises both numerical and categorical information, and it is widely used in a variety of industries, including marketing, finance, and healthcare. Clustering is a prominent data mining activity, and as mixed objects are regularly encountered in real-world datasets, grouping mixed datasets into meaningful groups is practically beneficial [63]. Although it may be challenging to directly perform mathematical operations, such as adding or averaging, to feature values of mixed data type [261]. It is expected that clustering algorithms will be adaptable enough to perform with any data type that the dataset contain [4].

6.2. Computational methods

In several fields of study, particularly data mining, clustering algorithms have emerged as one of the most important research area. No one approach can be anticipated to perform effectively for all types of data therefore, the algorithm used for clustering analysis is largely influenced by the type of data [13]. A range of clustering strategies, methodologies, and algorithms are proposed and implemented, with limitations. The enormous amount of data that is referred to as “Big Data” is increasing day-by-day makes the majority of conventional clustering techniques computationally expensive. Therefore, a linear time complexity or nearly linear time complexity algorithms are highly desirable [4]. There are some useful characteristics that need to be considered while designing novel clustering techniques. These techniques are important to deal with the running-time complexities of clustering algorithms, as discussed in subsequent sub-sections.

6.2.1. Hyper parameter selection/setting

The best algorithm and parameter setting must be chosen while addressing a clustering problem [262]. For example, the k-mean clustering approach, which has been utilized to address a variety of clustering issues, is the most widely used and simplest clustering algorithm. However, the user must first define k (NC) prior to clustering [36,251]. Choosing the right number of partitions (K) prior to clustering is a bit challenging. Therefore, a few automatic data clustering algorithms have been used recently and the majority of them are motivated by either natural or physically occurring phenomena, such as PSO, bee colony optimization algorithm, bacterial evolutionary algorithm, gravitational search algorithm etc. Clustering analysis is seen as an optimization problem that involves maximizing dissimilarity across clusters and minimizing dissimilarity within a cluster. The selection of the right hyperparameters can be used to optimize the optimization function [263].

6.2.2. Convergence –speed

The convergence to the local optimum and slow convergence velocity are two major clustering issues that have been addressed by employing two concepts from chaos theory and the acceleration approach [264]. Several nature inspired metaheuristic techniques with some innovations are used to improve the selection of initial centroids in KM [35]. A fast FCM technique has been reported by implementing the new rule of updating the cluster center in each iteration. A swarm-based optimization algorithm called the normative fish swarm algorithm (NFSA) is suggested as a useful global and local search method to find useful global optima with a faster convergence rate. A number of metaheuristic algorithms, inspired by both human and natural processes, have been used to successfully handle clustering problems [11].

6.2.3. Computational complexity

Despite their effectiveness, certain clustering techniques could be too computationally demanding to be used to huge datasets with a high-dimensional feature map. High-capacity GPUs can be used to improve the issue by increasing the output of computational resources [265].

6.2.4. Scalability and effectiveness

Clustering large datasets in a reasonable amount of time is a challenging task in big data mining. Therefore, there is a huge scope for improvements in terms of effectiveness and scalability. The capacity of an algorithm to manage an increasing amount of input by adding more resources to the system is known as scalability. The traditional methods use clustering algorithms without taking the system’s scalability into account. In contrast, scalable approaches make use of the system’s scalability using clustering algorithms. KM, k-modes and CLARA have high scalability whereas SLINK, PAM, k-prototype and K-medoids have low scalability [13]. There are two techniques to make clustering algorithms more scalable: (i) reducing the size of the dataset, and (ii) employing several physical processors to analyse distinct dataset subsets concurrently.

6.2.5. Clustering similarity and validation measures

The purpose of clustering is to decompose an unlabeled dataset in a well-organized way into similar groups in such a way that data points are more similar inside the group than across the groupings [266]. Similarity measures play a vital role in discovering compact and well-separated cluster prototypes [1]. Moreover, most of the clustering algorithms extract the subgroups from the entire dataset by making some predefined assumptions. These types of clustering algorithms initiate some sort of validation when considering the different subgroups present in the entire dataset [267]. Therefore, an essential step in the clustering process is validating the outcomes of the clustering algorithms [228]. The next subsections describe the various cluster similarity and validation measures used in both traditional and recently proposed - clustering algorithms.

6.2.5.1. Clustering similarity measures. A technique of determining the degree of association between the elements in a dataset is needed in order to cluster them. Similarity measures of clustering represent the degree of cohesiveness in an intra-cluster or separation in inter-cluster characteristics [4,7,12]. Distance metrics play an important role in accessing the similarity among the data objects, therefore distance-based clustering is a highly common method for grouping objects and has produced successful results [266,268]. The input data has a significant impact on how well existing clustering approaches perform. Various datasets often need numerous separation techniques and similarity metrics [269]. The clustering outcome might be impacted by the similarity measures chosen. This subsection provides an overview summary of the similarity measures that are often employed in both traditional and newly introduced clustering approaches.

- **Euclidean distance***:* Euclidean distance (ED) is a standard similarity metric used between data points that is applicable to numeric data of any dimension [1]. Moreover, it is also the default distance metric used by the KM clustering algorithm [4]. The ED determines the root of square differences between the coordinates of a pair of objects and can be generalized to higher dimensions as shown in Eq. (7) [268,270].

$$Dist_{X,Y} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \quad (7)$$

where $X_i and X_j$ are two objects in cluster k.

- **Manhattan distance:** The Manhattan distance is determined as the total of the absolute differences between the two objects X and Y shown in Eq. (8) [271].

$$Dist_{X,Y} = |X_{ik} - X_{jk}| \quad (8)$$

- **Chebyshev distance:** It represents the maximum separation between any two points X and Y in a single dimension. The formula for the Chebyshev distance between the two points is shown in Eq. (9) [272].

$$Dist_{X,Y} = max_k |X_{ik} - X_{jk}| \quad (9)$$

- **Minkowski distance:** A generalized measure known as the Minkowski distance, or derived from norm so- called L_p [273]. Manhattan, Euclidean and Chebyshev distances are special cases of Minkowski distance and can be obtained by setting $p = 1$, $p = 2$ and $p=\infty$ respectively [272]. Minkowski distance metric represented by eq. (10).

$$Dist_{X,Y} = \left(\sum_{k=1}^{d} |X_{ik} - X_{jk}|^p \right)^{1/p} \quad (10)$$

- **Cosine distance:** Cosine distance determines the similarity of two patterns by calculating the cosine angle between them. In text mining, it is frequently used to compare documents by calculating the angle between two vectors of n dimensions shown in Eq. (11) [274].

$$\ddot{I} = \frac{\cos^{-1}(A.B)}{\|A\|\|B\|} \quad (11)$$

- **Average distance:** In ED, the largest feature values would be dominant over the others. To improve the clustering results, an average distance is an updated version of ED. The average distance between two data points X and Y in n-dimension can be defined as Eq. (12) [275].

$$Dist_{X,Y} = \sqrt{\frac{1}{n} \sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \quad (12)$$

- **Weighted Euclidean distance:** It is a kind of ED that calculates the similarity between patterns when the weight (importance) of each feature is specified as shown in Eq. (13) [275]. Min Ji et al. [276] proposed a dynamic fuzzy cluster technique for time series using this metric.

$$Dist_{weighted} = \sqrt{\sum_{k=1}^{m} w_i (X_{ik} - X_{jk})^2} \quad (13)$$

- **Chord distance:** Chord distance is a variation in ED that can address the problem caused by the scale of measurements shown in Eq. (14). The chord distance is the length between two normalized points on a hypersphere with radius one [275].

$$D_{chord} = \left(2 - 2 \frac{\sum_{i=1}^{n} x_i y_i}{\|x\|_2 \|y\|_2} \right)^{1/2} \quad (14)$$

where $\|x\|_2$ is the L^2- norm .$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

- **Mahalanobis distance:** In contrast to Euclidean and Manhattan distances which are independent of the linked dataset to which two data points belong, the Mahalanobis distance is a data-driven metric [275]. By employing the squared Mahalanobis distance or performing a whitening adjustment to the data, Mahalanobis distance can reduce distortion brought on by linear correlation among features [31].The mathematical form of Mahalanobis distance is defined in Eq. (15). Mahalanobis distance is useful in data classification and clustering [277]. Using the Mahalanobis, Manhattan and Bhattacharyya distance approaches, the performance of gender clustering is examined on the Harvard-Haskins Database [278].

$$D_{mah} = \sqrt{(x-y) S^{-1} (x-y)^T} \quad (15)$$

where S is the covariance matrix of the dataset.

- **Multi view point based similarity measure**: In document clustering, where several perspectives may be employed to generate a more meaningful assessment of similarity, multi viewpoint-based similarity measure has good advantages. Multi viewpoint-based similarity measure is defined by Eq. (16) [4].

$$MVS(d_i, d_j | d_i, d_j \in S_r) = \frac{1}{n - n_r} \sum_{d_h} \cos(d_i - d_h, d_j - d_h) \|d_i - d_h\| \|d_j - d_h\| \quad (16)$$

where d_i and d_j are points in clusters S_r and d_h is view point.

- **Bilateral Slope-Based distance:** In time series clustering [279], a new suggested similarity metric is Bilateral Slope-Based distance (BSD). It combines a simple time series representation, the slope of each time series segment, ED and so called dynamic time warping. The definition of the BSD is shown in Eq. (17).

$$D_{BSD}\left(TS_i^{(1)}, TS_j^{(2)}\right) = \left| x_i^{(1)} - x_j^{(2)} \right| + \left| \sin\theta_i^{(1)} - \sin\theta_j^{(2)} \right| + \left| \sin\theta_{i-1}^{(1)} - \sin\theta_{j-1}^{(2)} \right| \quad (17)$$

- **Pearson correlation:** In order to cluster gene expression data, pearson correlation (PC) is frequently utilized [23]. This similarity metric determines how similar two gene expression patterns' shapes are. The PC defined by Eq. (18) where μ_x and μ_y are the means for x and y respectively [275].

$$Pearson(x,y) = \frac{\sum_{i=1}^{n} (x_i - \mu_x)\left(y_i - \mu_y\right)}{\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}} \quad (18)$$

- **Jaccard distance:** Jaccard distance, a traditional similarity measure on sets, useful in information retrieval, data mining, ML, and many other fields can all benefits from the use. The Jaccard distance between two finite sets A and B shown in Eq. (19) [280].

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \quad (19)$$

6.2.5.2. Cluster validation measures. One of the key concerns for clustering applications is clustering validation [281]. Cluster validation is a method for identifying a set of clusters that most closely matches natural partitions (NC) without using any class label [282]. One of the challenging issues necessary for the success of clustering applications is clustering validation, which measures the goodness of clustering results [283,284]. There are three groups of validity measures proposed for clustering algorithms: internal, external, and relative validations [285]. The two primary types of clustering validation are external clustering validation and internal clustering validation [282,283]. Both types consist of a variety of clustering validation indices, which are discussed

below in the next subsection. Relative validation is based on comparing partitions produced by the same algorithm with various parameter values or data subsets of datasets [285].

(i) Internal clustering validation criteria

Internal validation indexes have been employed recently to evaluate clustering results to determine the optimal NC. Internal cluster validation depends on the two criteria: compactness and separation [285–287]. Compactness measures the degree to which the objects in a cluster are related to one another. A collection of metrics evaluates cluster compactness using variance. Compactness is improved by lower variation. Additionally, a variety of distance-based methods are available for evaluating the cluster compactness including maximum or average pairwise distances and maximum or average center-based distances. Separation measures how effectively or clearly differentiated one cluster is from other clusters. Examples of metrics of separation include the pairwise distances between cluster centers and the pairwise minimum distances between objects in various clusters [287]. A brief mathematical definition of common internal clustering validation criteria is discussed below.

- **Sum of square error:** The most frequently used criteria metric for clustering is sum of square error (SSE) [1]. It's described as follows shown in Eq. (20).

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2 \tag{20}$$

where C_k is the collection of instances that constitute cluster k and μ_k is the cluster's vector mean.

- **Silhouette index:** A compact and well-separated cluster is identified using the silhouette coefficient [10,288,289]. Maurice Roux stated that divisive algorithm based on the silhouette index performs well with both synthetic and real-world datasets [47]. The silhouette index is defined as follows shown in Eq. (21).

$$S(i) = \frac{(b(i) - a(i))}{Max(a(i), b(i))} \tag{21}$$

where $a(i)$ represents the average distance between the i^{th} sample and all samples contained in $X_j (j = 1, \cdots C)$; $b(i)$ is the minimum average distance between the i^{th}s and all of the samples clustered in $X_k (k = 1,..c;\ k \neq j)$.

- **Dunn index:** In a partitioning, the Dunn index essentially finds for the ratio between the smallest intra-cluster distance and the largest inter-cluster distance. It is defined as follows shown in Eq. (22) [136].

$$Dunn = \min_{1 \le i \le c} \left(min \left(\frac{d(c_i, c_j)}{\max_{1 \le k \le c} d(X_k)} \right) \right) \tag{22}$$

where $d(c_i, c_j)$ is the distance between cluster X_i and X_j; $d(X_k)$ represents the distance between members of cluster (X_k) and c is the NC in the dataset.

- **Davies – Bouldin index (DB):** The DB index calculates the average similarity between any two clusters and their closest neighbors. The purpose of this index is to find compact and well-separated groupings of clusters. DB index is defined as follows shown in Eq. (23) [4,282].

$$DB = \frac{1}{c} \sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{d(x_i) + d(x_j)}{d(c_i, c_j)} \right\} \tag{23}$$

where c is the NC;i, j are the labels for the clusters, $d(x_i)$ and $d(x_j)$ are all entities in clusters i and j; $d(c_i, c_j)$ is the distance between the cluster centroids.

- **Calinski-Harabasz index:** By computing the distances between cluster points and their centroids, the calinski-harabasz validity index measures how compact or close the clusters are. This index is computed as shown in Eq. (24) [4].

$$CH = \frac{trace(S_B)}{trace(S_w)} \cdot \frac{n_p - 1}{n_p - k} \tag{24}$$

- **Bayesian information criterion (BIC):** To prevent overfitting, BIC [22] was developed and is defined as shown in Eq. (25).

$$BIC = -ln(L) + vln(n) \tag{25}$$

- **Novel validity index (NIVA index):** Novel validity index can be defined as shown in Eq. (26) [282].

$$NIVA(C) = \frac{Compac(C)}{SepxG(C)} \tag{26}$$

where $Compac(C)$ is the average compactness of the cluster C and $SepxG(C)$ is the average separability of cluster C.

- **Scatter criteria:** The scatter criteria can be defined as shown in Eq. (27) [1].

$$S_k = \sum_{x \in c_k} (x - \mu_k)(x - \mu_k)^T \tag{27}$$

- **Gamma Index:** For use in clustering situations, the gamma index is an adaptation of Goodman and Kruskal's gamma statistic [290]. It can be defined as shown in Eq. (28).

$$G = \frac{(S+) - (S-)}{(S+) + (S-)} \tag{28}$$

- **K-means cosine cloud similarity (K3CM):** K3CM which is statistical knowledge-based strategy for finding and grouping the most similar objects and forms cohesive clusters. This approach uses the 2CM method to determine k centroids in the first phase and the CK2M algorithm to refine the clusters in the second phase. The primary idea behind this similarity is to measure two n-dimensional vectors based on their angles. It can be formulated as shown in Eq. (29) [291]. K3CM is versatile in nature, which means it may be applied to various dataset to produce finer clusters. This quality makes it useful in cluster analysis and other related applications such as pattern recognition in data mining, information retrieval and ML etc.

$$2CM(\ddot{1}) = \frac{\vec{A} . \vec{B}}{\|\vec{A}\| \|\vec{B}\|} = \frac{\sum_{i=1}^{k} a_i b_i}{\sqrt{\sum_{i=1}^{k} (a_i)^2} \sqrt{\sum_{i=1}^{k} (b_i)^2}} \tag{29}$$

where $\vec{A} = [a_1 \cdots, a_k]$ and $\vec{B} = [b_1 \cdots, b_k]$ represent horizontal vectors.

