Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd
Anatolian Science

ISSN,e-ISSN: 2548-1304
Volume:9, Issue:1, pp:32-50, 2024
https://doi.org/10.53070/bbd.1421527

Review Paper

Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications

Tasnim ALASALI [1], Yasin ORTAKCI*[1]

[1] Faculty of Engineering, Department of Computer Engineering, Karabük University,78050, Karabük, Turkiye

(2028150011@ogrenci.karabuk.edu.tr, yasinortakci@karabuk.edu.tr)

Received:Jan.17,2024 Accepted:Mar.05,2024 Published:Jun.01,2024

Abstract— Leveraging clustering techniques is essential for data mining research and practical applications. It has traditionally functioned as a pivotal analytical technique, facilitating the organization of unlabeled data to extract meaningful insights. The inherent complexity of clustering challenges has led to the development of various clustering algorithms. Each of these algorithms is tailored to address specific data clustering scenarios. In this context, this paper thoroughly analyzes clustering techniques in data mining, including their challenges and applications in various domains. It also extensively explores the strengths and limitations characterizing distinct clustering methodologies, encompassing distance-based, hierarchical, grid-based, and density-based algorithms. Additionally, it explains numerous examples of clustering algorithms and their empirical results in various domains, including but not limited to healthcare, image watermarking, air pollution analysis, text document clustering, and the field of big data analytics. Furthermore, this paper presents a future direction for open issues, trending topics and techniques related to clustering in data mining.

Keywords : *Clustering; hierarchical; distance-based; grid-based; density-based; data mining.*

1. Introduction

Data mining is one of the most recent interdisciplinary area in computer science. Data mining is extracting useful information from vast data in warehouses, databases, or other information repositories. It automatically recognizes data patterns and extensive data sources (Hossain et al., 2019; Sharma et al., 2013). Data mining refers to the process of extracting or "mining" valuable information from massive volumes of data (Patel et al., 2014; C. Wang et al., 2013). Nowadays, people deal with vast amounts of data, often organizing and storing them as extensive datasets (Patel et al., 2014; Tiwari, 2021). Process discovery is the learning activity that contributes to developing process models from information systems' event logs (De Weerdt et al., 2013). Data mining allows for retrieving and examining intriguing insights, observable behaviors, or high-level data from various perspectives. The knowledge gained may be used for problem-solving, information management, process control, and query handling. These techniques allow decision-makers to make clear and objective choices to tackle the most pressing global issues.

The scholarly landscape shows various classification approaches to clustering in data mining (Fahad et al., 2014; Ghosal et al., 2020; Kaya & Schoop, 2022; Omar et al., 2023; Saxena et al., 2017a). This paper uses a similar presentation style to the classification framework outlined in (Saxena et al., 2017a) due to its comprehensive analysis and relevance to the current research. Figure 1 depicts diversity in clustering techniques that facilitate to identify specific and hidden data pattern (Hossain et al., 2019). The ultimate selection of the method is mainly determined by the data in the data repository and the intended model.

Clustering is a data analysis technique characterized by the grouping of objects where there is limited or no prior knowledge about the relationships between these objects within the given dataset (De Weerdt et al., 2013; Samoilenko & Osei-Bryson, 2019; W.-B. Xie et al., 2020). The main aim of clustering is to reveal any inherent classes or structures within the data. Moreover, clustering is often described as categorizing unlabeled data with minimal or no human guidance into distinct groups. These groups are formed so that objects belonging to the same cluster exhibit similar characteristics, setting them apart from objects in other clusters.

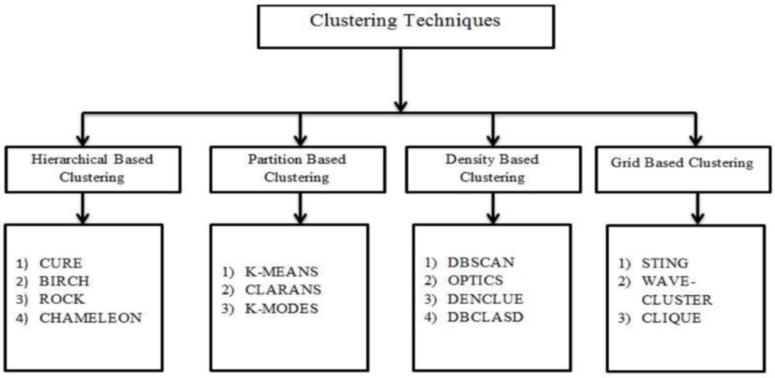

Figure 1. Data mining methods (Saxena et al., 2017a)

Clustering also falls under unsupervised learning, an integral aspect of machine learning. This involves using algorithms to identify patterns in datasets, which can be obtained through direct observation or the creation of simulated data. As stated in (A. K. Jain et al., 1999), the learning process in clustering involves the endeavor to classify data observations or independent variables without prior knowledge of a specific target variable.

Figure 2 summarizes the general procedure used while creating an unsupervised learning solution. In the first stage, the input raw data is represented by a cloud of multicolored dots, symbolizing the raw, unstructured, and possibly unclean data that serves as the input for the algorithm. The second stage is the algorithm phase, which involves interpretation and processing. The algorithm identifies patterns, categorizes data points, or formulates initial assessments during interpretation. Notably, the annotations indicate an unknown output and the absence of a training dataset, suggesting the potential utilization of unsupervised learning techniques or an exploratory nature inherent in the algorithm. In the processing step, the algorithm performs operations on the interpreted data to transform it into a structured and meaningful form. This could involve sorting, filtering, or applying mathematical models. The final stage is the 'Output', which displays the result of the algorithm. The algorithm has clustered the data into categories or patterns, as represented by the organized groups of dots.

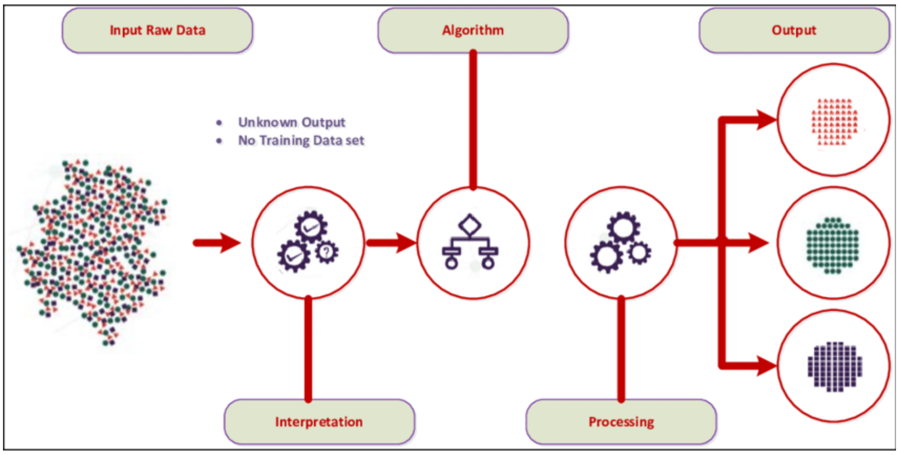

Figure 2. Unsupervised Learning (Nozari & Sadeghi, 2021)

The article is organized as follows: Section II explains the fundamental concepts and principles of clustering in data mining. Section III discusses the challenges associated with data clustering from a data mining perspective. This section addresses not only the technical hurdles but also the diverse challenges faced by researchers from various viewpoints. Section IV explores the wide range of practical applications of clustering algorithms in various domains. This section also aims to highlight the versatility and significant impact of clustering. Section V presents emerging trends and novel techniques in clustering that require more attention as hot topics in this field. Some important implications and discussion follow in Section VI, elaborating on the previously introduced issues. Section VII highlights potential areas for future research and ongoing challenges in the field. The paper concludes in Section VIII by summarizing the main points, restating the importance of clustering in data mining, and outlining general implications.

2. Clustering in Data Mining

Data quantities are growing exponentially in many scientific and industrial fields, and automated classification methods are already commonplace tools for data set exploration. Traditional clustering methods for automatic categorization help reveal a dataset's structure (Xu & Tian, 2015). Clustering algorithms are employed to partition a dataset into multiple groups, to organize the input dataset into a finite number of clusters based on shared attributes or characteristics. They can be applied to normalized and non-normalized data (Chakraborty & Nagwani, 2014). When dealing with normalized data, these algorithms typically require fewer iterations to converge. Consequently, in most scenarios, normalized data yields more favorable results than non-normalized ones (Chakraborty & Nagwani, 2014; Yedla et al., 2010). Clustering reduces a data set's dimensionality, and the primary goal of such algorithms to discern distinct groups within the data (P. K. Jain & Pamula, 2019).

Moreover, clustering encompasses various approaches, including hierarchical, partitional, grid-based, density-based, and model-based methods (Saxena et al., 2017b). It is worth noting that the efficacy of these techniques can vary based on the data type and volume used for clustering (S. R. A. Ahmed et al., 2019). Consequently, each clustering technique has a unique set of advantages and disadvantages. Tt is crucial to acknowledge that no single clustering method is universally applicable to all situations (Deng et al., 2011; Gu & Sheng, 2013). While there are over 100 known clustering techniques, it is noteworthy that only a limited subset of these algorithms has widespread usage (Ezugwu et al., 2022a). Considering this, the operational principles of widely-used clustering approaches are listed in Table 1 as:

Table 1. Comprehensive Overview of Clustering Methods

Clustering Type	**Algorithm**	**Description**	**Algorithm Steps**	**Time Complexity**	**Limitations And Considerations**
Distance-based (Kaya & Schoop, 2022)	**K-means**	K-means clustering is a method to group data items based on specific features into K clusters, with K being a positive integer number. Grouping is achieved by minimizing the distances between the cluster centroid and data points (Samoilenko & Osei-Bryson, 2019; C. Wang et al., 2013)	1. Initialize cluster centers. 2. Assign each data point to the nearest cluster. 3. Update the cluster center as the means of data points within that cluster. 4. Repeat steps 2-3 until the stopping criteria.	O (I * k * d * n) **n** is the number of data points. **k** is the number of clusters in the dataset. **d** is the number of dimensions for each data point. **I** represent the number of iterations needed to converge.	Sensitive to initialization. Sensitive to outliers. Appropriate for spherical clusters. Determining the optimal value of K.
	K-Medoids	K-Medoids clustering is like K-means but calculates medoids instead of means. It is particularly effective for small data sets but less for large ones.	K-medoids employ medoids (data points) instead of means as cluster representatives. The primary distinction between K-means and K-medoids is how they define cluster centers. In K-means, it uses the mean of data points, while in K-medoids, it designates one of the data points within the group as the center.	O(k(n-k)/2)	Less sensitive to outliers compared to K-means. However, it is not ideal for noisy data. Requires specifying K.
	Distributed K-Means	Distributed K-Means are designed for normalized data. It involves several steps: discovering max and min values of features, normalizing data, clustering using K-means, and computing centroids.	1. Distribute the data among multiple nodes. 2. Performs k-means independently on each node. REPEAT: 3. Let nodes share their centroid. 4. Calculate the global centroids UNTIL centroids of all nodes converge. 6. Assign the data points to their nearest centroid.	The total distributed complexity is O((n/p)* k * d * i + communication overhead).	Effective for normalized data. Not suitable for non-normalized data.

Hierarchical (Fahad et al., 2014)	(Agglomerative) or AGNES (agglomerative nesting)	Hierarchical methods (bottom-up) create a hierarchy of clusters. Agglomerative hierarchical clustering starts with each data point as a separate cluster and merges them iteratively based on proximity until a tree-like structure is formed.	1. Compute pattern similarity coefficients. 2. Initially, place each pattern in its cluster. 3. Merge two connected clusters into one, then recalculate inter-cluster similarity scores. 4. Repeat until there are stopping criteria.	O(n^2)	The high time complexity for large datasets. Accuracy depends on the chosen approach (top-down or bottom-up). AGNES can be quadratic in complexity.
	Divisive or DIANA (divisive analysis).	Divisive hierarchical clustering (top-down) begins with all data in one cluster and splits it into smaller clusters recursively based on dissimilarity.	1. Start with all patterns in one cluster. 2. Split the cluster using a flat clustering algorithm. 3. Apply this recursively.	O(n^2)	Top-down clustering is more accurate but computationally expensive.
Grid-based (Saxena et al., 2017a)	STING (Statistical Information Grid)	STING is a grid-based approach subdividing the zone into rectangular cells at different resolutions. Statistical information is precomputed and used for query responses. It involves a top-down approach starting from a selected stage, calculating cell confidence intervals, and gradually advancing to lower stages.	1. Exclude irrelevant cells from further consideration. 2. Upon completing the examination of the current layer, proceed to the subsequent lower level. 3. Iteratively execute the entire process until the lowest level is attained.	O(K) K represents the cumulative number of grid cells at the lowest hierarchical level.	Limited to identifying parallel or perpendicular cluster borders, not diagonal ones. Cluster sizes are constrained to the union of cells.
Density-based (Ghosal et al., 2020)	DBSCAN (Density-Based Spatial Clustering of Application with Noise)	**DBSCAN** identifies clusters based on dense data regions separated by lower-density areas. It does not require a predefined number of clusters and can find irregularly shaped clusters.	1. Select an unvisited point. 2. Remove the ε-neighborhood of the selected point. 3. If there are enough neighbors, start grouping; otherwise, label the point as an outlier. 4. If a point is part of a group, its ε-neighborhood also belongs to it. Repeat the above steps for all ε-neighborhood points. 5. Select an unvisited data point and repeat the process until all points are visited.	O(n^2) or O(n*log(n)) When utilizing active arrangements with low-dimensional data (d<=5), DBSCAN's computational complexity can be reduced to O(n log n) (Lv et al., 2016).	DBSCAN's performance hinges on ε and MinPts parameters. Time-consuming nearest neighbor search during cluster expansion. Sensitivity to initial points; struggles with varying data densities.