(ii) External clustering validation criteria

External clustering validation is based on comparison of partition generated by the clustering algorithm and the true cluster number of the datasets [285]. Some of the external clustering validation criteria are discussed below.

- **Entropy:** The data mining community frequently used the entropy and purity measures as external validation methods derived from the information retrieval community. The clusters' class labels' purity is measured using entropy. Therefore, the cluster entropy value will be 0 if all samples in the cluster have the same label. However, the entropy rises as the variety of class labels for the items in a cluster increases is defined as follows shown in Eq. (30) [282].

$$E_j = \sum_i p_{ij} log\left(p_{ij}\right) \tag{30}$$

- **Purity:** Entropy and purity have a lot in common. Purity measures the class with majority objects within cluster [292]. The purity of each cluster P_j is defined as shown in Eq. (31).

$$P_j = \frac{1}{n_j} max_i\left(n_j^i\right) \tag{31}$$

The total purity of the clustering solution is calculated as the weighted sum of the purities of each individual cluster and is given as follows shown in Eq. (32) [282].

$$Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j \tag{32}$$

In general, better clustering performance is indicated by lower entropy or higher purity values [292].

- **Fowlkes-Mallows index (FMI):** The similarity between the groups formed after the clustering procedure is determined by the FMI. More similarity between the clusters is indicated by a higher FMI value. The FMI is defined as shown in Eq. (33) [1].

$$FM = \sqrt{\frac{TP}{TP+FP}\frac{TP}{TP+FN}} \tag{33}$$

- **NMI measure:** The NMI of two objects can be defined as follows shown in Eq. (34) [4].

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{34}$$

where I(X,Y) denotes the mutual information between two random variables X and Y and H(X) denotes the entropy of X, X represents cluster prototypes while Y represents actual labels.

- **Jaccard Index:** The Jaccard similarity coefficient, commonly referred to as the Jaccard index, is a statistic used to assess the diversity and similarity of sample sets. It can be defined as follows shown in Eq. (35) [1].

$$JJ(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{TP}{TP+FP+FN} \tag{35}$$

- **Rand Index:** The Rand index can be defined as shown in Eq. (36) [1].

$$RAND = \frac{TP+TN}{TP+FP+FN+TN} \tag{36}$$

where TP: "true positives", TN: "true negatives", FP: "false positives" and FN: "false negatives". The Rand index ranges from 0 to 1.

- **F-measure:** The F-measure index uses the weighting recall parameter $\eta > 0$ to balance the false negatives shown in Eq. (37) [1].

$$F = \frac{(\eta^2+1).P.R}{\eta^2.p+R} \tag{37}$$

where P is the precision rate and R is the recall rate.

- **Mutual Information Based measure:** The mutual information based metric seeks to assess the connection between the partition outcome and the dataset's underlying structure. The mutual information –based measure can be defined as follows shown in Eq. (38) [1].

$$C = \frac{2}{m} \sum_{i=1}^{g} \sum_{h=1}^{k} m_{i,h} log_{g.k}\left(\frac{m_{i,h}.m}{m_{.,h}.m_{i,.}}\right) \tag{38}$$

where $C = \{C_1, C_2, \cdots C_g\}$ for m instances and target attribute z with domain $dom(z) = \{c_1, c_2, \cdots c_k\}$.

7. Emerging applications of clustering techniques

Several interdisciplinary disciplines can benefit from clustering methods. This section lists notable applications of cluster analysis that are being utilized effectively based on a literature review.

7.1. Web intelligence

The term "web intelligence," which was first used in the late 1999s, refers to the study and use of ML algorithms and information technology with a particular focus on the Web platforms. Communities of international academics are paying attention to common Web intelligence applications including online text categorization, web document clustering, web recommender for e-commerce, web usage profiling, and similar knowledge discovery activities. Tang et al. [293] applied nature-inspired algorithms on data generated by web activities for web intelligence. Shafiq et al. [294] provided a web user session clustering approach using PSO. In web user session clustering, the primary sources of session data are cookies, cache servers, and web server logs are used. Data about user activity is provided in common log format (CLF) or expanded CLF format by web logs. IP address, date and timestamp, items requested, protocol, status, bytes delivered, and software agent are common entries in logs. Recommender systems are a type of web intelligence technique [295]. Abraham et al. [262] developed an ACO-based clustering to find trends of users' search history to analyze their behavior. Ben et al. [296] presented a taxonomy of recommendation system and interface of user input data in e-commerce domain. Jayalakshm et al. [213] proposed taylor horse herd optimization based deep fuzzy clustering and laplace based K-nearest neighbor. The proposed method facilitates user for web page recommendation using interesting sub-graphs.

7.2. Speech processing

Humans naturally communicate with one another through speech. Many people exchange information in a real-world environment using mobile phones and other communication devices. Sonkamble et al. [297] proposed an algorithm called Modified KM LBG algorithm used to obtain a good codebook. They employed vector quantization, which is the primary and most effective approach for speech coding, data compression, speech recognition, voice synthesis, and speaker identification. A recent development in speech processing is expressive speech modelling, which includes the synthesis and recognition of emotional speech. Fuzzy clustering, including probabilistic, possibilistic and graded-possibilistic c-means clustering is used for emotion recognition from speech signals [297]. In order to improve the models, a number of changes were also performed, including feature selection, the usage of more clusters than classes and the grouping of classes with similar traits,

and ultimately, modifying the parameters of the probabilistic models. Vani et al. [298] analyzed a comparative experimental study of homogeneous and heterogeneous speech data with clean and noisy signals. Authors used additive and convoluted noisy speech signals as an input, mel-Frequency cepstral coefficients is used for feature extraction and PCA is used for feature selection and then clustering is performed based on three algorithms KM, FCM and Gaussian kernel based FCM.

7.3. Intelligent medical science

The emergence of massive data sources, including genomes, electronic health records (EHRs), mobile diagnostics, and wearable technology, as well as advances in AI applications, have opened the way to the diagnosis and treatment of a wide range of severe diseases. Clustering has been demonstrated to be a powerful ML tool in the intelligent healthcare industry. Medical data mining has a lot of potential for revealing hidden patterns in datasets related to the medical industry. Medical diagnosis is considered an important but challenging task that must be performed precisely and effectively. Alashwal et al. [299] concentrated on analyzing and reviewing clustering methods KM, KM-Mode, multi-layer clustering and hierarchical agglomerative clustering that have been applied to the Alzheimer's Disease Neuroimaging Initiative and the UC Irvine ML Repository of neurological diseases, particularly Alzheimer's disease. Yadav et al. [300] proposed Foggy KM algorithm for clustering of Lung Cancer Data. Data obtained from Sanjay Gandhi Postgraduate Institute of Medical Science Lucknow is taken into consideration for the study. The original data were high-dimensional; however, a number of attributes and instances were ultimately taken into consideration based on the recommendations of medical experts. Three factors—the Dunn index, the silhouette value and connectedness in comparison to the conventional KM algorithm are used to evaluate the proposed algorithm's validity. Greene et al. [301] investigated seven benchmark real-world datasets (breast,diabetes heart, iris, liver, lymph and thyroid) from the UCI library and two synthetic datasets to study ensemble clustering. Skin cancer has grown more quickly over the past decade. Infection and bacteria lead to skin disorders. In order to detect skin illnesses in their early stages, many techniques have been applied to image data so far. M. Kumar et al. [302] proposed an enhanced strategy to detect three types of skin cancers in early stages using FCM, ANN and DE algorithms. Healthcare research has grown exponentially over the years and has led to significant discoveries. Therefore, there is tremendous scope in the field of intelligent healthcare.

7.4. Image processing and segmentation

Image segmentation is a key challenge in image processing. Segmentation is a process of splitting images into regions to extract meaningful information. It may be regarded as the most important and essential technique for defining, characterizing, and visualizing areas of interest in medical imaging. There are several algorithms and methods available for segmenting images. KM and FCM clustering are typically the most often used clustering-based image segmentation algorithms due to their simplicity and speedy convergence to an optimal solution [113].

Segmenting medical images is thought to be a popular study area. Several methodologies and algorithms for image segmentation have been proposed by a number of researchers. Gopal et al. [303] designed an intelligent system of image processing using FCM along with GA, and PSO to diagnose brain tumor through Magnetic Resonance Imaging (MRI). Lam. et al. [39] proposed a novel clustering algorithm: adaptive fuzzy-KM clustering for image segmentation captured using digital cameras. The proposed approach achieved the best cluster center value for a more accurate segmentation process and successfully produced better segmented images than the traditional FCM and Moving KM algorithms. Maksoud et al. [219] proposed a method on segmentation of brain tumors using a hybrid clustering method. The proposed method integrates the FCM algorithm with the KM clustering algorithm. Glavan et al. [304] developed an innovative algorithm for applying CNN to extract information from X-ray images and apply labels. The algorithm examines every pixel in the image and attempts to divide it into two categories: bone and non-bone. Yang et al. [305] proposed modified and robust fuzzy clustering algorithm with an additive Gaussian noise for image segmentation.

Despite much research, segmentation is still a challenging work due to differences in image content, cluttered objects, image noise, non-uniform object texture, and other reasons. Although there are several algorithms and image segmentation approaches available, they depend on several factors, including the optimization algorithm, hyper-parameter value, data preparation approach, loss function, and more. Therefore, a quick and effective method for segmenting medical images still needs to be developed.

7.5. Information retrieval

Relational, graphical, and textual databases have grown rapidly and proliferated over the past few decades as a result of the accessibility of affordable and efficient information systems and storage technologies. While it is easier to collect and store information, retrieving it has become much more difficult, especially for large-scale databases. Information retrieval (IR) is the study of obtaining relevant information from a large pool of data. It facilitates the user's communication with an information system to deal with unstructured data as opposed to conventional database systems. IR has been used for clustering a variety of functions, including query expansion, document grouping [306], automatic document indexing, digital library, and automatic web-search result display. Digital libraries are progressively replacing traditional library systems. Mohammed et al. [307] presented metaheuristic-based approach i.e. firefly algorithm, for document clustering. Each document in this approach is represented by a single firefly, and its total weight is equal to the firefly's starting brightness. Shi [308] applied FCM clustering algorithm to group library borrowing records and reader information in digital library management system. It enhances resource utilization, library collection organization and analysis of the reader's behavior. P. Prabhu [309] discussed KM clustering algorithm for clustering documents.

7.6. Aviation and vehicular systems

The aviation community has become more and more interested in using ML techniques for safety analysis, incident and accident investigation, and defect identification. Clustering algorithms have been utilized to tackle safety challenges in aviation and automobile, such as air traffic control, aircraft safety anomalies, risk identification, and flight anomaly detection [310]. Li et al. [311] proposed a novel clustering-method for anomalies detection in routine airline operations. The novel approach used DBSCAN clustering algorithms to identify abnormal flights of distinctive data patterns, facilitated by data from the flight data recorder. The proposed algorithm performed well in continuous parameters. Mangortey et al. [235] proposed study of flight risk management using AHC algorithms, Divisive Analysis, Self Organizing Tree Algorithm, KM clustering algorithm, PAM, CLARA and Model-based clustering algorithm. The algorithms were used to categorize related flights and detect irregularities and abnormalities. Nazeri et al. [310] performed an experiment on aviation data to analyze the severe weather pattern on National Airspace System performance. Authors used KM clustering with K value ranging between 2 and 7 to form the clusters of similar weather days. Clustering results were found meaningful and may be used to solve air traffic control issues.

7.7. Bioinformatics: Microarrays gene expression clustering

In the advancement of microarray technology, researchers are now

able to analyse, identify, and track the amounts of mRNA transcripts for thousands of genes in a single experiment [312,313]. Gene expression (GE) data are often displayed as a real-valued matrix, with the columns representing the pattern of gene expression across all microarray experiments and the row objects representing GE measurements over numerous experiments [314]. There are two viewpoints on the applications of clustering algorithms in bioinformatics. The first aspect is based on the identification of similar gene expression patterns in DNA microarray studies [76]. The second component refers to clustering techniques that act directly on linear DNA or protein sequences [1]. Tasoulis et al. [312] applied k-windows clustering technique to cluster data and calculated number of cluster from GE microarray data. Cluster analysis is the most commonly employed computational technique to analyze microarray data. When applied to GE data, clustering techniques have shown promise for locating biologically significant groups of genes and samples. They have also proven helpful for addressing questions about gene function, gene regulation, and differentiation of GE under various conditions.

7.8. Financial system and economics

Analysis of financial data is becoming more and more important in the business world. Companies anticipate that they gather more data from daily operations, they will be able to draw out knowledge that will be valuable in making judgments about future customer‘s behavior and services. Labeled data of ground truth are relatively rare in many financial applications, such as credit assessment, fraud detection, and reject inference. Unsupervised models are therefore frequently employed to infer the patterns hidden in the data. Cai et al. [315] stated that several data mining techniques have been used by banking and financial institutions to improve their company performance. The purpose of clustering in financial systems is to create a comprehensive method to detect clusters in financial data and to broaden the scope of the clusters so that they are comprehensible [316]. Wenjie et al. [317] proposed Semantic-driven subtractive clustering, a parallel clustering technique that is based on a Hadoop MapReduce framework, was presented for mitigating the risk of customer churn.

7.9. Intelligent robotics

The Multi-Robot Task Allocation (MRTA) problem is a challenging subject in the robotics area with several practical applications. Asma et al. [318] proposed a cluster-based solution to this problem. ACD^{2}PSO is the name of a clustering method that the authors presented that is based on a dynamic distributed PSO approach. The newly developed clustering technique divides the robot tasks into clusters first using dynamic distributed particle swarm optimization (D^{2}PSO), and then allocates the robots to the clusters using the concept of multiple travelling salesman issues. In another research, Arslan et al. [319] developed a unique clustering application to deal with the issue of coordinated robot navigation. A sequence of hierarchy-preserving controllers for a general, hierarchical navigation framework that is provably accurate for collision-free motion design towards any given destination. Janati et al. [320] applied KM clustering algorithm for assigning large number of tasks to robot from multiple tasks in an efficient time. Clustering algorithm reduces the size of state space explored by partition the tasks into disjoint sets. Hence, the computational complexity is reduced by decreasing the size of the state space. After partitioning the tasks, the robots should be assigned to clusters in an optimum way based on the distance from robots to the nearest task of the clusters.

7.10. Text mining

Text mining, also known as text data mining, knowledge discovery in text, or intelligent text analysis, refers to extracting information from unstructured text and presenting the knowledge to users in a clear and concise manner. The main challenge in text mining is the unstructured nature of the data. After transforming textual data to a structured representation, text mining uses many of the same data mining techniques on the corpus of textual data [321]. Text mining consists of three activities: information retrieval, information extraction and data mining [322]. It is frequently used in a wide range of sectors, including journalism and media, telecommuting, power and other service sectors; information technology and the internet; banking services; insurance; and financial markets. It is also often utilized in political institutions, political analysts, government departments, and legal documentation. Many different tasks are included in text mining, including document clustering [323], document classification, text summarization [324,325], sentiment analysis [326,327], social network analysis [328], web page classification, author identification, plagiarism detection, spam/malware/phishing analysis, patent analysis [329,330] and financial decision making etc. With the wide range of applicability, the field of text mining is an active area of research. Nassirtoussi et al. [331] conducted survey on market prediction based on online text mining. The reviews focused on three important areas: pre-processing, ML and evaluation mechanisms.