Figure 3 encapsulates the dynamic research development in the field, as demonstrated by the scientific output over the last five years. This trend analysis reflects academic energy and highlights the increasing relevance of clustering techniques in the broader data mining landscape. Figure 3 also displays the temporal progression of empirical studies conducted on clustering methods within the domain of data mining from 2019 to 2023. A total of 49,712 research investigations have been carried out during this period. These research studies are sourced from the Web of Science (WoS) Core Collection, accessed through a user-friendly interface that facilitates query composition and the extraction of associated scholarly works. The results indicate a notable upward trajectory in the volume of studies centered around the clustering concept in data mining during the last five years. In 2019, there were 8,468 studies, which consistently rose, peaking in 2022 at 12,296. The increase in the cumulative number of studies conducted over the last five years indicates a growing research interest in this area.

3. Open Issues and Challenges in Data Clustering

Data clustering faces numerous challenges and unresolved issues, particularly with datasets' increasing complexity and dimensionality(Thudumu et al., 2020). The primary challenges regarding data clustering identified in recent research were elaborated below:

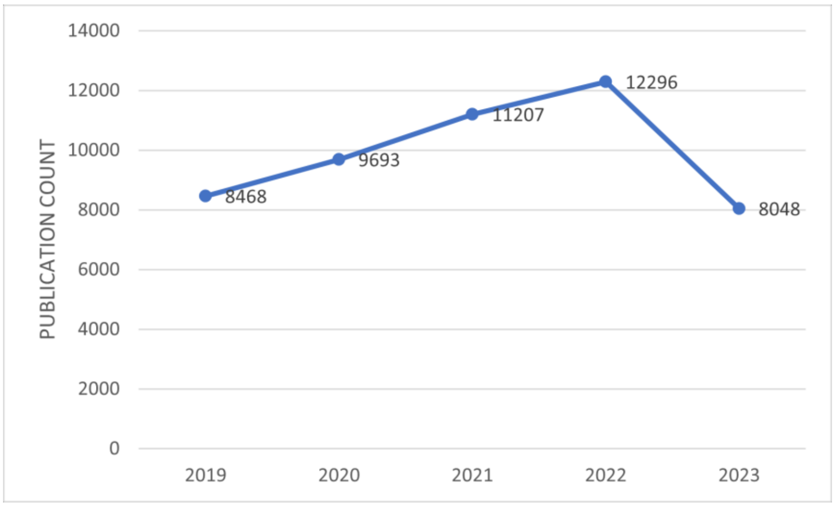

Figure 3. Evolution of publication counts on clustering in data mining (last five years)

3.1. High-Dimensional Data

In recent years, digital data production has grown exponentially in various domains. Data content has started to become vast, heterogeneous, complex, and rapidly increasing in dimensionality (Gao et al., 2017). High Dimensional Data (HDD) applications have been identified in biomedical research, web technologies, education, medicine, business, and social media platforms (Curiskis et al., 2020). HDD is found in various formats, including text (Ikotun et al., 2023a), digital images (Richards & Richards, 2022), speech signals (Chalapathi et al., 2022), and video content (Souiden et al., 2022). Machine learning algorithms face significant challenges in data mining due to the high dimensionality of this data, making tasks more complicated (Ayesha et al., 2020a). These challenges include:

3.1.1. Inadequacies in Similarity Metrics

Considering the appropriate distance metric when working with high-dimensional spaces is important(Ghazal, 2021). In high-dimensional spaces, traditional similarity metrics like the Euclidean distance may lose effectiveness (Thrun & Ultsch, 2021). This is because the spatial distance between the nearest and farthest points tends to become similar as the number of dimensions increases (Chakraborty & Das, 2020). Therefore, current similarity metrics struggle to effectively discriminate between high-dimensional data points. These metrics often fail to capture the intricacies of the data's underlying geometry, necessitating the development of new approaches.

On the other hand, instead of Euclidean distance, Manhattan distance (L1 norm) may offer better performance by considering the separate absolute differences along each dimension (Rehman & Khan, 2021). This metric can be more robust in spaces where the Euclidean distance becomes inflated due to the curse of dimensionality.

3.1.2. Visualization Limitations

Visualizing high-dimensional data is impractical without first reducing its dimensionality through techniques such as PCA (Principal Component Analysis) and t-SNE (t-Distributed Stochastic Neighbor Embedding) (Pareek & Jacob, 2021). These methods compress data into two or three dimensions for visualization. However, this compression can lead to the loss of critical information and may not truly represent the complex relationships in the original high-dimensional space. These visualizations can provide valuable insights, but they may oversimplify or distort the structure of the data.

3.1.3. Correlation between Features

High-dimensional datasets frequently have correlated features that can distort the distribution of data points. This correlation can mislead clustering algorithms, causing them to form clusters based on these correlations rather than the underlying patterns (Ray et al., 2021). To cluster effectively in such contexts, algorithms must recognize and adjust for the influence of feature correlation, ensuring that clusters reflect genuine data structures.

3.1.4. Loss of Variance after Dimension Reduction

Dimension reduction techniques are essential for managing high-dimensional data, making it more manageable for analysis. However, these methods inherently involve trading off some data variance, which measures the information content (Duan et al., 2023a; Li et al., 2023). Reducing dimensions can discard information that might have been crucial for accurately clustering the data (Ukey et al., 2023). Balancing the need to reduce dimensions to preserve as much information as possible is a significant challenge in high-dimensional data clustering. The deep learning paradigm can help in this issue. Autoencoders compress data into a lower-dimensional space by learning an encoded representation that preserves much of the original data's relevant information (Chadebec et al., 2022; Wickramasinghe et al., 2021). The compressed representation can be used for more effective clustering. This addresses the sparsity and visualization issues that are inherent to high-dimensional data.

However, several strategies have been proposed to address the challenges of clustering high-dimensional data.

- Subspace and Ensemble Clustering: Subspace and ensemble clustering methods address the complexity of high-dimensional data by tailoring the clustering process to the internal structure of the dataset. Subspace clustering identifies clusters within smaller, more relevant portions of the total feature space, recognizing that not every dimension contributes equally to meaningful clustering (Qu et al., 2023; L. Wang et al., 2023). This technique is particularly adept at uncovering hidden clusters that may not be visible when considering the full range of dimensions due to noise or irrelevance of certain features. Ensemble clustering, on the other hand, aggregates the results of multiple clustering algorithms, leveraging the diversity of these approaches to gain a more accurate and stable view of the underlying structure of the data(G. He et al., 2022; Shi et al., 2023). Ensemble clustering effectively deals with the problems caused by the variability and complexity of high-dimensional datasets by combining the best parts of different clustering methods(Shi et al., 2023). This creates a robust solution that is more likely to accurately reflect how the data is grouped. These strategies work well together to handle the complex aspects of high-dimensional data clustering. They use the specificity of subspace methods and various ensemble techniques to get better clustering results.
- Incorporating Domain Knowledge: Incorporating domain knowledge into the clustering process is a critical strategy that significantly improves the quality and relevance of the identified clusters (Duan et al., 2023b). This approach involves a nuanced understanding and application of specific knowledge related to the field or domain under study. By carefully selecting the most relevant features to the domain and adjusting the distance metrics to reflect the meaningful relationships within the data more accurately (Qoku & Buettner, 2023), clustering algorithms can be fine-tuned to produce more accurate results. Managing the dimensionality reduction process to keep information vital to the domain emphasizes the most critical dimensions, leading to more accurate and understandable clustering results. This careful focus on the most relevant aspects of the data, informed by deep domain expertise, enables the discovery of truly insightful patterns and relationships that might otherwise remain hidden.

Addressing the complexity of clustering in high-dimensional data requires a multi-faceted approach that combines innovative algorithmic strategies with insights from domain knowledge. These mitigation strategies represent ongoing research areas, with developments in algorithm design, dimensionality reduction, and computational methods playing a critical role in advancing the field. The successful application of these strategies can significantly improve the accuracy and efficiency of clustering high-dimensional data, opening new avenues for discovery and analysis in various domains.

3.2. Absence of an Overall Evaluation Metric

In data clustering, a notable challenge is that many algorithms can produce multiple solutions for the same data set, complicating the task of accurately evaluating these solutions. This multiplicity is often due to the initial conditions or parameters set by the algorithm, such as the placement of initial centroids in K-means clustering. As a result, each algorithm run may produce a different clustering result, especially in complex data sets with no explicit or distinct groupings. This variability introduces uncertainty in determining which solution most accurately reflects the underlying structure of the data (Mussabayev et al., 2023; Saklani et al., 2023).

Evaluating the accuracy of the generated solutions is a big issue in unsupervised learning tasks such as clustering (Ezugwu et al., 2022b). Due to the lack of ground truth labels in clustering, traditional accuracy metrics designed for classification tasks are not directly applicable. Instead, alternative methods such as internal validity indices (e.g., silhouette scores, Davies-Bouldin index). However, these methods have limitations and may not fully capture the clustering quality (Fakir & El Iklil, 2021). For example, they may favor more compact and well-separated clusters, which may not always match the natural groupings in the data, especially in cases where clusters are not spherical or have varying densities.

The challenge of dealing with multiple solutions and evaluating their accuracy emphasizes the need for robust methodologies in clustering analysis. It is essential to incorporate multiple runs and consider a range of evaluation metrics to obtain a comprehensive view of the algorithm's performance. Furthermore, it highlights the importance of domain expertise in interpreting clustering outcomes. Understanding the data context can offer vital insights into the most relevant and meaningful solution. A balanced approach, combining mathematical expertise with practical insights, is necessary to tackle these challenges.

3.3. Computational Limits

Processing large datasets and clustering algorithms are inherently complex, leading to challenges with computational limits. The issue revolves around balancing the computational resources required by clustering algorithms with the available processing power (Wright & Ma, 2022). Clustering algorithms can be computationally intensive, especially those designed for large volumes of high-dimensional data. They often require significant memory and processing time, mainly as the dataset grows in size and complexity. For example, the K-means algorithm, famous for its simplicity and effectiveness in many scenarios, exhibits a computational complexity that scales unfavorably with increased data points (N) and the number of dimensions (D). The time complexity of K-means is O(NKD), where K is the number of clusters. As any of these parameters increase, the computation time also increases, making the algorithm impractical for large datasets or those with high dimensions (Kharchenko, 2021).

Advanced methods are being used to overcome computational problems in data clustering. One such method is the application of dimensionality reduction algorithms (Ayesha et al., 2020b; Guo et al., 2022; R. Liu et al., 2020; Reddy et al., 2020), including t-Distributed Stochastic Neighbor Embedding (t-SNE) (H. Liu et al., 2021) and Uniform Manifold Approximation and Projection (UMAP) (Yu et al., 2023). These algorithms reduce high-dimensional data into lower-dimensional spaces, simplifying the clustering task and reducing computational overhead.

Utilizing distributed computing paradigms and creating and combining sophisticated algorithms is also possible. Distributed computing frameworks such as Apache Hadoop (Lydia et al., 2020) and Apache Spark (N. Ahmed et al., 2020) have revolutionized the scalability of clustering processes. These frameworks enable parallel processing by distributing data and computations across multiple computing nodes. By leveraging such architectures, clustering algorithms can process vast datasets in a fraction of the time previously required. Apache Spark's MLlib is a prime example of a library that implements clustering algorithms optimized for distributed environments (JayaLakshmi & Kishore, 2022).

Algorithms like DBSCAN and HDBSCAN are designed to cluster large datasets efficiently by minimizing distance calculations and adapting to the dataset's structure to optimize computational resources. Despite the existence of dimensionality reduction techniques, distributed computing frameworks, and some adaptable clustering algorithms, the quest for even more efficient clustering algorithms remains crucial.

3.4. Issues in Feature Weighting

Traditional clustering methods often treat all features as equally important, leading to potential inaccuracies in the outcomes in real-world datasets as features have varying relevance (Ikotun et al., 2023b). To improve clustering success, it is essential to use distance metrics that account for variable feature relevance. By assigning higher weights to more relevant features, the clustering algorithm better captures the true similarity between data points, leading to more accurate clustering results.

One modern approach to tackle this challenge is Feature Weighting (Oskouei et al., 2021; Sinaga et al., 2021; Sun et al., 2022). This method assigns weights to features based on their relevance to the specific clustering objective; thereby, the similarity between the instances can be reflected more accurately. Algorithms such as Weighted K-Means (J. Xie et al., 2023) or Feature Weighted Fuzzy C-Means (Kuo et al., 2021; Mohammadi et al., 2023) exemplify this approach by considering different features with varying impact, leading to more accurate similarity calculations and improved clustering outcomes. Another advanced strategy is using adaptive algorithms, which dynamically adjust their parameters or feature weights based on the characteristics of the dataset. For example, Adaptive Resonance Theory and its variants can learn incrementally from the data, adjusting their clustering strategy to emphasize more informative dimensions and de-emphasize less relevant ones (da Silva et al., 2022). In addition, dimension reduction techniques can be incorporated into the pre-processing of the data before clustering. Specifically, feature selection-oriented techniques like Principal Feature Analysis can highlight significant features. This ensures that the clustering algorithm focuses on features that carry the most information about the data's structure.

As a result, clustering algorithms benefit significantly from incorporating feature weighting techniques, suggesting a need for further research in this area.

4. Diverse Applications of Clustering Algorithm

This section explores the diverse applications of clustering algorithms across various domains, including healthcare, social network analysis, and market segmentation. It also highlights the practical implementations of these algorithms in real-world scenarios. Clustering is often used as a first step in several data mining tasks, including summarizing data for classification (Iam-On & Boongoen, 2015), finding patterns (M. He & Chen, 2024; lahmood HAMEED & DAKKAK, 2022; Shah et al., 2023), formulating and testing hypotheses (Price et al., 2021;

Rubarth et al., 2021), compressing data (CERNIAN et al., 2011; Pham et al., 2010; Z. Xie et al., 2009), reducing the number of dimensions (Bahadori & Charkari, 2018), and finding outliers (ALASALI & DAKKAK, 2023; Dakkak et al., 2015; Mayanglambam et al., 2023). It is also an essential part of collaborative filtering (Kannout et al., 2023), recommendation systems (Ambikesh et al., 2023), exploring multimedia data (Kadiravan et al., 2021), investigating biological data (Mrukwa & Polanska, 2022), closely looking at social networks (Jeong et al., 2023; Kim et al., 2023; Marqués-Sánchez et al., 2023), and identification of changing trends (Dakkak et al., 2021; Sabitha & Bansal, 2017). Table 3 presents a detailed discussion of different research on clustering algorithms and their applications. It shows the outcomes and contributions of each study, providing valuable insights for researchers and practitioners in data clustering.