7.11. Video surveillance

Due to ongoing security concerns, the necessity of monitoring both public and private areas is growing in prominence today. This highlights the necessity for surveillance systems. A surveillance system is useful in object tracking, behaviour analysis, motion detection, face recognition, and classification tasks [332]. Although textual data appears to be rising in volume on the internet and in other automated systems, there are some occasions when it may not be possible to learn as much from them as from video files. A compact video file may include more information than text documents or other media files such as audio and image files. Asad et al. [333] developed an improved version of the HAC algorithm, which can cluster human face images for security purposes. Ranjith et al. [334] proposed a novel approach known as new anomaly detection based on DBSCAN that groups the motions of moving objects of various sizes and shapes. It is used to detect anomalies based on moving vehicle trajectories.

7.12. Marketing and business analytics

Market segmentation has been the main application of cluster analysis in marketing [65]. Market segmentation is the division of a large market into more manageable groupings or clusters [335]. In the business environment, segmentation strategy has been widely used due to the heterogeneity of customer demands. Marketing researchers utilized the KM algorithm to cluster customers in tourism applications [336,337], banking applications [338], telecommunications [339] or customer behaviour relating to their weight loss and beauty preferences [340]. The goal is to cluster consumers and more precisely target each of them [6]. The market segmentation technique provides a lot of advantages. The most obvious benefit is that decision-makers may more accurately target smaller market patterns by adopting specialized marketing strategies and making better use of resources. Moreover, market segmentation strengthens connections between customers and businesses. The automation of customer recommendation systems has been used by marketers to process customer reviews. For market analysis, clustering techniques are now frequently used. These techniques use customer reviews that have been captured and grouped into reviews with similar preferences [4]. Zahra et al. [341] proposed a novel KM algorithm for recommender systems. The proposed algorithm is based on the novel centroid based selection rather than random selection. Piggott [342] created a new market segmentation model based on airlines using data clustering by KM, EM, x-means, hierarchical, and random clustering. Big data analytics and ML algorithms are being studied for their potential application to marketing-related problems

including forecasting, segmentation, and knowledge-based decision support systems in the hotel industry [343]. Fuzzy techniques' membership functions make it possible for a single data point to belong to many groups, and in certain marketing applications, examining group overlap is the key to examining business initiatives [199]. Abdullah Alghamdi [343] developed a novel hybrid approach to hotel selection that incorporates attribute selection, fuzzy clustering, and a neuro-fuzzy system for social data analysis. The method is applied on the dataset obtained from TripAdvisor. The author employs optimal feature selection based on correlation approach, fuzzy KM for data clustering and the adaptive neuro-fuzzy inference system approach for forecasting consumers' overall satisfaction in each cluster.

7.13. Object and character recognition

The essential process of object recognition allows access from vision to other cognitive functions including categorization, language, and reasoning. Character recognition such as handwriting recognition (HR) has been a prominent field of research. One of the well-known early choices for character recognition was neural networks [344]. 3D object recognition is essential for many indoor applications to comprehend the surroundings [345]. Several clustering techniques and strategies have been developed to aggregate views of 3D objects for object recognition in range data. Using clustering techniques, certain studies—including [1,4,31] recognized lexemes in handwritten text for writer identification HR. Gaur et al. [346] proposed method for recognition of handwritten Hindi characters using KM clustering and SVM, which organizes images into k-clusters. Similar type of study proposed by Sheshadri et al. [347] to recognize Kannada characters using KM clustering. Pourmohammad et al. [344] proposed an efficient way of character recognition using KM algorithm.

7.14. Patter recognition in data mining

Organizations may now collect huge amounts of information because to rapid advancements in data collecting and storage technologies. However, it has proven to be quite difficult to extract valuable information. Data mining is a collection of mathematical and statistical techniques that automatically extract new patterns, useful information, and knowledge from massive databases [348]. Clustering is a well-known fundamental data mining task for information extraction. Nonetheless, a number of researchers have created and provided several clustering methods with varying applicability for various domains. It is useful to a broad range of business intelligence applications such as outlier/fraud detection, customer profiling, targeted marketing, work flow management, and store layout. Benabdellah et al. [349] examined and compared the suitability of AHC, KM, SOM, and DBSCAN to the sparse datasets originating from industries.

7.15. Data transfer through network

Several network management activities, including flow prioritization, traffic shaping/policing, and diagnostic monitoring, rely on the correct detection and categorization of network traffic based on application type. Social media websites and forums are being used to distribute vast amounts of user-generated data online. To achieve high-speed data transfer and avoid delays, an efficient transmission system is required to categorize traffic by taking advantage of the unique traits that applications have when they communicate on a network. Clustering is frequently used to effectively find similar groupings of traffic to overcome these problems. UML algorithms, for example, the autoclass algorithm, DBSCAN, and KM, have been utilized to group homogeneous traffic detection utilizing transport layer data based on the application's distinguishing qualities as they transmit across a given network [350].

7.16. Urban development

The 21st century's most important demographic trend is urbanization. Urbanization is expected to increase from 55 % of the population in 2018 to 68 % by 2050 [351]. Smart technology is being quickly embraced by contemporary society. IoT applications, particularly in smart cities, are quickly evolving. Data is regularly produced by smart city technology. IoT data that is gathered on the cloud is typically diverse and unstructured [352]. The challenge of service location is critical in many contexts, including the positioning of sensors in a network, the location of warehouses, the location of public parks and the placement of brand-specific retail outlets. UL has a successful track record of deciphering the complexity of cities. An ant clustering algorithm-based KM method is used in macroscopic of highway transportation hubs [353]. Moreover, generalized density-based clustering algorithm called GFDBSCAN can be used to group geographical areas according to the needs and preferences that customers have stated for services such as retail establishments, ATMs, bank branch operations, public utilities etc. [354]. Logesh et al. [21] proposed a novel urban trip recommendation using hybrid quantum-induced PSO clustering in smart cities. The satisfaction rate, TP rate, recall, f-measure, and accuracy of the suggested recommendation approach have been assessed using real-world, large-scale datasets of Yelp and TripAdvisor. Ran et al. [355] proposed an innovative noise-enabled KM clustering algorithm and applied it to locate urban hotspots. The suggested technique performs better at clustering, reliably produces clustering results, and efficiently captures urban hotspots.

7.17. Privacy protection

Nowadays, more and more data is being sent through the Internet. As a result, data security is seen as a serious challenge when transferring data via the Internet [356]. The most important issues that need to be taken into account in the current digital world are privacy concerns against unauthorized access to information. Clustering methods that divide data into groups and examine those groupings can assist improve data security and reliability [4]. A network traffic monitoring system called IDS may detect malicious activity and issues alert. Network administrators can use IDS to identify typical and unusual activity in network traffic packets. As compared to the overall detection rate of attacks, the existing IDS solutions often have low detection accuracy for particular attack types. To overcome the limitations of existing methods, Alfoudi et al. [357] proposed novel hyper-clustering model for dynamic intrusion detection based on DBSCAN and cosine similarity. Chiou et al. [356] proposed a novel spectral clustering method to analyse network traffic and classify clients as victims or not. Sheng et al. [358] proposed a heuristic clustering technique to classify unknown traffic in supervisory control and data acquisition networks. Data security is the study of protecting digital information from unauthorized access, alteration or theft throughout its lifespan.

7.18. Human activity recognition using wearable sensors

There has been a recent increase in research into UML algorithms for human activity recognition (HAR) [359]. The task of HAR is important in many applications, including ubiquitous healthcare, smart environments, security and surveillance. The most useful applications of HAR have been in the fall detection, behavioral and psychological monitoring, stress detection, gait abnormality detection [360] and other applications. HAR can be generically categorized into two categories: sensor-based and vision-based [361]. It is essential to gather accurate application data in order to provide more advanced services. The challenge of labeling datasets during or after collection, which is typically tedious, time-consuming, and expensive, is one limitation of wearable sensors-based HAR [362]. UML approaches that process information from wearable sensors and/or cameras strategically placed in the

surroundings are typically used to recognize human activities. Data clustering is the most critical aspect of UML technique. Clustering is helpful to automatic recognition of human activity after analyzing these unlabeled data. Kwon et al. [359] presented UML techniques for recognizing everyday human activities such as standing, sitting, walking, jogging, and lying down on mobile devices. The authors formed groups of k = 5 activities using KM clustering, a mixture of Gaussian (GMM), average-linkage hierarchical agglomerative clustering, and DBSCAN, and found that GMM showed the maximum accuracy [359]. Ma. et al. [363] proposed a multi-task deep clustering framework for HAR. The unlabeled dataset is divided into groups using the KM clustering algorithm, which generates pseudo labels for the instances. A DNN classifier is trained for the classification of human activities using the latent features and pseudo labels. The HAR framework, which includes sensor-based data gathering, data preparation, and UML approaches for recognizing human actions, has emerged as an active topic in ubiquitous computing research.

7.19. Other applications

Clustering has several applications. Here are some other useful applications found in various fields such as network anomaly detection [222], operation research [364], customer churn prediction [223], IDS [225], computer graphics, social media analysis, image clustering [365], social computing [366] and astronomy [44] etc. The distribution of papers in clustering applications is shown in Fig. 13.

8. Concluding remarks

Data clustering, a prominent data mining approach, aims to uncover inherent patterns within datasets by grouping similar data points into clusters. Recent years have witnessed significant advancements in this field, leading to the emergence of innovative clustering methodologies. This study contributes valuable insights into data clustering by systematically reviewing 367 articles published between 1999 and 2024. It comprehensively covers a diverse array of clustering methods and their taxonomies, recent advancements in the field, metrics for measuring similarity and validating clustering results, challenges encountered in cluster analysis, and varied applications of clustering algorithms. By consolidating this information, the study serves as a pivotal resource for researchers, analysts, and data scientists seeking to develop novel, adaptable, and efficient clustering techniques.

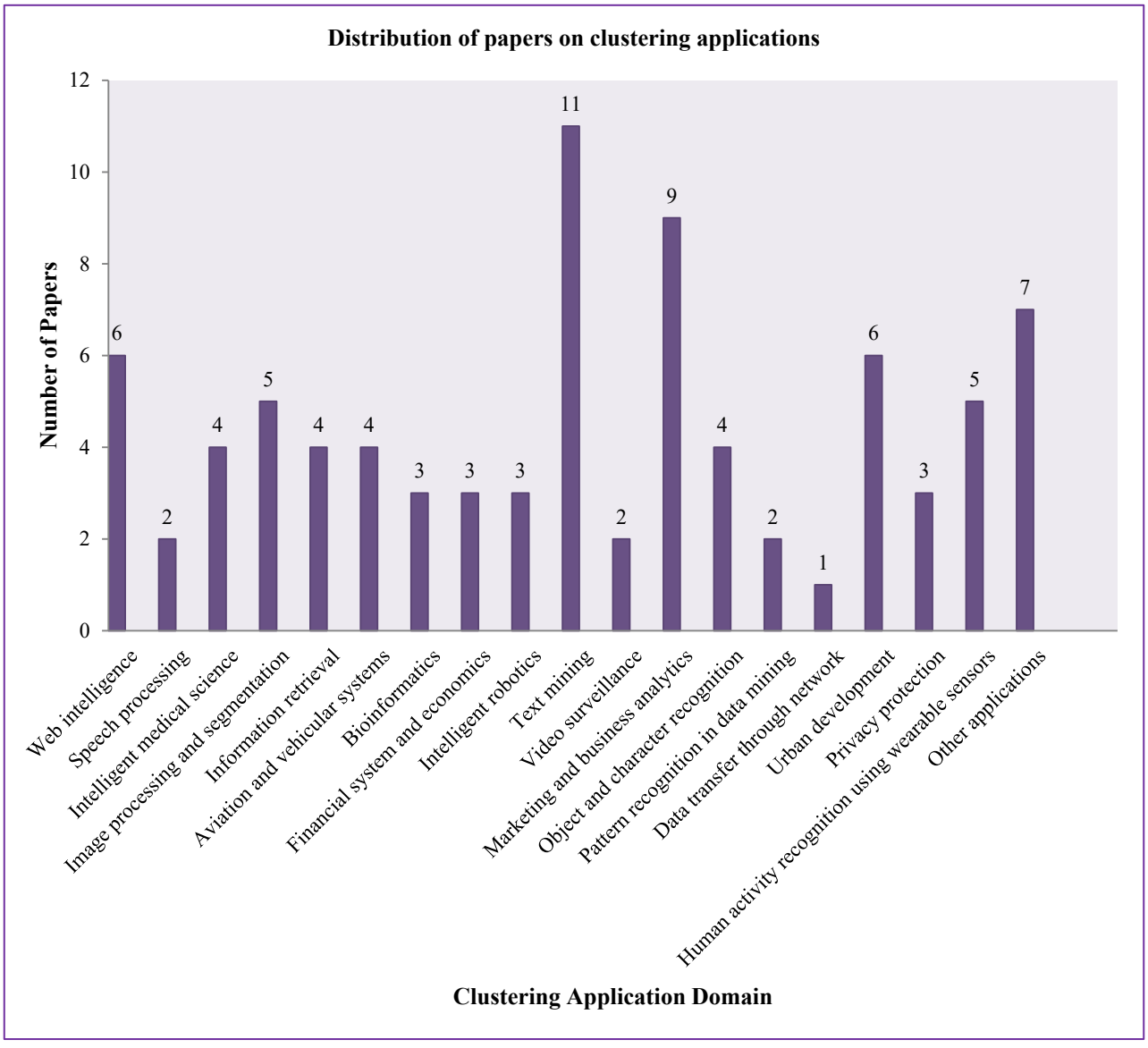


Fig. 13. Papers on the distribution of clustering applications in different domains.

According to the studies presented in this review paper, numerous clustering algorithms have been developed recently across various domains. However, none of these algorithms provides a general solution to common clustering challenges. Both hierarchical and partitioning-based methods face limitations, particularly in terms of computational speed and throughput, which pose challenges when applied to big data contexts. In order to address these limitations, researchers can further expand their research work in the field of data clustering, particularly in leveraging parallel computing concepts and emerging metaheuristic-based clustering methods inspired by nature. Recent advancements in clustering techniques such as multiview, mixed data type, automatic, and time series clustering are gaining attention but require more thorough investigation. Additionally, new strategies need to be devised to enhance the generation and distribution of quality solutions, expedite convergence, and improve both exploitation and generalization capabilities. These advancements are essential for making data clustering more effective in addressing complex real-world problems.

CRediT authorship contribution statement

Jaswinder Singh: Writing – review & editing, Writing – original draft. **Damanpreet Singh:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. Prakash, A review of clustering techniques and developments, Neurocomputing 267 (2017) 664–681.

[2] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, Comprehensive survey on hierarchical clustering algorithms and the recent developments, no. 222. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10366-3.

[3] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recogn. Lett. 31 (8) (2010) 651–666, https://doi.org/10.1016/j.patrec.2009.09.011.

[4] A.E. Ezugwu, A.M. Ikotun, O.O. Oyelade, L. Abualigah, A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects, Eng. Appl. Artif. Intel. 110 (2022) 104743, https://doi.org/10.1016/j.engappai.2022.104743.

[5] M.E. Celebi, H.A. Kingravi, P.A. Vela, A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Syst. Appl. 40 (1) (2013) 200–210, https://doi.org/10.1016/j.eswa.2012.07.021.

[6] S. Subudhi, S. Panigrahi, Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection, J. King Saud Univ. - Comput Inf. Sci. 32 (5) (2020) 568–575, https://doi.org/10.1016/j.jksuci.2017.09.010.

[7] A.E.E. Abiodun M. Ikotun, Enhanced firefly-K-means clustering with adaptive mutation and central limit theorem for automatic clustering of high-dimensional datasets, Appl. Sci. (2022).

[8] Y. Duan, C. Liu, S. Li, X. Guo, C. Yang, An automatic affinity propagation clustering based on improved equilibrium optimizer and t-SNE for high-dimensional data, Inf. Sci. (ny) 623 (2023) 434–454, https://doi.org/10.1016/j.ins.2022.12.057.

[9] V.K. Dehariya, S.K. Shrivastava, R.C. Jain, Clustering of image data set using K-means and fuzzy K-means algorithms, in: 2010 Int. Conf. Comput. Intell. Commun. networks, pp. 386–391, 2010, doi: 10.1109/CICN.2010.80.