Table 3. Overview of research papers on clustering algorithms and their application in various fields

Ref.	Year	Algorithm	Approach	Outcomes	Field of Application
(Shrifan et al., 2022)	2022	K-means Modification	Utilizing Tukey's rule and a new distance metric	Improved clustering accuracy and centroid convergence. The proposed distance metric outperforms most existing metrics. Significantly enhances overall clustering accuracy by up to 80.57% on nine standard multivariate datasets.	Data Clustering
(Rahayu et al., 2020)	2020	K-Means Algorithm	Using CRISP-DM methodology for clustering	Classification of Rabies vulnerability.	Veterinary/Health care
(Alomari et al., 2023)	2023	ACUTE (Efficient and Scalable Spatial Data Clustering)	Utilizing topological relations for spatial object clustering, reducing the need for pairwise distance calculations	Enhanced efficiency and scalability in spatial data clustering. Extensive testing against synthetic and real-world datasets shows superior precision, recall, and error rates compared to state-of-the-art techniques.	Spatial Data Clustering
(Anam et al., 2023)	2023	K-means Clustering Algorithm	Combining with the Bat Algorithm (BA)	Developing a classification model for diagnosing diabetes mellitus using K-means clustering optimized with BA. Experimental results show improved performance compared to standard K-means in all evaluation metrics. Slightly higher computational time.	Healthcare/Diabetes Diagnosis
(Vandhana & Anuradha, 2021)	2021	Ensemble Clustering	Enhanced ensemble clustering for air pollution data	Addressing issues of cluster shape and number of clusters in air pollution data. Overcoming bias and variance in traditional clustering through ensembling. Improved clustering performance compared to basic clustering algorithms. Identification of healthy and unhealthy regions to control contamination. Handling uncertain objects in clustering without prior data information.	Air Pollution Analysis, Public Health
(Hassan et al., 2023)	2023	Hybrid Genetic Algorithm and K-means Clustering	Integration of genetic algorithm and k-means clustering for colored image watermarking	Optimal cluster centroids were obtained by integrating genetic algorithm and k-means clustering. Improved distribution of cover and watermark pixels into clusters for reduced perceptible changes in the watermarked image. Adopt the method of the least significant bit for concealment. Enhanced imperceptibility and resistance against common attacks in image watermarking.	Image Watermarking
(Karthikeyan et al., 2020)	2020	K-means Clustering and Hierarchical Clustering	Differentiation and comparison of clustering techniques (K-means and hierarchical clustering)	Evaluation based on execution time and memory usage. K-means is suitable for larger datasets with minimum execution time and memory usage. Agglomerative clustering is optimal for smaller datasets with overall minimal memory consumption.	Data Mining, Clustering, Fleet Management

(Bansal et al., 2017)	2017	Improved K-Means	Enhancement of the K-Means clustering algorithm to automatically determine the number of clusters and assign clusters to un-clustered points.	Automatic determination of the number of clusters. Improved accuracy and reduced clustering time. Application in cancer prediction.	Data Mining, Clustering, Healthcare
(Maia et al., 2020)	2020	MicroTEDAclus Algorithm	Development of an evolving algorithm (MicroTEDAclus) for clustering data streams from arbitrary distributions, with a focus on robustness and efficiency.	Competitive performance for online clustering of data streams with arbitrary shapes. Robust parameter settings require tuning of only one variable. – Ability to work with larger data sets, including high-dimensional datasets. Linear complexity in dimension-dependent processing steps. Potential for further optimization, including mechanisms for reducing the number of micro-clusters. Application to various tasks like fault detection, forecasting, or classification.	Data Stream Analysis
(Han et al., 2017)	2017	Bird Flock Gravitational Search Algorithm	Development of a novel data clustering algorithm based on a modified Gravitational Search Algorithm.	Introduction of a mechanism inspired by collective bird response to enhance diversity. Exploration of a more expansive search space, escaping suboptimal solutions. Evaluation of 13 benchmark datasets, outperforming various other clustering algorithms.	Data Mining, Clustering, Machine Learning
(Dhas et al., 2022)	2022	Distributed-Parallel Particle Swarm Optimization with k-means	Proposal of a distributed-parallel particle swarm optimization algorithm combined with k-means for clustering large data sets.	Improved clustering performance for large data sets. Reduced computational steps.	Data Clustering, Big Data Analysis
(Kuwil et al., 2019)	2019	Critical Distance Clustering Algorithm	Distance-based, Robust, and Dynamic Clustering	Improved clustering performance, simplicity, flexibility, and robustness. Effective handling of outliers. Parameter-free cluster generation. Cluster validity indicators. The findings demonstrate that the novel algorithm surpasses well-known clustering methods, including MST-based clustering, K-means, and DBSCAN.	Data Clustering
(Purwandari et al., 2020)	2020	K-means, Spectral Clustering, and Agglomerative Clustering	Application of data mining techniques for measuring customer satisfaction in a family restaurant	Agglomerative clustering proves effective in segmenting customers based on shared characteristics. The study provides insights into customer satisfaction measurement and offers recommendations for restaurant improvement.	Customer Satisfaction Management
(Sahoo et al., 2023)	2023	Opposition Learning Based Improved Bee Colony Optimization	Enhancing the rate of convergence and quality of clustering in data mining by introducing “opposite bees” created using opposition-based learning. These bees explore the solution space alongside mainstream bees via the Bee Colony Optimization-based clustering algorithm. A steady-state selection procedure, crossover, and mutation operations are employed to balance explorations and prevent local optima.	The results show that this algorithm performs better than newly proposed benchmarks, especially regarding the rate of convergence, the quality of the clustering, and the ability to explore and exploit.	Data mining and data engineering with a focus on data clustering for classification and regression problems.

(Krishnaswamy et al., 2023)	2023	(MR-HDBCC) Map Reduce-based Hybrid Density-Based Clustering and Classification Algorithm	Developing an efficient algorithm for clustering and classifying big data utilizing Map Reduce.	The MR-HDBCC technique has been validated using benchmark datasets and has proven effective in several ways. The algorithm offers an efficient solution for clustering and classifying large datasets, especially those with complicated patterns and noisy data.	Big data analytics
(Tejasree & Chandra Mohan, 2023)	2023	(IDBEFM) Improved Differential Bond Energy Algorithm with Fuzzy Merging	Enhancing text document clustering by addressing the issue of sparse and uninformative features.	The test results show that the IDBEFM technique is better than standard clustering algorithms for text document clustering. It handles the challenges of sparse and uninformative features in text documents effectively.	Text document clustering mainly focused on addressing issues related to sparse and uninformative features.
(Açmalı & Ortakcı, 2021)	2021	Metaheuristic approach in data clustering	Since data clustering is an np-hard optimization problem, this study offers various metaheuristic optimization algorithms to solve clustering problems.	The clustering performance of different metaheuristic algorithms is compared with each other and with K-Means. The Grey Wolf Optimisation Algorithm produces better clustering results than K-Means in three benchmark datasets. The Harmony Search algorithm produces the worst results in the group.	Metaheuristic, Clustering, Data Mining
(Ortakci, 2017)	2017	Paralellizaiton of PSO in data clustering	Particle swarm optimization (PSO) is an optimization algorithm based on populations suitable for data clustering applications. Given its population-centric nature, this study explores a parallel PSO approach to enhance clustering speed.	A CPU-parallelized version of the PSO algorithm (P-PSO) is presented in clustering. P-PSO produced better results than classic PSO regarding both speed and clustering success.	Metaheuristic, Parallelization, Clustering

5. Trending Techniques and Application Areas of Clustering in Data Mining

Recent advancements have led to the development of innovative clustering algorithms that address scalability, accuracy, and the processing of complex, high-dimensional data. These techniques enhance the detection of hidden patterns and facilitate the analysis of large datasets across distributed environments. This section discusses state-of-the-art clustering methods and their recent applications, highlighting their crucial role in transforming raw data into useful insights across various fields.

5.1. Deep Learning-Based Clustering

This approach combines traditional clustering methods with new deep-learning techniques to improve feature extraction and identify complex patterns in data. For instance, techniques such as autoencoders in Deep Embedded Clustering (DEC) enable more effective dimensionality reduction and clustering. It is renowned for its exceptional ability to mine meaningful patterns from data without supervision. The DEC framework has been expanded and enhanced through various studies to address its initial limitations, such as the lack of discriminative feature learning and the challenge of clustering mixed data types. One significant extension is the Contrastive Deep Embedded Clustering (CDEC), which introduces contrastive learning to DEC (Amirizadeh & Boostani, 2021; Sheng et al., 2022). By constructing positive and negative samples and maximizing the distance between them, CDEC enhances the model's ability to capture representative and discriminative features, substantially improving clustering effectiveness across several public datasets.

Another framework proposed by (Lee et al., 2022) addresses the challenges of mixed data by introducing a method to handle both numerical and categorical features. They use soft-target updates to improve convergence stability and performance on mixed data benchmarks. Deep Embedding Clustering Based on Contractive Autoencoder (DECCA) has been introduced to handle high-dimensional documents (Diallo et al., 2021). DECCA utilizes contractive autoencoders to preserve important information and maintain data point locality, further demonstrating DEC's adaptability to diverse data types.

5.2. Scalable Clustering Algorithms

Scalable clustering algorithms are essential for processing vast amounts of data efficiently in the era of big data. Apache Spark, a distributed computing framework, has scaled traditional algorithms like K-means for big data applications. For example, the Scalable Fuzzy Clustering with Anchor Graph (SFCAG) method, introduced by (C. Liu et al., 2022), dramatically alters traditional fuzzy clustering by integrating the anchor graph technique. This method addresses scalability through efficient anchor selection and a sparse graph construction. Unlike traditional methods that require extensive distance calculations and iterative solution updates, SFCAG uses a trace ratio model for fast membership matrix learning, resulting in linear time complexity relative to data size. This improvement speeds up the clustering process and maintains high effectiveness and scalability across diverse datasets. Similarly, the POFCM algorithm, developed by (Pérez-Ortega et al., 2023), enhances the conventional Fuzzy C-Means clustering technique by adopting a parallel processing methodology employing OpenMP. It is particularly impactful for extensive datasets characterized by numerous clusters and dimensions, resulting in significantly faster solution times and superior parallel efficiency compared to the conventional method.

In addition, Mortensen et al. (2023) presented the MARIGOLD algorithm, which significantly enhances K-Means clustering for high-dimensional data. MARIGOLD reduces the need for repetitive Euclidean distance computations by integrating a unique combination of a tight distance-bounding scheme, stepwise calculations over a multiresolution transform, and leveraging the triangle inequality. This approach improves the scalability and efficiency of K-Means in handling high-dimensional datasets and demonstrates the potential for real-time applications such as ARPES experiments.

5.3. Ensemble Clustering

Ensemble clustering is an advanced technique in data mining that aims to improve clustering performance by combining multiple clustering algorithms. This approach integrates the strengths of various algorithms to improve robustness and applicability, addressing the limitations of single clustering algorithms. Many studies utilize ensemble clustering techniques to enhance clustering performance in data mining (Chatterjee & Das, 2023; J. Chen et al., 2023; Elgarhy et al., 2023; Q. Huang et al., 2023; Nie et al., 2023). The effectiveness of these techniques is showcased through some innovative approaches.

One innovative method in this field is Fuzzy-Rough Induced Spectral Ensemble Clustering. This approach recognizes the inherent differences in the reliability and significance of clusters produced by base clustering algorithms (Yue et al., n.d.). It uses fuzzy-rough sets to manage the imprecision and vagueness commonly found in real-world data. On the other hand, the Ensemble Clustering with Attentional Representation (ECAR) method uses neural networks to improve ensemble clustering by capturing higher-order fusion information from related groups of base partitions (Hao et al., 2023). It employs an attentional network to encode the significance of each sample’s association with its group, achieving adaptive refinement of base partition weights. This leads to enhanced clustering diversity, consistency, and outperformance of existing methods.

The ensemble hierarchical clustering algorithm proposed by (Q. Huang et al., 2023) prioritizes the selection of primary clusters based on merit, using Normalized Mutual Information criteria for evaluation. The selected clusters are then re-clustered into hyper-clusters, with instances assigned based on similarity defined by merit and cluster size. The algorithm has demonstrated superior performance on UCI datasets.

5.4. Graph-Based Clustering

Graph-based clustering uses graph theory to explore complex relationships within data, providing a nuanced perspective that traditional clustering techniques may lack. This approach treats data points as nodes in a graph, with the connections (edges) between them representing relationships or similarities. Algorithms like Spectral Clustering and Graph Convolutional Networks (Phan & Nguyen, 2024) use the graph’s structure to partition it into clusters based on the connections. Spectral Clustering utilizes the eigenvalues of the Laplacian matrix to obtain a low-dimensional representation that simplifies cluster separation. Graph Convolutional Networks and their derivatives are recognized for their capacity to integrate node feature data directly into the clustering procedure. These methods have demonstrated significant effectiveness in domains where data is inherently relational or networked, such as social media analysis (Ollagnier et al., 2023; Phan et al., 2023; Utku et al., 2023; Zhong et al., 2023), protein-protein interaction networks (H. Chen et al., 2023; Fu et al., 2024), and recommendation systems (Choudhary et al., 2023; F. Wang et al., 2023). This reflects a growing interest in utilizing the rich structural and feature-based information encoded in graphs for clustering tasks.

5.5. Incremental Clustering

Incremental clustering is a modern data mining approach that can adapt to new data without reprocessing the entire dataset (Laohakiat & Sa-Ing, 2021). It is beneficial for scenarios where data arrives in streams or batches over time, requiring a flexible clustering process that can update existing clusters with new information as it becomes available. Incremental clustering algorithms, such as online K-Means (Abernathy & Celebi, 2022), Incremental DBSCAN (Azhir et al., 2021; Bhattacharjee & Mitra, 2020), and Density-Based Stream Clustering (DBSTREAM) method (Bechini et al., 2020; Faroughi et al., 2023), can efficiently integrate new data points into pre-existing cluster frameworks. This maintains up-to-date cluster assignments and centroids with minimal computational overhead. Unlike traditional clustering techniques that require a static dataset, these algorithms can run on dynamic datasets.