[10] A.M. Bagirov, R.M. Aliguliyev, N. Sultanova, Finding compact and well-separated clusters: clustering using silhouette coefficients, Pattern Recogn. 135 (2023), https://doi.org/10.1016/j.patcog.2022.109144.

[11] A.E. Ezugwu, Nature - inspired metaheuristic techniques for automatic clustering: a survey and performance study, Springer International Publishing (2020), https://doi.org/10.1007/s42452-020-2073-0.

[12] A. Fahad, et al., A survey of clustering algorithms for big data: taxonomy and empirical analysis, IEEE Trans. Emerg. Top. Comput. 2 (3) (2014) 267–279, https://doi.org/10.1109/TETC.2014.2330519.

[13] M.A. Mahdi, K.M. Hosny, I. Elhenawy, Scalable clustering algorithms for big data : a review 9 (2021) 80015–80027, doi: 10.1109/ACCESS.2021.3084057.

[14] M. Mittal, L. M. Goyal, D. Jude, H. Jasleen, Clustering approaches for high-dimensional databases : a review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., no. December 2018, pp. 1–14, 2019, doi: 10.1002/widm.1300.

[15] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: a new perspective, Neurocomputing 300 (2018) 70–79, https://doi.org/10.1016/j.neucom.2017.11.077.

[16] Q. Li, S. Wang, X. Zeng, B. Zhao, Y. Dang, How to improve the accuracy of clustering algorithms, Inf. Sci. (Ny)., vol. 627, no. June 2022, pp. 52–70, 2023, doi: 10.1016/j.ins.2023.01.094.

[17] A. José-García, W. Gómez-Flores, Automatic clustering using nature-inspired metaheuristics: a survey, Appl. Soft Comput. J. 41 (2016) 192–213, https://doi.org/10.1016/j.asoc.2015.12.001.

[18] Y. Liu, X. Wu, Y. Shen, Automatic clustering using genetic algorithms, Appl. Math Comput. 218 (4) (2011) 1267–1279, https://doi.org/10.1016/j.amc.2011.06.007.

[19] S.M. Javidan, A. Banakar, K.A. Vakilian, Y. Ampatzidis, Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning, Smart Agric. Technol. 3(June 2022) 100081, 2023, doi: 10.1016/j.atech.2022.100081.

[20] A. Rahman, Z. Islam, Knowledge-based systems a hybrid clustering technique combining a novel genetic algorithm with, Knowledge-Based Syst. 71 (2014) 345–365, https://doi.org/10.1016/j.knosys.2014.08.011.

[21] R. Logesh, V. Subramaniyaswamy, V. Vijayakumar, X. Gao, V. Indragandhi, A hybrid quantum-induced swarm intelligence clustering for the urban trip recommendation in smart city, Futur. Gener. Comput. Syst. 83 (2018) 653–673, https://doi.org/10.1016/j.future.2017.08.060.

[22] D. Xu, Y. Tian, A comprehensive survey of clustering algorithms, Ann. Data Sci. 2 (2) (2015) 165–193, https://doi.org/10.1007/s40745-015-0040-1.

[23] R. Xu, S. Member, D.W. Ii, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.

[24] B.F. Azevedo, A. Maria, A.C.R. Ana, Hybrid approaches to optimization and machine learning methods : a systematic literature review 113(7). Springer US, 2024. doi: 10.1007/s10994-023-06467-x.

[25] A.M. Ikotun, A.E. Ezugwu, L. Abualigah, B. Abuhaija, J. Heming, K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data, Inf. Sci. (NY) 622 (2023) 178–210, https://doi.org/10.1016/j.ins.2022.11.139.

[26] P. Bhattacharjee, Panthadeep, Mitra, A survey of density based clustering algorithms, Front. Comput. Sci. 15(1) (2021), doi: https://doi.org/10.1007/s11704-019-9059-3.

[27] E. Hancer, B. Xue, M. Zhang, A survey on feature selection approaches for clustering, Artif. Intell. Rev. 53 (6) (2020) 4519–4545, https://doi.org/10.1007/s10462-019-09800-w.

[28] Y. Yang, H. Wang, Multi-view clustering: a survey, Big Data Min. Anal. 1 (2) (2018) 83–107, https://doi.org/10.26599/BDMA.2018.9020003.

[29] S. Bandaru, A.H.C. Ng, K. Deb, Data mining methods for knowledge discovery in multi-objective optimization: Part A - Survey, Expert Syst. Appl. 70 (2017) 139–159, https://doi.org/10.1016/j.eswa.2016.10.015.

[30] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) (2006) 2399–2434.

[31] A.K. Jain, Data clustering: a review, Adv. Mach. Learn. Data Min. Astron. 31 (3) (1999) 543–561, https://doi.org/10.1201/b11822-19.

[32] C.C. Aggarwal, C. Zhai, A survey of text clustering algorithm, Min. Text Data (2012) 77–128, https://doi.org/10.1007/978-1-4614-3223-4.

[33] E.R. Hruschka, R.J.G.B. Campello, A.A. Freitas, C. Ponce, L.F. De Carvalho, A survey of evolutionary algorithms for clustering, IEEE Trans. Syst. Man Cybern. 39 (2) (2009) 133–155, https://doi.org/10.1109/TSMCC.2008.2007252.

[34] T.W. Liao, Clustering of time series data — a survey, Pattern Recogn. 38 (2005) 1857–1874, https://doi.org/10.1016/j.patcog.2005.01.025.

[35] S. Jagannath, G. Panda, A survey on nature inspired metaheuristic algorithms for partitional clustering, Swarm Evol. Comput. 16 (2014) 1–18, https://doi.org/10.1016/j.swevo.2013.11.003.

[36] K. Bindra, A. Mishra, A detailed study of clustering algorithms, 6th Int. Conf. Reliab. infocom Technol. Optim., 2017, pp. 371–376.

[37] I. Assent, Clustering high dimensional data, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2(August) (2012) 340–350, doi: 10.1002/widm.1062.

[38] A. Alam, M. Muqeem, S. Ahmad, Comprehensive review on clustering techniques and its application on high dimensional data, Int. J. Comput. Sci. Netw. Secur. (2021) 237–244, https://doi.org/10.22937/IJCSNS.2021.21.6.31.

[39] D. Lam, D.C. Wunsch, Clustering 20, 2014. doi: 10.1016/B978-0-12-396502-8.00020-6.

[40] C.D. Nguyen, K.J. Cios, GAKREM: A novel hybrid clustering algorithm, Inf. Sci. (Ny) 178 (2008) 4205–4227, https://doi.org/10.1016/j.ins.2008.07.016.

[41] T. Barton, T. Bruna, P. Kordik, Chameleon 2: an improved graph-based clustering algorithm, ACM Trans. Knowl. Discov. from Data 13 (1) (2019) 1–27, https://doi.org/10.1145/3299876.

[42] A. Agarwal, R.K. Roul, A novel hierarchical clustering algorithm for online resources, vol. 708. Springer Singapore, 2018. doi: 10.1007/978-981-10-8636-6_49.

[43] D.P. Dabhi, M.R. Patel, M.R.P. Dipak, P. Dabhi, Extensive survey on hierarchical clustering methods in data mining, Int. Res. J. Eng. Technol. 03 (11) (2016) 659–665.

[44] H. Yu, X. Hou, Hierarchical clustering in astronomy, Astron. Comput. 41 (2022) 100662, https://doi.org/10.1016/j.ascom.2022.100662.

[45] Y. Jeon, J. Yoo, J. Lee, S. Yoon, S. Member, NC-link: A new linkage method for efficient hierarchical clustering of large-scale data, IEEE Access 5 (2017) 5594–5608, https://doi.org/10.1109/ACCESS.2017.2690987.
[46] A.M. Jarman, Hierarchical cluster analysis: comparison of single linkage, complete linkage, average linkage and centroid linkage method, Res. Gate (2020) 1–13, https://doi.org/10.13140/RG.2.2.11388.90240.
[47] M. Roux, A comparative study of divisive and agglomerative hierarchical clustering algorithms, J. Classif. 35 (August) (2018) 345–366, https://doi.org/10.1007/s00357-018-9259-9.
[48] F. Murtagh, Ward's hierarchical agglomerative clustering method: which algorithms implement ward' s criterion ? J. Classif. 295 (October) (2014) 274–295, https://doi.org/10.1007/s00357-014-9161-z.
[49] J. Brier, lia dwi jayanti, "SLINK: An optimally efficient algorithm for the single-link cluster method, Comput. J., vol. 21, no. 1, pp. 30–34, 1973, [Online]. Available: http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203.
[50] R.T. Ng, J. Han, I.C. Society, CLARANS : A Method for Clustering Objects for Spatial Data Mining, IEEE Trans. Knowl. Data Eng. 14(5) (2002) 1003–1016, doi: https://doi.org/10.1109/TKDE.2002.1033770.
[51] T. Sun, C. Shu, F. Li, H. Yu, L. Ma, Y. Fang, An efficient hierarchical clustering method for large datasets with map-reduce, 2009 Int Conf. Parallel Distrib. Comput. Appl. Technol. (2009) 494–499, https://doi.org/10.1109/PDCAT.2009.46.
[52] Sudipto Guha, C.F. StanfordTsai, Z.C. Chen, C.W. Tsai, CURE: An efficient clustering algorithm for large databases, in: Proc. IEEE Int. Conf. Syst. Man Cybern., vol. 5, pp. 446–451, 2002, doi: 10.1109/ICSMC.2002.1176400.
[53] P.A. Vijaya, M.N. Murty, D.K. Subramanian, Leaders – subleaders: an efficient hierarchical clustering algorithm for large data sets, Pattern Recogn. Lett. 25 (2004) 505–513, https://doi.org/10.1016/j.patrec.2003.12.013.
[54] S. Guha, R. Rastogi, K. Shim, Rock: a robust clustering algorithm for categorical attributes, Inf. Syst. 25 (5) (2000) 345–366, https://doi.org/10.1016/S0306-4379(00)00022-3.
[55] T. Xiong, S. Wang, A. Mayers, DHCC: divisive hierarchical clustering of categorical data, Data Min. Knowl. Discov. (2012) 103–135, https://doi.org/10.1007/s10618-011-0221-2.
[56] G. Karypis, E. Han, V. Kumar, Chameleon : Hierarchical Clustering Using Dynamic Modeling, Computer (Long. Beach. Calif), 1999, pp. 68–75, doi: https://doi.org/10.1109/2.781637.
[57] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large database, ACM SIGMOD Rec. 25 (2) (1996) 103–114, https://doi.org/10.1145/235968.233324.
[58] S. Horng, M. Su, Y. Chen, T. Kao, R. Chen, J. Lai, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert Syst. Appl. 38 (1) (2011) 306–313, https://doi.org/10.1016/j.eswa.2010.06.066.
[59] P. Pappula, U.N. Dulhare, A study on monothetic Divisive Hierarchical Clustering Method, Int. J. Adv. Sci. Technol. Eng. Manag. Sci., no. August, 2017.
[60] M. Chavent, Y. Lechevallier, O. Briant, DIVCLUS-T: A monothetic divisive hierarchical clustering method, Comput. Stat. Data Anal. 52 (2007) 687–701, https://doi.org/10.1016/j.csda.2007.03.013.
[61] C. Zhong, D. Miao, R. Wang, X. Zhou, DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points, Pattern Recogn. Lett. 29 (16) (2008) 2067–2077, https://doi.org/10.1016/j.patrec.2008.07.002.
[62] O. Pasi Franti, Virmajoki, V. Hautama, Fast agglomerative clustering using a k -nearest neighbor graph, in: IEEE Trans. Pattern Anal. Mach. Intell. 28(11) (2006) 1875–1881, doi: https://doi.org/10.1109/ICRA.2014.6907776.
[63] D.T. Dinh, V.N. Huynh, S. Sriboonchitta, Clustering mixed numerical and categorical data with missing values, Inf. Sci. (Ny) 571 (2021) 418–442, https://doi.org/10.1016/j.ins.2021.04.076.
[64] W. Wei, J. Liang, X. Guo, P. Song, Y. Sun, Hierarchical division clustering framework for categorical data, Neurocomputing 341 (2019) 118–134, https://doi.org/10.1016/j.neucom.2019.02.043.
[65] S.K. Popat, M. Emmanuel, Review and comparative study of clustering techniques, Int. J. Comput. Sci. Inf. Technol. 5 (1) (2014) 805–812.
[66] Y. Xiao, J. Yu, Partitive clustering (K-means family), Wiley Interdiscip Rev. Data Min. Knowl. Discov. 2 (June) (2012) 209–225, https://doi.org/10.1002/widm.1049.
[67] C.B.N. Cir, G. Cleuziou, N. Essoussi, Overview of overlapping partitional clustering methods, Partitional Clust. Algorithms (2015) 245–275, https://doi.org/10.1007/978-3-319-09259-1.
[68] A. Ahmad, A k -mean clustering algorithm for mixed numeric and categorical data, J. Syst. Sci. Complex. 63 (2007) 503–527, https://doi.org/10.1016/j.datak.2007.03.016.
[69] D.J. Bora, A comparative study between fuzzy clustering algorithm and hard clustering algorithm, arXiv Prepr. arXiv, vol. 10, no. 2, pp. 108–113, 2014, doi: https://doi.org/10.48550/arXiv.1404.6059.
[70] A. Taher, S.A. El-said, A. Ella, Fuzzy and hard clustering analysis for thyroid disease, Comput. Methods Programs Biomed. 111 (1) (2013) 1–16, https://doi.org/10.1016/j.cmpb.2013.01.002.
[71] Y. Chen, S. Sanghavi, H. Xu, Improved graph clustering, IEEE Trans. Inf. Theory 60 (10) (2014) 6440–6455, https://doi.org/10.1109/TIT.2014.2346205.
[72] P. Foggia, et al., A graph-based clustering method and its applications, Adv. Brain, Vision, Artif. Intell. Second Int. Symp. BVAI 2 (2007) 277–287.
[73] C. Science, An enhanced density based spatial clustering of applications with noise, 2009 IEEE Int Adv. Comput. Conf. (2009) 6–7, https://doi.org/10.1109/IADCC.2009.4809235.
[74] S. Kamran Khan, Fong, S.U. Rehman, K. Aziz, I. Science, DBSCAN : Past, Present and Future," fifth Int. Conf. Appl. Digit. Inf. web Technol. (ICADIWT 2014), pp. 232–238, 2014, doi: https://doi.org/10.1109/ICADIWT.2014.6814687.
[75] M. Hahsler, M. Piekenbrock, D. Doran, "dbscan : Fast Density-Based Clustering with R, J. Stat. Softw., vol. 91, no. 1, 2019, doi: 10.18637/jss.v091.i01.
[76] R. Maheshwari, S. Kumar, A. Chandra, DCSNE: density-based clustering using graph shared neighbors and entropy, Pattern Recogn. 137 (2023) 109341, https://doi.org/10.1016/j.patcog.2023.109341.
[77] R.J.G.B. Campello, P. Kröger, J. Sander, A. Zimek, Density-based clustering, Data Min. Knowl. Discov., no. August, pp. 1–15, 2019, doi: 10.1002/widm.1343.
[78] A. Idrissi, A multi-criteria decision method in the DBSCAN algorithm for better clustering, Int. J. Adv. Comput. Sci. Appl. 7 (2) (2016) 377–384.
[79] B. Borah, D.K. Bhattacharyya, An improved sampling-based DBSCAN for large spatial databases, Int. Conf. Intell. Sens. Inf. Process., pp. 92–96, 2004, doi: https://doi.org/10.1109/ICISIP.2004.1287631.
[80] H. Rehioui, A. Idrissi, M. Abourezq, F. Zegrari, DENCLUE-IM : A New Approach for Big Data Clustering, Procedia - Procedia Comput. Sci., vol. 83, no. Ant 2016, pp. 560–567, 2022, doi: 10.1016/j.procs.2016.04.265.
[81] M. Ankerst, M.M. Breunig, H. Kriegel, OPTICS: Ordering points to identify the clustering structure, ACM SIGMOD Rec. (1999) 49–60, https://doi.org/10.1145/304181.304187.
[82] P. Liu, D. Zhou, N. Wu, VDBSCAN: Varied density based spatial clustering of applications with noise, 2007 Int Conf. Serv. Syst. Serv. Manag. (2007) 1–4, https://doi.org/10.1016/j.patrec.2008.07.002.
[83] B. Liu, A fast density-based clustering algorithm for large databases, Int. Conf. Mach. Learn. Cybern., no. August, pp. 996–1000, 2006, doi: https://doi.org/10.1109/ICMLC.2006.258531.
[84] O. Uncu, W.A. Gruver, D.B. Kotak, S. Memmber, GRIDBSCAN : GRId density-based spatial clustering of applications with noise, 2006 IEEE Int. Conf. Syst. Man Cybern. (2006) 2976–2981, https://doi.org/10.1109/ICSMC.2006.384571.
[85] A. Degirmenci, O. Karal, Efficient density and cluster based incremental outlier detection in data streams, Inf. Sci. (Ny) 607 (2022) 901–920, https://doi.org/10.1016/j.ins.2022.06.013.
[86] X. Wei, An overview on density peaks clustering, Neurocomputing (2023) 1–34, https://doi.org/10.21203/rs.3.rs-2428649/v1.
[87] C. Bouveyron, C. Brunet-saumard, Model-based clustering of high-dimensional data: a review, Comput. Stat. Data Anal. 71 (2014) 52–78, https://doi.org/10.1016/j.csda.2012.12.008.
[88] E.R. C. FRALEY, How many clusters ? Which clustering method ? Answers via model-based cluster analysis, Comput. J. 41(8) (1998) 578–588, doi: https://doi.org/10.1093/comjnl/41.8.578.
[89] U. Kokate, A. Deshpande, P. Mahalle, P. Patil, Data stream clustering techniques, applications, and models: comparative analysis and discussion, Big Data Cogn. Comput. 2 (2018), https://doi.org/10.3390/bdcc2040032.
[90] P.D. Mcnicholas, Model-based clustering, J. Classif. 373 (November) (2016) 331–373, https://doi.org/10.1007/s0035.
[91] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Mach. Learn., pp. 139–172, 1987.
[92] A.E.R. Chris Fraley, MCLUST: software for model-based cluster analysis, Icml (1999) 297–306.
[93] M. Yang, C. Lai, C. Lin, A robust EM clustering algorithm for Gaussian mixture models, Pattern Recogn. 45 (11) (2012) 3950–3961, https://doi.org/10.1016/j.patcog.2012.04.031.
[94] G.A. Carpenter, S. Grossberg, A massively parallel architecture for a self-organizing neural pattern recognition machine, Comput. Vision, Graph. Image Process. 115 (1987) 54–115, https://doi.org/10.1016/S0734-189X(87)80014-2.
[95] T. Kohonen, The self-organizing map, Proc. IEEE 78 (9) (1990) 1464–1480, https://doi.org/10.1109/5.58325.
[96] S. Aghabozorgi, A. Seyed, T.Y. Wah, Time-series clustering – a decade review, Inf. Syst. 53 (2015) 16–38, https://doi.org/10.1016/j.is.2015.04.007.
[97] E.E.A.E.J.C. SOLTANOLKOTABI, MAHDI and Stanford, Robust subspace clustering, Inst. Math. Stat. 42(2) (2014) 669–699, doi: 10.1214/13-AOS1199.
[98] H. Kriegel, P. Kr, A. Zimek, Subspace clustering, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2(August) 351–364, 2012, doi: 10.1002/widm.1057.
[99] H. Rakesh Agrawal, Road, S. Jose, Automatic subspace clustering of high dimensional data for data mining applications, Proc. 1998 ACM SIGMOD Int. Conf. Manag. data, pp. 94–105, 1998, doi: https://doi.org/10.1145/276304.276314.
[100] L. Parsons, Subspace clustering for high dimensional data: a review, Acm Sigkdd Explor. Newsl. 6 (1) (2004) 90–105, https://doi.org/10.1145/1007730.1007731.
[101] X. Peng, J. Feng, J.T. Zhou, Y. Lei, S. Yan, Deep subspace clustering, IEEE Trans. Neural Networks Learn. Syst. 31 (12) (2020) 5509–5521, https://doi.org/10.1109/TNNLS.2020.2968848.
[102] P.R. Rakesh Agrawal, J. Gehrke, D. Gunopulos, Automatic subspace clustering of high dimensional data, Data Min. Knowl. Discov. (2005) 5–33.
[103] D. Karaboga, C. Ozturk, A novel clustering approach: artificial bee colony (ABC) algorithm, Appl. Soft Comput. 11 (2011) 652–657, https://doi.org/10.1016/j.asoc.2009.12.025.
[104] J. Peng, A cutting algorithm for the minimum sum-of-squared error clustering, in: Proc. 2005 SIAM Int. Conf. Data Min., pp. 150–160, 2005, doi: https://doi.org/10.1137/1.9781611972757.14.
[105] K. Hammouda, A comparative study of data clustering techniques, Univ. Waterloo, Ontario, Canada, 2000, pp. 1–21.
[106] T.M.K. Trupti, M. Kodinariya, D.P.R. Makwana, Review on determining number of cluster in K-means clustering, Int. J. 2013 (2016) 90–95.