Due to its scalability, incremental clustering is highly effective for processing real-time data. This is a critical capability for managing extensive data streams, such as sensor networks, social media analytics, and real-time market analysis (Ran et al., 2023). These algorithms ensure that the clustering output accurately reflects the changing nature of the data by utilizing techniques such as micro-cluster maintenance and dynamic cluster merging. Incremental clustering techniques aim to minimize the impact of noise and outlier data points, improving the reliability of clustering results. This approach reflects a shift towards more adaptive and efficient data mining strategies that can handle the challenges posed by dynamic and voluminous datasets.

5.6. Multi-View Clustering

Multi-view clustering is an advanced technique in data mining and machine learning (M.-S. Chen et al., 2022), which aims to use multiple data sources to gain diverse perspectives. This method assumes that integrating data from different viewpoints can provide a more comprehensive and nuanced understanding of the patterns within the data. This process aggregates and synthesizes information to perform clustering, resulting in a more accurate and meaningful grouping of data points. Multi-view clustering is helpful in scenarios where single-source data may not provide a complete picture (Haris et al., 2024). This is especially true in multimedia information retrieval, social network analysis, and bioinformatics, where data can be multimodal and multi-faceted.

It uses various algorithms to handle the challenges of integrating heterogeneous data, such as scale, distribution, and type discrepancies. These algorithms aim to find a consensus clustering that best represents the data across all views. They achieve this by leveraging co-training, multi-kernel learning, and cross-view consensus. This approach overcomes the limitations of single-view clustering approaches (Haris et al., 2024).

Moreover, it improves clustering performance by using complementary information from various views, thus making clustering analysis resilient against noise and missing values in any individual perspective enhances analysis robustness against noise and missing values in any single view. Multi-view clustering represents a significant stride towards more comprehensive and reliable data analysis techniques in the era of big data and complex information systems.

6. Discussion and Implications

Our investigation indicates that clustering techniques show different levels of accuracy and uncertainty, particularly when dealing with noisy data. Notably, the K-Means clustering algorithm outperforms its counterparts mainly when applied to large datasets. K-Means is adept at generating high-quality clusters, and its performance improves as the number of clusters increases. However, it is essential to note that K-Means is unsuitable for categorical data, while hierarchical clustering algorithms perform well with categorical data. As a solution, agglomerative and divisive hierarchical algorithms are preferred for categorical records. However, to improve the performance of K-Means for categorical data types, an adaptive approach has been suggested for K-Means. This involves assigning rank values to categorical characteristics, which converts categorical data into numeric format. On the other hand, it should be noted that density-based approaches produce very promising results when the datasets are not sufficiently disjoint (Ali & Kadhum, 2017; Yuan & Yang, 2019).

High-dimensional data presents complex challenges. The recent surge in data generation has increased the demand for advanced clustering techniques. Although strategies for clustering high-dimensional data and integrating domain-specific information into the process are being developed, further research is required in this field. Similarly, as the evaluation criteria in clustering can not consistently produce dependable results, they reduce the reliability of clustering results. To address this issue, it is necessary to develop new evaluation criteria that are not dependent on the data set and take into account more than one criterion. Additionally, feature selection can improve the performance of clustering algorithms by reducing data size and eliminating noise and irrelevant information. Additionally, it enables faster and more accurate clustering of big data while using less memory and computer resources. Therefore, it is crucial to explore new feature selection methods for data clustering.

While there is no one-size-fits-all clustering algorithm, recent approaches such as Apache Spark-supported scalable clustering algorithms, ensemble clustering, and graph-based clustering have increased the success of clustering. These approaches represent a significant shift towards more dynamic, robust, and efficient frameworks in data mining. On the other hand, the innovation of incremental and multi-view clustering introduces scalable and adaptable strategies for real-time data stream analysis.

7. Future Research Directions and Unresolved Challenges

Although data mining is a valuable tool, it can present challenges when used on a large scale. These issues include performance, data complexity, strategies, and methods. To realize the full potential of data mining, it is critical to recognize and address its challenges. In this context, future research direction can be summarized as follows:

- Efficiency and Scalability of Algorithms: To successfully extract insights from the vast amount of data, data mining algorithms must exhibit both efficiency and scalability. Creating a parallel formulation of an improved rough k-means algorithm is suggested to open up the path to a more compelling future. This work would significantly improve algorithmic performance and accelerate data extraction procedures.
- Privacy and Security Concerns: Information mining raises concerns about data security, privacy, and management. Data disclosure without permission, such as a retailer revealing customer purchase details, underscores the need for enhanced data protection. To improve data privacy and security in cloud environments, it is recommended to establish a unified system implementing blockchain technology within any clustering algorithm.
- Handling Complex Data Types: Managing complex data types, such as graphical, temporal, and spatial data, is necessary in today's data landscape. New clustering approaches should be investigated to handle different and complex data types.
- Performance Enhancement: The data mining framework's efficiency depends on the algorithms and methodologies used. However, existing strategies often fall short, negatively impacting performance. This fact suggests that new algorithms, hardware architectures, and data structures are required to improve the performance of the clustering process. Future research efforts should explore sophisticated techniques tailored to the distinctive attributes of complex data types, such as network and graph data, which require focused attention and innovation in data mining.
- Future work should involve collaborations across disciplines that converge data mining with areas such as artificial intelligence, privacy-preserving techniques, and advanced data visualization. These collaborations aim to expand the range of data mining applications while addressing the ongoing challenges in data mining's evolving landscape.

8. Conclusion

This paper provides a comprehensive and up-to-date overview of data mining clustering techniques, including their challenges and various applications, making it a valuable resource for researchers and practitioners. It explores various clustering methodologies, highlighting their strengths and weaknesses, explicitly emphasizing distance-based, hierarchical, grid-based, and density-based methods. This paper covers the latest trends and developments in clustering algorithms, focusing on deep, subspace, and hybrid clustering.

This study analyzes several well-known clustering algorithms, including K-means, K-medoids, AGNES, DIANA, DBSCAN, STING, and BFGSA. It also extensively explores the wide range of applications of clustering algorithms in areas such as health care, air pollution analysis, image watermarking, text document clustering, and big data analysis.

The contribution of this paper extends to the existing body of literature on data mining clustering techniques. It reviews the existing landscape and identifies open challenges and opportunities for future research. The paper details many plausible approaches to address the challenges of dealing with mixed and categorical data, working with high-dimensional and complicated datasets, creating scalable algorithms, and ensuring accurate and reliable clustering results. It provides a broad overview for researchers interested in data clustering techniques with practical and theoretical information applications as well as presenting possible future research directions.

References

Abernathy, A., & Celebi, M. E. (2022). The incremental online k-means clustering algorithm and its application to color quantization. *Expert Systems with Applications*, *207*, 117927.

Açmalı, Ş. S., & Ortakcı, Y. (2021). Clustering Performance Analysis of Traditional and New-Generation Meta-Heuristic Algorithms. *Manchester Journal of Artificial Intelligence and Applied Sciences*, *2*(2).