[107] T.M. Kodinariya, P.R. Makwana, Review on determining of cluster in K-means [Online].Available: Int. J. Adv. Res. Comput. Sci. Manag. Stud. 1 (6) (2013) 90–95 https://www.researchgate.net/publication/313554124.
[108] X. Liu, W. Gla, J.A.K. Suykens, S. Member, B. De Moor, Y. Moreau, Optimized data fusion for kernel k-means clustering, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 1031–1039, https://doi.org/10.1109/TPAMI.2011.255.
[109] K.A.A. Nazeer, S.D.M. Kumar, Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids K, in: 2011 Second Int. Conf. Emerg. Appl. Inf. Technol., 2011, pp. 38–41, doi: 10.1109/EAIT.2011.57.
[110] D. Aloise, A. Deshpande, P. Hansen, NP-hardness of Euclidean sum-of-squares clustering, 2009, pp. 245–248, doi: 10.1007/s10994-009-5103-0.
[111] A. Pérez-Ortega, J., Almanza-Ortega, N. N., Vega-Villalobos, A., Pazos-Rangel, R., Zavala-Díaz, C., Martínez-Rebollar, The K-means algorithm evolution. Introduction to Data Science and Machine Learning, 2019.
[112] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, A local search approximation algorithm for k-means clustering, Comput. Geom. Theory Appl. 28 (2004) 89–112, https://doi.org/10.1016/j.comgeo.2004.03.003.
[113] T.P. Karaikal, Selection of optimal number of clusters and centroids for K-means and Fuzzy C-means Clustering : A Review," 2020 5th Int. Conf. Comput. Commun. Secur., 2020, pp. 5–8, doi: https://doi.org/10.1109/ICCCS49678.2020.9276978.
[114] A.E. Ezugwu, M.B. Agbaje, A comparative performance study of hybrid firefly algorithms for automatic data clustering, IEEE Access 8 (2020) 121089–121118, https://doi.org/10.1109/ACCESS.2020.3006173.
[115] X. Wu, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. (2008) 1–37, https://doi.org/10.1007/s10115-007-0114-2.
[116] H. Ismkhan, I-k-means −+ : An iterative clustering algorithm based on an enhanced version of the k -means, Pattern Recogn. 79 (2018) 402–413, https://doi.org/10.1016/j.patcog.2018.02.015.
[117] W. Tong, S. Liu, X. Gao, Neurocomputing A density-peak-based clustering algorithm of automatically determining the number of clusters, Neurocomputing (2020), https://doi.org/10.1016/j.neucom.2020.03.125.
[118] M. Capó, A. Pérez, J.A. Lozano, An efficient approximation to the K-means clustering for massive data, Knowledge-Based Syst. 117 (2017) 56–69, https://doi.org/10.1016/j.knosys.2016.06.031.
[119] C. Zhang, D. Ouyang, J. Ning, An artificial bee colony approach for clustering, Expert Syst. Appl. 37 (7) (2010) 4761–4767, https://doi.org/10.1016/j.eswa.2009.11.003.
[120] S.J. Redmond, C. Heneghan, A method for initialising the K-means clustering algorithm using kd-trees, Pattern Recogn. Lett. 28 (2007) 965–973, https://doi.org/10.1016/j.patrec.2007.01.001.
[121] T. Md Shamsur Rahim, Ahmed, An initial centroid selection method based on radial and angular coordinates for K-means algorithm, in: 2017 20th Int. Conf. Comput. Inf. Technol., pp. 22–24, 2017, doi: https://doi.org/10.1109/ICCITECHN.2017.8281801.
[122] A.E. Ezugwu, A.K. Shukla, M.B. Agbaje, O.N. Oyelade, Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature, Neural Comput. Appl. 4 (2021) 6247–6306, https://doi.org/10.1007/s00521-020-05395-4.
[123] J. Saha, J. Mukherjee, CNAK: Cluster number assisted K-means, Pattern Recogn. 110 (2021) 107625, https://doi.org/10.1016/j.patcog.2020.107625.
[124] K. P. Sinaga, M. Yang, Unsupervised K-means clustering algorithm 8 (2020), doi: 10.1109/ACCESS.2020.2988796.
[125] H.T. Dashti, T. Simas, R.A. Ribeiro, A. Assadi, A. Moitinho, MK-means - Modified K-means clustering algorithm, in: 2010 Int. Jt. Conf. Neural Networks (IJCNN), pp. 1–6, 2010, doi: https://doi.org/10.1109/IJCNN.2010.5596300.
[126] A.M. Dan Pellege, X-means: extending K-means with efficient estimation of the number of clusters, Icml (2000) 727–734.
[127] H. Harb, A. Makhoul, R. Couturier, An enhanced K-means and ANOVA-based clustering approach for similarity aggregation in underwater wireless sensor networks, IEEE Sens. J. 15 (10) (2015) 5483–5493, https://doi.org/10.1109/JSEN.2015.2443380.
[128] A.M. Ikotun, M.S. Almutari, A.E. Ezugwu, K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions, Appl. Sci. 11 (23) (2021) pp, https://doi.org/10.3390/app112311246.
[129] G. Komarasamy, An optimized K-means clustering technique using bat algorithm, vol. 84(2) (2012) 263–273.
[130] S. Ye, X. Huang, Y. Teng, Y. Li, K-means clustering algorithm based on improved cuckoo search algorithm and its application, 2018 IEEE 3rd Int Conf. Big Data Anal. 1 (2018) 422–426.
[131] E.A. Pambudi, A.Y. Badharudin, A.P. Wicaksono, Enhanced K-means by using grey wolf optimizer for brain MRI segmentation, ICTACT J. Soft Comput. 11 (03) (2021) 2353–2358, https://doi.org/10.21917/ijsc.2021.0336.
[132] B. Niu, Q. Duan, J. Liu, L. Tan, Y. Liu, A population-based clustering technique using particle swarm optimization and k-means, Nat. Comput. 16 (1) (2017) 45–59, https://doi.org/10.1007/s11047-016-9542-9.
[133] S.Z. Selim, M.A. Ismail, K-means-type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1984) 81–87, https://doi.org/10.1109/TPAMI.1984.4767478.
[134] L.M. Abualigah, A.T. Khader, E.S. Hanandeh, Hybrid clustering analysis using improved krill herd algorithm, Appl. Intell., pp. 4047–4071, 2018, doi: https://doi.org/10.1007/s10489-018-1190-6 Hybrid.
[135] X. Yang, Firefly algorithm, stochastic test functions and design optimisation, Int. J. Bio-Inspired Comput. 2 (2) (2010) 78–84, https://doi.org/10.1504/IJBIC.2010.032124.
[136] S.J. Nanda, G. Panda, A survey on nature inspired metaheuristic algorithms for partitional clustering, Swarm Evol. Comput. 16 (2014) 1–18, https://doi.org/10.1016/j.swevo.2013.11.003.
[137] S. Katoch, S.S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future, Multimed. Tools Appl. (2021) 8091–8126, https://doi.org/10.1007/s11042-020-10139-6.
[138] U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique, Pattern Recogn. 33 (2000) 1455–1465, https://doi.org/10.1016/S0031-3203(99)00137-5.
[139] K. Krishna, M.N. Murty, Genetic K-means algorithm, IEEE Trans. Syst. Man, Cybern. Part B 29 (3) (1999) 433–439, https://doi.org/10.1109/3477.764879.
[140] M. Wang, Y. Tseng, H. Chen, K. Chao, Expert systems with applications A novel clustering algorithm based on the extension theory and genetic algorithm, Expert Syst. Appl. 36 (4) (2009) 8269–8276, https://doi.org/10.1016/j.eswa.2008.10.010.
[141] R.H. Sheikh, Genetic algorithm based clustering: a survey, 2008 first Int Conf. Emerg. Trends Eng. Technol. 2 (6) (2008) 314–319, https://doi.org/10.1109/ICETET.2008.48.
[142] M. Sarkar, B. Yegnanarayana, D. Khemani, A clustering algorithm using an evolutionary programming-based approach, Pattern Recogn. Lett. 18 (1997) 975–986, https://doi.org/10.1016/S0167-8655(97)00122-0.
[143] Y. Ding, X. Fu, Neurocomputing Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm, Neurocomputing 188 (2016) 233–238, https://doi.org/10.1016/j.neucom.2015.01.106.
[144] J. Handl, B. Meyer, Ant-based and swarm-based clustering, Swarm Intell. (2007) 95–113, https://doi.org/10.1007/s11721-007-0008-7.
[145] D. Martens, B. Baesens, T. Fawcett, Editorial survey : swarm intelligence for data mining, Mach. Learn., no. August 2010 (2011) 1–42, doi: 10.1007/s10994-010-5216-5.
[146] B.H. Nguyen, B. Xue, M. Zhang, A survey on swarm intelligence approaches to feature selection in data mining, Swarm Evol. Comput. 54 (2020) 100663, https://doi.org/10.1016/j.swevo.2020.100663.
[147] M. Dorigo, G. Di Car, Ant colony optimization: a new meta-heuristic, in: Proc. 1999 Congr. Evol. Comput., pp. 1470–1477, 1999, doi: https://doi.org/10.1109/CEC.1999.782657.
[148] T.S. Oscar Cordon, F. Herrera, A review on the ant colony optimization metaheuristics: basic, models and new trends, Mathw. Soft Comput. 9 (2002).
[149] Y. Gu, L.O. Hall, Kernel based fuzzy ant clustering with partition validity, in: 2006 IEEE Int. Conf. Fuzzy Syst., 2006, pp. 61–65, doi: https://doi.org/10.1109/FUZZY.2006.1681695.
[150] P.M. Kanade, L. Hail, Fuzzy ant clustering by centroid positioning, 2004 IEEE Int Conf. Fuzzy Syst. (2004) 371–376.
[151] S. Kaes, M. Hossain, S.A. Ema, H. Sohn, Rule-based classification based on ant colony optimization: a comprehensive review, Appl. Comput. Intell. Soft Comput. (2022), https://doi.org/10.1155/2022/2232000.
[152] R.S. Parpinelli, An ant colony based system for data mining: applications to medical data, Proc. 3rd Annu. Conf. Genet. Evol. Comput. San Fr. (1999).
[153] L. Xing, Y. Chen, P. Wang, Q. Zhao, J. Xiong, A knowledge-based ant colony optimization for flexible job shop scheduling problems, Appl. Soft Comput. J. 10 (3) (2010) 888–896, https://doi.org/10.1016/j.asoc.2009.10.006.
[154] A. Maroosi, B. Amiri, A new clustering algorithm based on hybrid global optimization based on a dynamical systems approach algorithm, Expert Syst. Appl. 37 (8) (2010) 5645–5652, https://doi.org/10.1016/j.eswa.2010.02.047.
[155] W. Verbeke, D. Martens, C. Mues, B. Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Syst. Appl. 38 (3) (2011) 2354–2364, https://doi.org/10.1016/j.eswa.2010.08.023.
[156] S. Misra, S.K. Dhurandher, M.S. Obaidat, K. Verma, P. Gupta, Simulation modelling practice and theory a low-overhead fault-tolerant routing algorithm for mobile ad hoc networks: a scheme and its simulation analysis, Simul. Model. Pract. Theory 18 (5) (2010) 637–649, https://doi.org/10.1016/j.simpat.2010.01.008.
[157] J. Handl, B. Meyer, Improved ant-based clustering and sorting in a document retrieval interface, Int. Conf. Parallel Probl. Solving from Nat. (2002) 913–923.
[158] A. Ramos, V. Abraham, Antids: self-organized ant-based clustering model for intrusion detection system, Soft Comput. as Transdiscipl. Sci. Technol. Proc. fourth IEEE Int. Work. WSTST'05, 2005, pp. 977–986.
[159] H. Azzag, G. Venturini, A. Oliver, C. Guinot, A hierarchical ant based clustering algorithm and its use in three real-world applications, Eur. J. Oper. Res. 179 (2007) 906–922, https://doi.org/10.1016/j.ejor.2005.03.062.
[160] S. Tulin Inkaya, Kayalıgil, N. Evin, Ant colony optimization based clustering methodology, Appl. Soft Comput. 28 (2015) 301–311, doi: 10.1016/j.asoc.2014.11.060.
[161] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proc. ICNN'95-international Conf. neural networks, 1995, pp. 1942–1948, doi: https://doi.org/10.1109/ICNN.1995.488968.
[162] F. Yang, T. Sun, C. Zhang, An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization, Expert Syst. Appl. 36 (6) (2009) 9847–9852, https://doi.org/10.1016/j.eswa.2009.02.003.
[163] D. Sedighizadeh, E. Masehian, Particle swarm optimization methods, taxonomy and applications, Int. J. Comput. Theory Eng. 1 (5) (2009) 486–502.
[164] X. Cui, T. E. Potok, P. Palathingal, Document clustering using particle swarm optimization, in: Proc. 2005 IEEE Swarm Intell. Symp., 2005, pp. 1–7, doi: https://doi.org/10.1109/SIS.2005.1501621.
[165] S. Vancouver, W. Centre, W. Jatmiko, K. Sekiyama, and T. Fukuda, A PSO-based mobile sensor network for odor source localization in dynamic environment:

theory, simulation and measurement, in: 2006 IEEE Int. Conf. Evol. Comput., pp. 1036–1043, 2006, doi: https://doi.org/10.1109/CEC.2006.1688423.
[166] A.P. Engelbrecht, Dynamic clustering using particle swarm optimization with application in image segmentation, Pattern Anal. Appl. (2006) 332–344, https://doi.org/10.1007/s10044-005-0015-5.
[167] S. Das, A. Abraham, S.K. Sarkar, A hybrid rough set – particle swarm algorithm for image pixel classification, 2006 Sixth Int Conf. Hybrid Intell. Syst. (2006) 2–6, https://doi.org/10.1109/HIS.2006.264909.
[168] S. Paterlini, T. Krink, Differential evolution and particle swarm optimisation in partitional clustering, Comput. Stat. Data Anal. 50 (2006) 1220–1247, https://doi.org/10.1016/j.csda.2004.12.004.
[169] A.A.A. Esmin, R.A. Coelho, S. Matwin, A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data, Artif. Intell. Rev. (2015) 23–45, https://doi.org/10.1007/s10462-013-9400-4.
[170] H.Z. Junyan Chen, Research on application of clustering algorithm based on PSO for the web usage pattern, 2007 Int Conf. Wirel. Commun. Netw. Mob. Comput. (2007) 3705–3708, https://doi.org/10.1109/WICOM.2007.916.
[171] L. Chuang, C. Hsiao, C. Yang, Chaotic particle swarm optimization for data clustering, Expert Syst. Appl. 38 (12) (2011) 14555–14563, https://doi.org/10.1016/j.eswa.2011.05.027.
[172] R.J. Kuo, Y.J. Syu, Z. Chen, F.C. Tien, Integration of particle swarm optimization and genetic algorithm for dynamic clustering, Inf. Sci. (Ny) 195 (2012) 124–140, https://doi.org/10.1016/j.ins.2012.01.021.
[173] M. Alswaitti, M. Albughdadi, N. Ashidi, M. Isa, Density-based particle swarm optimization algorithm for data clustering, Expert Syst. Appl. 91 (2018) 170–186, https://doi.org/10.1016/j.eswa.2017.08.050.
[174] G. Dobbie, Y. Sing, P. Riddle, S. Ur, Research on particle swarm optimization based clustering: a systematic review of literature and techniques, Swarm Evol. Comput. 17 (2014) 1–13, https://doi.org/10.1016/j.swevo.2014.02.001.
[175] D. Karaboga, An idea based on honey bee swarm for numerical optimization, Tech. report-tr06, 2005.
[176] M.D.O. Dusan Teodorovic, Panta Lucic, Goran Markovic, Bee colony optimization: principles and applications, in: 2006 8th Semin. Neural Netw. Appl. Electr. Eng., 2006, pp. 151–156, doi: https://doi.org/10.1109/NEUREL.2006.341200.
[177] S.S. Ilango, S.V.M. Kaliappan, Optimization using artificial bee colony based clustering approach for big data, Cluster Comput. 22 (s5) (2019) 12169–12177, https://doi.org/10.1007/s10586-017-1571-3.
[178] E. Hancer, C. Ozturk, D. Karaboga, Artificial bee colony based image clustering method, 2012 IEEE Congr. Evol. Comput., 2012, pp. 1–5, doi: https://doi.org/10.1109/CEC.2012.6252919.
[179] A. Kumar, D. Kumar, S.K. Jarial, A review on artificial bee colony algorithms and their applications to data clustering, Cybern. Inf. Technol. 17 (3) (2017) 3–28, https://doi.org/10.1515/cait-2017-0027.
[180] P. Das, D. K. Das, S. Dey, A modified bee colony optimization (MBCO) and its hybridization with k-means for an application to data clustering, Appl. Soft Comput. J. 70 (2018) 590–603, doi: 10.1016/j.asoc.2018.05.045.
[181] J. Ji, W. Pang, Y. Zheng, Z. Wang, Z. Ma, A novel artificial bee colony based clustering algorithm for categorical data, PLoS One (2015) 1–17, https://doi.org/10.1371/journal.pone.0127125.
[182] Y.G. Yugal kumar, Sahoo, A two-step artificial bee colony algorithm for clustering, Neural Comput. Appl. 28(3) (2017) 537–551, doi: 10.1007/s00521-015-2095-5.
[183] O. Isaac, A. Jantan, A. Esther, State-of-the-art in arti fi cial neural network applications: a survey, Heliyon no. October (2018) e00938.
[184] J. Xiao, Y. Tian, S. Member, L. Xie, A hybrid classification framework based on clustering, IEEE Trans. Ind. Informatics 16 (4) (2020) 2177–2188, https://doi.org/10.1109/TII.2019.2933675.
[185] T. Fu, Engineering applications of artificial intelligence a review on time series data mining, Eng. Appl. Artif. Intel. 24 (1) (2011) 164–181, https://doi.org/10.1016/j.engappai.2010.09.007.
[186] S. Zolhavarieh, S. Aghabozorgi, Y.W. Teh, A review of subsequence time series clustering, Sci. World J. 2014 (2014), https://doi.org/10.1155/2014/312521.
[187] K. Chan, A.W. Fu, Efficient time series matching by wavelets, Proc. 15th IEEE Int. Conf. Data Eng. (1999) 126–133, https://doi.org/10.1109/ICDE.1999.754915.
[188] Y.M. Christos Faloutsos, M. Ranganathan, Fast subsequence matching in time-series databases 2 (1994) 419–429, doi: https://doi.org/10.1145/191843.191925.
[189] J. Abonyi, B. Feil, S. Nemeth, P. Arva, Modified Gath – Geva clustering for fuzzy segmentation of multivariate time-series, Fuzzy Set. Syst. 149 (2005) 39–56, https://doi.org/10.1016/j.fss.2004.07.008.
[190] V. Kavitha, M. Punithavalli, Clustering time series data stream – a literature survey," arXiv Prepr. arXiv 8(1) (2010).
[191] C.C. Aggarwal, T.J.W.R. Ctr, J. Han, J. Wang, A framework for clustering evolving data streams, in: Proc. 2003 VLDB Conf., pp. 81–92, 2003.
[192] J.A. Silva, R.C. Barros, E.R. Hruschka, J.O. Ao, Data stream clustering: a survey, ACM Comput. Surv. 46 (1) (2013) 1–23, https://doi.org/10.1145/2522968.2522981.
[193] S. Ding, F. Wu, J. Qian, H. Jia, Research on data stream clustering algorithms, Artif. Intell. Rev. (2015) 593–600, https://doi.org/10.1007/s10462-013-9398-7.
[194] A. Bifet, G. Holmes, N. Zealand, R. Gavaldà, New ensemble methods for evolving data streams, Proc. 15th ACM SIGKDD Int Conf. Knowl. Discov. Data Min. (2009) 139–148, https://doi.org/10.1145/1557019.1557041.
[195] L. Fu, P. Lin, A.V. Vasilakos, S. Wang, An overview of recent multi-view clustering, Neurocomputing 402 (2020) 148–161, https://doi.org/10.1016/j.neucom.2020.02.104.
[196] G. Chao, S. Sun, J. Bi, A survey on multiview clustering, IEEE Trans. Artif. Intell. 2 (2) (2021) 146–168, https://doi.org/10.1109/TAI.2021.3065894.
[197] M.S. Yang, A survey of fuzzy clustering, Math. Comput. Model. 18 (11) (1993) 1–16, https://doi.org/10.1016/0895-7177(93)90202-A.
[198] H. Wang, J. Wang, G. Wang, A survey of fuzzy clustering validity evaluation methods, Inf. Sci. (Ny) 618 (2022) 270–297, https://doi.org/10.1016/j.ins.2022.11.010.
[199] J. Li, H.W. Lewis, Fuzzy clustering algorithms – review of the applications, 2016 IEEE Int. Conf. Smart Cloud, 2016, doi: 10.1109/SmartCloud.2016.14.
[200] L.A. Zadeh, I. Introduction, U.S. Navy, Fuzzy Sets *, Inf. Control 353 (1965) 338–353, https://doi.org/10.1016/S0019-9958(65)90241-X.
[201] M.H.F. Zarandi, A fuzzy clustering model for fuzzy data with outliers, Int. J. Fuzzy Syst. Appl. 1 (June) (2011) 29–42, https://doi.org/10.4018/ijfsa.2011040103.
[202] E.H.P. Uspini, New approach to clustering, Inf. Control 32 (1969) 22–32, https://doi.org/10.1016/S0019-9958(69)90591-9.
[203] E.H. Ruspini, J.C. Bezdek, J.M. Keller, Fuzzy clustering: a historical perspective, IEEE Comput. Intell. Mag. no. February (2019) 45–55, https://doi.org/10.1109/MCI.2018.2881643.
[204] R. Suganya, R. Shanthi, Fuzzy C- means algorithm- a review, Int. J. Sci. Res. Publ. 2 (11) (2012) 1–3.
[205] J.C. Bezdek, FCM: The fuzzy c-means clustering algorithm, Comput. Geosci. 10 (2) (1984) 191–203, https://doi.org/10.1016/0098-3004(84)90020-7.
[206] D.C. Park, I. Dagher, Gradient based fuzzy c-means (GBFCM) algorithm, in: Proc. 1994 IEEE Int. Conf. Neural Networks, pp. 1626–1631, 1901, doi: https://doi.org/10.1109/ICNN.1994.374399.
[207] W. xin X. Zhong dong Wu, Fuzzy C-means clustering algorithm based on kernel method, IEEE Comput. Intell. Mag., 2003, doi: https://doi.org/10.1109/ICCIMA.2003.1238099.
[208] R.J. Kuo, T.C. Lin, F.E. Zulvia, C.Y. Tsai, A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis, Appl. Soft Comput. J. 67 (2018) 299–308, https://doi.org/10.1016/j.asoc.2018.02.039.
[209] D. Zhang, M. Ji, J. Yang, Y. Zhang, F. Xie, A novel cluster validity index for fuzzy clustering based on bipartite modularity, Fuzzy Set. Syst. 253 (2014) 122–137, https://doi.org/10.1016/j.fss.2013.12.013.
[210] R. Winkler, Fuzzy C-means in high dimensional spaces, Int. J. Fuzzy Syst. Appl. 1 (March) (2011) 1–16, https://doi.org/10.4018/ijfsa.2011010101.
[211] A. Stetco, X. Zeng, J. Keane, Expert systems with applications fuzzy C-means ++: fuzzy C-means with effective seeding initialization, Expert Syst. Appl. 42 (21) (2015) 7541–7548, https://doi.org/10.1016/j.eswa.2015.05.014.
[212] A. Kumar, D. Kumar, S.K. Jarial, A hybrid clustering method based on improved artificial bee colony and fuzzy C-means algorithm, Int. J. Artif. Intell 15 (2) (2017) 40–60.
[213] N. Jayalakshmi, V. Sangeeta, A. Srinuvasu, Advances in Engineering Software Taylor Horse Herd Optimized Deep Fuzzy clustering and Laplace based K-nearest neighbor for web page recommendation, Adv. Eng. Softw., vol. 175, no. August 2022, p. 103351, 2023, doi: 10.1016/j.advengsoft.2022.103351.
[214] T.P.S.P. Prabhusundhar, Prediction of rice disease using modified feature weighted fuzzy clustering (MFWFC) based segmentation and hybrid classification model, Int. J. Syst. Assur. Eng. Manag. (2023) 1–13, https://doi.org/10.1007/s13198-022-01835-7.
[215] J.B. Raja, S.C. Pandian, Computer Methods and Programs in Biomedicine PSO-FCM based data mining model to predict diabetic disease, Comput. Methods Programs Biomed. 196 (2020) 105659, https://doi.org/10.1016/j.cmpb.2020.105659.
[216] O.I. Abiodun, et al., Comprehensive review of artificial neural network applications to pattern recognition, IEEE Access 7 (2019) 158820–158846, https://doi.org/10.1109/ACCESS.2019.2945545.
[217] K. Du, Clustering: A neural network approach, Neural Netw. 23 (2010) 89–107, https://doi.org/10.1016/j.neunet.2009.08.007.
[218] B.K. Francis, S.S. Babu, Predicting academic performance of students using a hybrid data mining approach, J. Med. Syst. (2019) 1–15, https://doi.org/10.1007/s10916-019-1295-4.
[219] E. Abdel-maksoud, M. Elmogy, R. Al-awadi, Brain tumor segmentation based on a hybrid clustering technique, Egypt. Informatics J. (2015) 71–81, https://doi.org/10.1016/j.eij.2015.01.003.
[220] R. Sonawane, H. Patil, Biomedical Signal processing and control automated heart disease prediction model by hybrid heuristic-based feature optimization and enhanced clustering, Biomed. Signal Process. Control 72 (2022) 103260, https://doi.org/10.1016/j.bspc.2021.103260.
[221] R. Jain, A hybrid clustering algorithm for data mining, arXiv Prepr. arXiv, 2012, doi: 10.48550/arXiv.1205.5353.
[222] S.R. Gaddam, V.V. Phoha, S. Member, K.S. Balagani, K-means + ID3: a novel method for supervised anomaly detection by cascading K-means clustering and ID3 decision tree learning methods, IEEE Trans. Knowl. Data Eng. 19 (3) (2007) 345–354, https://doi.org/10.1109/TKDE.2007.44.
[223] P. Taylor, I. Bose, X. Chen, Journal of organizational computing and hybrid models using unsupervised clustering for prediction of customer churn, J. Organ. Comput. Electron. Commer. (2009) 131–151, https://doi.org/10.1080/10919390902821291.
[224] A. Kaur, S.K. Pal, A.P. Singh, Hybridization of K-means and firefly algorithm for intrusion detection system, Int. J. Syst. Assur. Eng. Manag. 9 (4) (2018) 901–910, https://doi.org/10.1007/s13198-017-0683-8.
[225] W.L. Al-yaseen, Z.A. Othman, M. Zakree, A. Nazri, Hybrid modified K-means with C4. 5 for intrusion detection systems in multiagent systems, Sci. World J. (2015), https://doi.org/10.1155/2015/294761.

[226] Q. Huang, R. Gao, H. Akhavan, An ensemble hierarchical clustering algorithm based on merits at cluster and partition levels, Pattern Recogn. 136 (2023), https://doi.org/10.1016/j.patcog.2022.109255.
[227] K. Chowdhury, D. Chaudhuri, A. Kumar, An entropy-based initialization method of K-means clustering on the optimal number of clusters, Neural Comput. Appl. 33 (12) (2021) 6965–6982, https://doi.org/10.1007/s00521-020-05471-9.
[228] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, An extensive comparative study of cluster validity indices, Pattern Recogn. 46 (2013) 243–256, https://doi.org/10.1016/j.patcog.2012.07.021.
[229] X. Li, W. Liang, X. Zhang, S. Qing, A cluster validity evaluation method for dynamically determining the near-optimal number of clusters, Soft. Comput. 24 (12) (2020) 9227–9241, https://doi.org/10.1007/s00500-019-04449-7.
[230] G. John, O. George, A. Thopil, Data clustering : application and trends, no. November. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10325-y.
[231] N. Bolshakova, F. Azuaje, Cluster validation techniques for genome expression data, Signal Process. 83 (4) (2003) 825–833, https://doi.org/10.1016/S0165-1684(02)00475-9.
[232] L. Wang, G. Cui, X. Cai, Fuzzy clustering optimal k selection method based on multi-objective optimization, Soft. Comput. 27 (3) (2023) 1289–1301, https://doi.org/10.1007/s00500-022-07727-z.
[233] C. Patil, I. Baidari, Estimating the optimal number of clusters k in a dataset using data depth, Data Sci. Eng. 4 (2) (2019) 132–140, https://doi.org/10.1007/s41019-019-0091-y.
[234] D. Chang, X. Zhang, C. Zheng, D. Zhang, A robust dynamic niching genetic algorithm with niche migration for automatic clustering problem, Pattern Recogn. 43 (4) (2010) 1346–1360, https://doi.org/10.1016/j.patcog.2009.10.020.
[235] E. Mangortey et al., Application of machine learning techniques to parameter selection for flight risk identification, AIAA Scitech 2020 Forum, vol. 1 PartF, no. January, 2020, doi: 10.2514/6.2020-1850.
[236] G.T. Reddy, M.P.K. Reddy, K. Lakshmanna, R. Kaluri, D.S. Rajput, Analysis of Dimensionality reduction techniques on big data, IEEE Access 8 (2020) 54776–54788, https://doi.org/10.1109/ACCESS.2020.2980942.
[237] J.P. Cunningham, Linear dimensionality reduction: survey, insights, and generalizations, J. Mach. Learn. Res. 16 (2015) 2859–2900.
[238] F. Kabir, T. Chen, S.A. Ludwig, A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction, Healthc. Anal., vol. 3, no. November 2022, p. 100125, 2023, doi: 10.1016/j.health.2022.100125.
[239] A.G. Hussien, F.A. Hashim, Enhanced COOT optimization algorithm for dimensionality reduction, in: 2022 Fifth Int. Conf. women data Sci. prince sultan Univ. (WiDS PSU), 2022, pp. 43–48, doi: 10.1109/WiDS-PSU54548.2022.00020.
[240] J. Wang, F. Biljecki, Unsupervised machine learning in urban studies: a systematic review of applications, Cities 129 (2022) 103925, https://doi.org/10.1016/j.cities.2022.103925.
[241] C.O.S. Sorzano, J. Vargas, A.P. Montano, A survey of dimensionality reduction techniques, arXiv Prepr. arXiv, 2014, pp. 1–35, doi: 10.48550/arXiv.1403.2877.
[242] E. Postma, E. Postma, Dimensionality reduction : a comparative review, J. Mach. Learn. Res. (2009).
[243] G. Szepannek, clustMixType : user-friendly clustering of mixed-type data in R, R J 10 (December) (2018) 200–208.
[244] S. Behzadi, N.S. Müller, C. Plant, C. Böhm, Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm, Int. J. Data Sci. Anal. 10 (3) (2020) 233–248, https://doi.org/10.1007/s41060-020-00216-2.
[245] G.G. Yin, Shuang, Applications of clustering with mixed type data in life insurance, Risks (2021) 1–19, https://doi.org/10.3390/risks9030047.
[246] R. Zhang, X. Liu, A Novel hybrid high-dimensional PSO Clustering algorithm based on the cloud model and entropy, Appl. Sci. (2023), https://doi.org/10.3390/app13031246.
[247] X. Hu, X. Xiong, Y. Wu, M. Shi, P. Wei, C. Ma, A hybrid clustered SFLA-PSO algorithm for optimizing the timely and real-time rumor refutations in online social networks, Expert Syst. Appl. 212 (8) (2023) pp, https://doi.org/10.1016/j.eswa.2022.118638.
[248] B. Mirzaei, B. Nikpour, H. Nezamabadi-pour, CDBH : A clustering and density-based hybrid approach for imbalanced data classification, Expert Syst. Appl. 164 (2020) (2021) 114035, doi: 10.1016/j.eswa.2020.114035.
[249] J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (15) (2005) 3201–3212, https://doi.org/10.1093/bioinformatics/bti517.
[250] L. Mohammad, A. Tajudin, E. Said, A new feature selection method to improve the document clustering using particle swarm optimization algorithm, J. Comput. Sci. 25 (2018) 456–466, https://doi.org/10.1016/j.jocs.2017.07.018.
[251] P. Agarwal, M.A. Alam, R. Biswas, Issues, challenges and tools of clustering algorithms, arXiv Prepr. arXiv, 2011, doi: 10.48550/arXiv.1110.2610.
[252] S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, Knowledge-Based Syst. 98 (2016) 1–29, https://doi.org/10.1016/j.knosys.2015.12.006.
[253] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, CSIRO Math. Inf. Sci. 304 (1998) 283–304.
[254] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, Dmkd 3 (1997) 34–39.
[255] Z. Huang, M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data, IEEE Trans. Fuzzy Syst. 7 (4) (1999) 446–452.
[256] N. Singh, A comprehensive study of challenges and approaches for clustering high a comprehensive study of challenges and approaches for clustering high dimensional data, Int. J. Comput. Appl. 4 (March) (2014), https://doi.org/10.5120/15995-4844.
[257] M. Rostami, K. Berahmand, S. Forouzandeh, A novel community detection based genetic algorithm for feature selection, J Big Data (2021) 1–27, https://doi.org/10.1186/s40537-020-00398-3.
[258] R.N. Davé, R. Krishnapuram, Robust clustering methods: a unified view, IEEE Trans. Fuzzy Syst. 5 (2) (1997) 270–293, https://doi.org/10.1109/91.580801.
[259] C.B. Hurley, Clustering visualizations of multidimensional data clustering visualizations of," J. Comput. Graph. Stat., no. November 2014, pp. 37–41, 2012, doi: 10.1198/106186004X12425.
[260] L. Xu, Y. Xu, T.W.S.C.Ã, "PolSOM : A new method for multidimensional data visualization, Pattern Recognit. 43(4) (2010) 1668–1675, doi: 10.1016/j.patcog.2009.09.025.
[261] A. Ahmad, S.S. Khan, Survey of state-of-the-art mixed data clustering algorithms, IEEE Access 7 (2019) 31883–31902, https://doi.org/10.1109/ACCESS.2019.2903568.
[262] A. Abraham, V. Ramos, Web usage mining using artificial ant colony clustering and linear genetic programming, in: 2003 Congr. Evol. Comput. CEC 2003 - Proc., vol. 2, pp. 1384–1391, 2003, doi: 10.1109/CEC.2003.1299832.
[263] E. Ditton, A. Swinbourne, T. Myers, Selecting a clustering algorithm: a semi-automated hyperparameter tuning framework for effective persona development, Array 14 (2022) 100186, https://doi.org/10.1016/j.array.2022.100186.
[264] G. Krishnasamy, A.J. Kulkarni, R. Paramesran, Expert systems with applications a hybrid approach for data clustering based on modified cohort intelligence and K-means, Expert Syst. Appl. March, 2014, doi: 10.1016/j.eswa.2014.03.021.
[265] A.S. Shirkhorshidi, S. Aghabozorgi, Big data clustering: a review, Proc. Int. Conf. Comput. Sci. Its Appl. Guimarães, Port. (2014) 707–720.
[266] J. Irani, Clustering techniques and the similarity measures used in clustering: a survey, Int. J. Comput. Appl. no. January (2016) 9–14, https://doi.org/10.5120/ijca2016907841.
[267] L.J. Deborah, R. Baskaran, A. Kannan, A survey on internal validity measure for cluster validation, Int. J. Comput. Sci. Eng. Surv. 1 (2) (2010) 85–102.
[268] S. Cha, Comprehensive survey on distance/similarity measures between probability density functions, City 1 (4) (2007) pp.
[269] E. Aljalbout, V. Golkov, Y. Siddiqui, M. Strobel, D. Cremers, Clustering with deep learning : taxonomy and new methods," arXiv Prepr. arXiv, 2018, pp. 1–12, doi: 10.48550/arXiv.1801.07648.
[270] A. Singh, A. Rana, U. Pradesh, K-means with three different distance metrics, Int. J. Comput. Appl. 67 (10) (2013) 13–17.
[271] C.X. Gao et al., An overview of clustering methods with guidelines for application in mental health research, Psychiatry Res., 327(2022) (2023) 115265, doi: 10.1016/j.psychres.2023.115265.
[272] J. Arora, K. Khatter, M. Tushir, Fuzzy c-means clustering strategies : a review of distance measures, Softw. Eng. Proc. CSI, pp. 153–162, 2018, doi: 10.1007/978-981-10-8848-3.
[273] C. Procopiuc, J.S. Park, Fast algorithms for projected clustering, 1999, pp. 61–72, doi: 10.1145/304181.304188.
[274] S. Pandit, S. Gupta, A comparative study on distance measuring approaches for clustering, Int. J. Res. Comput. Sci. 2 (1) (2011) 29–31.
[275] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, A comparison study on similarity and dissimilarity measures in clustering continuous data, PLoS One 12 (2015) 1–20, https://doi.org/10.1371/journal.pone.0144059.
[276] M. Ji, F. Xie, Y. Ping, A dynamic fuzzy cluster algorithm for time series, Abstr. Appl. Anal. (2013), https://doi.org/10.1155/2013/183410.
[277] S. Xiang, F. Nie, C. Zhang, Learning a mahalanobis distance metric for data clustering and classification 41 (2008) 3600–3612, doi: 10.1016/j.patcog.2008.05.018.
[278] M. Gomathy, K. Meena, K.R. Subramaniam, Gender clustering and classification algorithms in speech processing: a comprehensive performance analysis, Int. J. Comput. Appl. 51 (20) (2012) 9–17, https://doi.org/10.5120/8156-1533.
[279] H. Kamalzadeh, A. Ahmadi, S. Mansour, Clustering time-series by a novel slope-based similarity measure considering particle swarm optimization, Appl. Soft Comput. J. 96 (2020) 106701, https://doi.org/10.1016/j.asoc.2020.106701.
[280] S. Kosub, A note on the triangle inequality for the Jaccard distance, Pattern Recogn. Lett. 120 (2019) 36–38, https://doi.org/10.1016/j.patrec.2018.12.007.
[281] J. Xie, Z. Xiong, Q. Dai, X. Wang, Y. Zhang, A new internal index based on density core for clustering validation, Inf. Sci. (Ny) 506 (2020) 346–365, https://doi.org/10.1016/j.ins.2019.08.029.
[282] E. Rendón, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus External cluster validation indexes, Int. J. Comput. Commun. 5 (1) (2011) 27–34.
[283] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1650–1654, https://doi.org/10.1109/TPAMI.2002.1114856.
[284] D. Moulavi, P.A. Jaskowiak, R.J.G.B. Campello, A. Zimek, Density-based clustering validation, in: Proc. 2014 SIAM Int. Conf. data Min., 2014, pp. 839–847, doi: https://doi.org/10.1137/1.9781611973440.96.
[285] M. Brun, et al., Model-based evaluation of clustering validation measures, Pattern Recogn. 40 (2007) 807–824, https://doi.org/10.1016/j.patcog.2006.06.026.
[286] D.N. Campo, G. Stegmayer, D.H. Milone, A new index for clustering validation with overlapped clusters, Expert Syst. Appl. 64 (2016) 549–556, https://doi.org/10.1016/j.eswa.2016.08.021.
[287] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, S. Wu, Understanding and Enhancement of internal clustering validation measures, IEEE Trans. Cybern. 43 (3) (2013) 982–994, https://doi.org/10.1109/TSMCB.2012.2220543.

[288] A.P. Reynolds, G. Richards, B.D.E.L.A. Iglesia, Clustering rules: a comparison of partitioning and hierarchical clustering algorithms, J. Math. Model. Algorithms (2006) 475–504, https://doi.org/10.1007/s10852-005-9022-1.
[289] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.
[290] Z. Borut and K. R. Z, Validity index for clusters of different sizes and densities, Pattern Recognit. Lett. 32 (2011) 221–234, doi: 10.1016/j.patrec.2010.08.007.
[291] M. Aslam et al., Cloud migration framework clustering method for social decision support in modernizing the legacy system, Trans. Emerg. Telecommun. Technol. (2024) 1–21, doi: 10.1002/ett.4863.
[292] J. Wu, J. Chen, H. Xiong, M. Xie, External validation measures for K -means clustering: a data distribution perspective, Expert Syst. Appl. 36 (3) (2009) 6050–6061, https://doi.org/10.1016/j.eswa.2008.06.093.
[293] T. Rui, S. Fong, X. S. Yang, S. Deb, Nature-inspired clustering algorithms for web intelligence data, in: Proc. 2012 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. Work. WI-IAT 2012, pp. 147–153, 2012, doi: 10.1109/WI-IAT.2012.83.
[294] S. Alam, G. Dobbie, P. Riddle, Particle swarm optimization based clustering of Web usage data, Proc. - 2008 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IAT Work. 2008, pp. 451–454, 2008, doi: 10.1109/WIIAT.2008.292.
[295] Q. Li, B.M. Kim, Clustering approach for hybrid recommender system, in: Proc. - IEEE/WIC Int. Conf. Web Intell. WI 2003, pp. 33–38, 2003, doi: 10.1109/WI.2003.1241167.
[296] J. Ben Schafer, J. Konstan, J. Riedl, Recommender systems in e-commerce, ACM Int. Conf. Proceeding Ser. (1999) 158–166, https://doi.org/10.1145/336992.337035.
[297] D.D.D. Balwant A. Sonkamble, Speech recognition using vector quantization through modified K-means LBG algorithm, Comput. Eng. Intell. Syst. 3(7) (2012) 137–145.
[298] H.Y. Vani, M.A. Anusuya, M.L. Chayadevi, Fuzzy clustering algorithms - comparative studies for noisy speech signals, Ictact J Soft Comput (2019) 1920–1926, https://doi.org/10.21917/ijsc.2019.0267.
[299] H. Alashwal, M. El Halaby, J.J. Crouse, A. Abdalla, The application of unsupervised clustering methods to Alzheimer's disease, Front. Comput. Neurosci. 13 (May) (2019) 1–9, https://doi.org/10.3389/fncom.2019.00031.
[300] A.K. Yadav, D. Tomar, S. Agarwal, Clustering of lung cancer data using foggy K-means 1 (2018) 13–18, doi: 10.1109/ICRTIT.2013.6844173.
[301] D. Greene, A. Tsymbal, N. Bolshakova, P. Cunningham, Ensemble clustering in medical diagnostics, Proc. IEEE Symp. Comput. Med. Syst. 17 (2004) 576–581, https://doi.org/10.1109/cbms.2004.1311777.
[302] M. Kumar, M. Alshehri, R. Alghamdi, P. Sharma, V. Deep, A DE-ANN inspired skin cancer detection approach using fuzzy C-means clustering, Mob. Networks Appl. 25 (2020) 1319–1329.
[303] N.N. Gopal, M. Karnan, Diagnose brain tumor through MRI using image processing clustering algorithms such as fuzzy C means along with intelligent optimization techniques, 2010 IEEE Int. Conf. Comput. Intell. Comput. Res. (2010) 1–4, https://doi.org/10.1109/ICCIC.2010.5705890.
[304] C. Cernazanu-glavan, S. Holban, Segmentation of bone structure in X-ray images using convolutional neural network, Adv. Electr. Comput. Eng (2013) 87–94.
[305] Z. Yang, F.L. Chung, W. Shitong, Robust fuzzy clustering-based image segmentation, Appl. Soft Comput. J. 9 (1) (2009) 80–84, https://doi.org/10.1016/j.asoc.2008.03.009.
[306] R. Janani, S. Vijayarani, Text document clustering using spectral clustering algorithm with particle swarm optimization, Expert Syst. Appl. 134 (2019) 192–200, https://doi.org/10.1016/j.eswa.2019.05.030.
[307] A.J. Mohammed, Y. Yusof, H. Husni, Document clustering based on firefly algorithm, J. Comput. Sci. 11 (2015) 453–465, https://doi.org/10.3844/jcssp.2015.453.465.
[308] Y. Shi, Application of FCM clustering algorithm in digital library management system, Electron 11(23) (2022), doi: 10.3390/electronics11233916.
[309] P. Prabhu, Document Clustering for Information Retrieval – A General Perspective, Res. Gate, no. August 2011, 2019.
[310] Z. Nazeri, J. Zhang, Mining aviation data to understand impacts of severe weather on airspace system performance, in: Proc. - Int. Conf. Inf. Technol. Coding Comput. ITCC, 2002, pp. 518–523, 2002, doi: 10.1109/ITCC.2002.1000441.
[311] L. Li, S. Das, R.J. Hansman, R. Palacios, A.N. Srivastava, Analysis of flight data using clustering techniques for detecting abnormal operations, J. Aerosp. Inf. Syst. 12 (9) (2015) 587–598, https://doi.org/10.2514/1.I010329.
[312] D. K. Tasoulis, V. P. Plagianakos, M. N. Vrahatis, Unsupervised clustering of bioinformatics data, in: Eur. Symp. Intell. Technol. Hybrid Syst. their Implement. Smart Adapt. Syst., no. June, pp. 47–53, 2004.
[313] J.H. Do, D.K. Choi, Clustering approaches to identifying gene expression patterns from DNA microarray data, Mol. Cells 25 (2) (2008) 279–288.
[314] G. Kerr, H.J. Ruskin, M. Crane, P. Doolan, Techniques for clustering gene expression data 38 (2008) 283–293, doi: 10.1016/j.compbiomed.2007.11.001.
[315] F. Cai, "Clustering Approaches for Financial Data Analysis: a Survey," arXiv Prepr. arXiv, 2016, doi: 10.48550/arXiv.1609.08520.
[316] T. Li, G. Kou, Y. Peng, P.S. Yu, L. Fellow, An integrated cluster detection, optimization, and interpretation approach for financial data, IEEE Trans. Cybern. 52 (2) (2022) 13848–13861, https://doi.org/10.1109/TCYB.2021.3109066.
[317] W. Bi, M. Cai, M. Liu, G. Li, A big data clustering algorithm for mitigating the risk of customer churn, IEEE Trans. Ind. Informatics 12 (3) (2016) 1270–1281, https://doi.org/10.1109/TII.2016.2547584.
[318] A. Asma, B. Sadok, PSO-based dynamic distributed algorithm for automatic task clustering in a robotic swarm, Procedia Comput. Sci. 159 (2019) 1103–1112, https://doi.org/10.1016/j.procs.2019.09.279.
[319] O. Arslan, S. Member, D.P. Guralnik, D.E. Koditschek, Coordinated robot navigation via hierarchical clustering, IEEE Trans. Rob. 32 (2) (2016) 352–371, https://doi.org/10.1109/TRO.2016.2524018.
[320] F. Janati, F. Abdollahi, S. S. Ghidary, M. Jannatifar, J. Baltes, S. Sadeghnejad, Multi-robot task allocation using clustering method, 2017, pp. 233–247, doi: 10.1007/978-3-319-31293-4.
[321] B.S. Kumar, V. Ravi, Knowle dge-base d systems a survey of the applications of text mining in financial domain, Knowledge-Based Syst. 114 (2016) 128–147, https://doi.org/10.1016/j.knosys.2016.10.003.
[322] J. Thomas, S. Ananiadou, Applications of text mining within systematic reviews, Res. Synth. Methods (2011) 1–14, https://doi.org/10.1002/jrsm.27.
[323] A. Huang, Similarity measures for text document clustering, Proc. sixth new Zeal Comput. Sci. Res. Student Conf. no. April (2008) 9–56.
[324] R.M. Alguliyev, COSUM : Text summarization based on clustering and optimization, Expert Syst., no. August 2018, pp. 1–17, 2019, doi: 10.1111/exsy.12340.
[325] A. Agrawal, U. Gupta, Extraction based approach for text summarization using k-means clustering, Int. J. Sci. Res. Publ. 4 (11) (2014) 9–12.
[326] N. Öztürk, S. Ayvaz, Telematics and informatics sentiment analysis on Twitter: a text mining approach to the Syrian refugee crisis, Telemat. Informatics 35 (1) (2018) 136–147, https://doi.org/10.1016/j.tele.2017.10.006.
[327] S. Wakade, C. Shekar, K.J. Liszka, C. Chan, Text mining for sentiment analysis of twitter data, Proc. Int. Conf. Inf. Knowl. Eng. (2012).
[328] F. Bonchi, C. Castillo, A. Gionis, Social network analysis and mining for business applications, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–37, https://doi.org/10.1145/1961189.1961194.
[329] Y. Tseng, C. Lin, Y. Lin, Text mining techniques for patent analysis, Inf. Process. Manag. 43 (2007) 1216–1247, https://doi.org/10.1016/j.ipm.2006.11.011.
[330] A. Abbas, L. Zhang, S.U. Khan, A literature review on the state-of-the-art in patent analysis, World Pat. Inf. 37 (2014) 3–13, https://doi.org/10.1016/j.wpi.2013.12.006.
[331] A. Khadjeh, S. Aghabozorgi, T. Ying, D. Chek, L. Ngo, Text mining for market prediction: asystematic review, Expert Syst. Appl. 41 (16) (2014) 7653–7670, https://doi.org/10.1016/j.eswa.2014.06.009.
[332] O. Elharrouss, N. Almaadeed, S. Al-maadeed, Journal of visual communication and image representation a review of video surveillance systems, J. Vis. Commun. Image Represent. 77 (2021) 103116, https://doi.org/10.1016/j.jvcir.2021.103116.
[333] R. Mustafa, M.S. Hossain, An efficient strategy for face clustering use in video surveillance system, in: 2019 Jt. 8th Int. Conf. Informatics, Electron. Vis. 2019 3rd Int. Conf. Imaging, Vis. Pattern Recognit. (icIVPR), 2019, pp. 12–17, doi: 10.1109/ICIEV.2019.8858532.
[334] R. Ranjith, J.J. Athanesious, V. Vaidehi, Anomaly detection using DBSCAN clustering technique for traffic video surveillance, Seventh Int. Conf Adv. Comput. (2015) 1–6, https://doi.org/10.1109/ICoAC.2015.7562795.
[335] H. Liu, C. Ong, Variable selection in clustering for marketing segmentation using genetic algorithms, Expert Syst. Appl. 34 (2008) 502–510, https://doi.org/10.1016/j.eswa.2006.09.039.
[336] G. Arimond, A. Elfessi, A Clustering method for categorical data in tourism market segmentation research, J. Travel Res. 39 (2001) 391–397, https://doi.org/10.1177/004728750103900405.
[337] S. Dolnicar, F. Leisch, Segmenting markets by bagged clustering, Australas. Mark. J. 12 (1) (2004) 51–65, https://doi.org/10.1016/S1441-3582(04)70088-9.
[338] M. Namvar, A two phase clustering method for intelligent customer segmentation, 2010 Int. Conf. Intell. Syst. Model. Simul., pp. 215–219, 2010, doi: 10.1109/ISMS.2010.48.
[339] Q. Lin, Mobile customer clustering based on call detail records for marketing campaigns, in: 2009 Int. Conf. Manag. Serv. Sci., pp. 1–4, 2009, doi: 10.1109/ICMSS.2009.5302716.
[340] K. Kim, H. Ahn, A recommender system using GA K -means clustering in an online shopping market, Expert Syst. Appl. 34 (2008) 1200–1209, https://doi.org/10.1016/j.eswa.2006.12.025.
[341] S. Zahra, M. Ali, A. Khalid, M. Awais, U. Naeem, A. Prugel-bennett, Novel centroid selection approaches for KMeans-clustering based recommender systems, Inf. Sci. (Ny) 320 (2015) 156–189, https://doi.org/10.1016/j.ins.2015.03.062.
[342] R. Copy, B.J. Piggott, Master thesis identification of business travelers through clustering algorithms, 2015.
[343] A. Alghamdi, A hybrid method for big data analysis using fuzzy clustering, feature selection and adaptive neuro-fuzzy inferences system techniques: case of mecca and medina hotels in Saudi Arabia, Arab. J. Sci. Eng. 48 (2) (2023) 1693–1714, https://doi.org/10.1007/s13369-022-06978-0.
[344] S. Pourmohammad, R. Soosahabi, A.S. Maida, An efficient character recognition scheme based on K-means clustering, In: 2013 5th Int. Conf. Model. Simul. Appl. Optim., pp. 1–6, 2013, doi: 10.1109/ICMSAO.2013.6552640.
[345] H. Yu, J. Su, G. Cai, Y. Piao, N. Liu, M. Huang, International journal of applied earth observation and geoinformation 3DSAC: size adaptive clustering for 3D object detection in point clouds, Int. J. Appl. Earth Obs. Geoinf. 118 (October 2022) (2023) 103231, https://doi.org/10.1016/j.jag.2023.103231.
[346] A. Gaur, Handwritten Hindi character recognition using K- means clustering and SVM, in: 2015 4th Int. Symp. Emerg. trends Technol. Libr. Inf. Serv., 2015, pp. 65–70, doi: 10.1109/ETTLIS.2015.7048173.

[347] K. Sheshadri, P.K.T. Ambekar, D.P. Prasad, R.P. Kumar, An OCR system for Printed Kannada using k-means clustering, 2010 IEEE Int. Conf. Ind. Technol. (2010) 183–187, https://doi.org/10.1109/ICIT.2010.5472676.
[348] J. Yang, et al., Brief introduction of medical database and data mining technology in big data era, J. Evid. Based Med. no. January (2020) 1–13, https://doi.org/10.1111/jebm.12373.
[349] A.C. Benabdellah, A. Benghabrit, I. Bouhaddou, A survey of clustering algorithms for an industrial context, Procedia Comput. Sci. 148 (2019) 291–302, https://doi.org/10.1016/j.procs.2019.01.022.
[350] J. Erman, M. Arlitt, A. Mahanti, I.C. Methodologies, P. Recognition, Traffic classification using clustering algorithms, in: Proc. 2006 SIGCOMM Work. Min. Netw. data, pp. 281–286, 2006, doi: 10.1145/1162678.1162679.
[351] C. Tonne, et al., Defining pathways to healthy sustainable urban development, Environ. Int. 146 (2021), https://doi.org/10.1016/j.envint.2020.106236.
[352] M. Hosseinzadeh, A. Hemmati, A. Masoud, Clustering for smart cities in the internet of things : a review 25(6). Springer US, 2022. doi: 10.1007/s10586-022-03646-8.
[353] Y. Meng, Application of K-means algorithm based on ant clustering algorithm in macroscopic planning of highway transportation hub, 2007 First IEEE Int Symp. Inf. Technol. Appl. Educ. (2007) 483–488, https://doi.org/10.1109/ISITAE.2007.4409331.
[354] N.R. Kisore, C.H.B. Koteswaraiah, Improving ATM coverage area using density based clustering algorithm and voronoi diagrams, Inf. Sci. (Ny) 376 (2017) 1–20, https://doi.org/10.1016/j.ins.2016.09.058.
[355] X. Ran, X. Zhou, M. Lei, W. Tepsan, A novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots, Appl. Sci. (2021), https://doi.org/10.3390/app112311202.
[356] T.C.S.T.Y. Lin, Network security management with traffic pattern clustering, 2010 IEEE Int Conf. Comput. Intell. Comput. Res. (2014) 1757–1770, https://doi.org/10.1007/s00500-013-1218-0.
[357] A.S. Alfoudi, et al., Hyper clustering model for dynamic network intrusion detection, IET Commun. (2022), https://doi.org/10.1049/cmu2.12523.
[358] C. Sheng, Y. Yao, W. Li, W. Yang, Y. Liu, Unknown Attack traffic classification in SCADA network using heuristic clustering technique, IEEE Trans. Netw. Serv. Manag. (2023), https://doi.org/10.1109/TNSM.2023.3238402.
[359] Y. Kwon, K. Kang, C. Bae, Unsupervised learning for human activity recognition using smartphone sensors, Expert Syst. Appl., no. May, 2014, doi: 10.1016/j.eswa.2014.04.037.
[360] G. Paragliola, Gait anomaly detection of subjects with Parkinson's disease using a deep time series-based approach, IEEE Access 6 (2018) 73280–73292, https://doi.org/10.1109/ACCESS.2018.2882245.
[361] A. Ferrari, D. Micucci, M. Mobilio, P. Napoletano, On the personalization of classification models for human activity recognition, IEEE Access 8 (2020) 32066–32079, https://doi.org/10.1109/ACCESS.2020.2973425.
[362] A.O. Ige, M. Halim, M. Noor, A survey on unsupervised learning for wearable sensor-based activity recognition, Appl. Soft Comput. 127 (2022) 109363, https://doi.org/10.1016/j.asoc.2022.109363.
[363] H. Ma, Z. Zhang, W. Li, S. Lu, Unsupervised human activity representation learning with multi-task deep clustering, Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol. 5 (1) (2021) 1–25.
[364] B. Baesens, C. Mues, D. Martens, J. Vanthienen, 50 years of data mining and OR: upcoming trends and challenges, J. Oper. Res. Soc. 60 (SUPPL. 1) (2009) 16–23, https://doi.org/10.1057/jors.2008.171.
[365] C. Ozturk, E. Hancer, D. Karaboga, Improved clustering criterion for image clustering with artificial bee colony algorithm, Pattern Anal. Appl. 18 (3) (2015) 587–599, https://doi.org/10.1007/s10044-014-0365-y.
[366] Y. Lei, Y. Zhou, J. Shi, Overlapping communities detection of social network based on hybrid C-means clustering algorithm, Sustain. Cities Soc. 47 (December 2018) (2019) 101436, https://doi.org/10.1016/j.scs.2019.101436.