Ahmed, N., Barczak, A. L. C., Susnjak, T., & Rashid, M. A. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench. *Journal of Big Data*, *7*(1), 1–18.

Ahmed, S. R. A., Al Barazanchi, I., Jaaz, Z. A., & Abdulshaheed, H. R. (2019). Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set. *Periodicals of Engineering and Natural Sciences*, *7*(2), 448–457.

Alasalı, T., & Dakkak, O. (2023). Explorıng The Landscape Of Sdn-Based Ddos Defense: A Holıstıc Examınatıon Of Detectıon And Mıtıgatıon Approaches, Research Gaps And Promısıng Avenues For Future Exploratıon. *International Journal of Advanced Natural Sciences and Engineering Researches*, *7*(4), 327–349.

Ali, H. H., & Kadhum, L. E. (2017). K-means clustering algorithm applications in data mining and pattern recognition. *International Journal of Science and Research (IJSR)*, *6*(8), 1577–1584.

Alomari, H. W., Al-Badarneh, A. F., Al-Alaj, A., & Khamaiseh, S. Y. (2023). Enhanced Approach for Agglomerative Clustering Using Topological Relations. *IEEE Access*, *11*, 21945–21967.

Ambikesh, G., Rao, S. S., & Chandrasekaran, K. (2023). A grasshopper optimization algorithm-based movie recommender system. *Multimedia Tools and Applications*, 1–22.

Amirizadeh, E., & Boostani, R. (2021). CDEC: a constrained deep embedded clustering. *International Journal of Intelligent Computing and Cybernetics*, *14*(4), 686–701.

Anam, S., Fitriah, Z., Hidayat, N., & Maulana, M. H. A. A. (2023). Classification Model for Diabetes Mellitus Diagnosis based on K-Means Clustering Algorithm Optimized with Bat Algorithm. *International Journal of Advanced Computer Science and Applications*, *14*(1).

Ayesha, S., Hanif, M. K., & Talib, R. (2020a). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58.

Ayesha, S., Hanif, M. K., & Talib, R. (2020b). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, *59*, 44–58.

Azhir, E., Navimipour, N. J., Hosseinzadeh, M., Sharifi, A., & Darwesh, A. (2021). An efficient automated incremental density-based algorithm for clustering and classification. *Future Generation Computer Systems*, *114*, 665–678.

Bahadori, S., & Charkari, N. M. (2018). Increasing Efficiency of Time Series Clustering by Dimension Reduction Techniques. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY*, *18*(5), 164–170.

Bansal, A., Sharma, M., & Goel, S. (2017). Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining. *International Journal of Computer Applications*, *157*(6), 975–8887.

Bechini, A., Marcelloni, F., & Renda, A. (2020). TSF-DBSCAN: A novel fuzzy density-based approach for clustering unbounded data streams. *IEEE Transactions on Fuzzy Systems*, *30*(3), 623–637.

Bhattacharjee, P., & Mitra, P. (2020). BISDBx: towards batch-incremental clustering for dynamic datasets using SNN-DBSCAN. *Pattern Analysis and Applications*, *23*(2), 975–1009.

CERNIAN, A., CARSTOIU, D., & OLTEANU, A. (2011). Clustering Heterogeneous Web Data using Clustering by Compression. *Cluster Validity, 13th Intl. Symp. on Symbolic and Numeric Algorithms for Scientific Computing*.

Chadebec, C., Thibeau-Sutre, E., Burgos, N., & Allassonnière, S. (2022). Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 2879–2896.

Chakraborty, S., & Das, S. (2020). Detecting meaningful clusters from high-dimensional data: A strongly consistent sparse center-based clustering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), 2894–2908.

Chakraborty, S., & Nagwani, N. K. (2014). Analysis and study of Incremental DBSCAN clustering algorithm. *ArXiv Preprint ArXiv:1406.4754*.

Chalapathi, M. M., Kumar, M. R., Sharma, N., & Shitharth, S. (2022). Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal. *Security and Communication Networks*, *2022*.

Chatterjee, S., & Das, A. (2023). An ensemble algorithm using quantum evolutionary optimization of weighted type-II fuzzy system and staged Pegasos Quantum Support Vector Classifier with multi-criteria decision making system for diagnosis and grading of breast cancer. *Soft Computing*, *27*(11), 7147–7178.

Chen, H., Cai, Y., Ji, C., Selvaraj, G., Wei, D., & Wu, H. (2023). AdaPPI: identification of novel protein functional modules via adaptive graph convolution networks in a protein–protein interaction network. *Briefings in Bioinformatics*, *24*(1), bbac523.

Chen, J., Li, D., Huang, R., Chen, Z., & Li, W. (2023). Aero-engine remaining useful life prediction method with self-adaptive multimodal data fusion and cluster-ensemble transfer regression. *Reliability Engineering & System Safety*, *234*, 109151.

Chen, M.-S., Lin, J.-Q., Li, X.-L., Liu, B.-Y., Wang, C.-D., Huang, D., & Lai, J.-H. (2022). Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, *7*(3), 225–241.

Choudhary, C., Singh, I., & Kumar, M. (2023). Community detection algorithms for recommendation systems: techniques and metrics. *Computing*, *105*(2), 417–453.

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, *57*(2), 102034.

da Silva, L. E. B., Rayapati, N., & Wunsch, D. C. (2022). iCVI-ARTMAP: Using incremental cluster validity indices and adaptive resonance theory reset mechanism to accelerate validation and achieve multiprototype unsupervised representations. *IEEE Transactions on Neural Networks and Learning Systems*.

Dakkak, O., Arif, S., & Nor, S. A. (2015). Resource allocation mechanisms in computational grid: A survey. *Asian Research Publishing Network (ARPN)*, *10*.

Dakkak, O., Fazea, Y., Nor, S. A., & Arif, S. (2021). Towards accommodating deadline driven jobs on high performance computing platforms in grid computing environment. *Journal of Computational Science*, *54*, 101439.

De Weerdt, J., Vanden Broucke, S., Vanthienen, J., & Baesens, B. (2013). Active trace clustering for improved process discovery. *IEEE Transactions on Knowledge and Data Engineering*, *25*(12), 2708–2720.

Deng, M., Liu, Q., Cheng, T., & Shi, Y. (2011). An adaptive spatial clustering algorithm based on Delaunay triangulation. *Computers, Environment and Urban Systems*, *35*(4), 320–332.

Dhas, C. S. G., Yuvaraj, N., Kousik, N. V, & Geleto, T. D. (2022). D-PPSOK clustering algorithm with data sampling for clustering big data analysis. In *System Assurances* (pp. 503–512). Elsevier.

Diallo, B., Hu, J., Li, T., Khan, G. A., Liang, X., & Zhao, Y. (2021). Deep embedding clustering based on contractive autoencoder. *Neurocomputing*, *433*, 96–107.

Duan, Y., Liu, C., Li, S., Guo, X., & Yang, C. (2023a). An automatic affinity propagation clustering based on improved equilibrium optimizer and t-SNE for high-dimensional data. *Information Sciences*, *623*, 434–454.

Duan, Y., Liu, C., Li, S., Guo, X., & Yang, C. (2023b). An automatic affinity propagation clustering based on improved equilibrium optimizer and t-SNE for high-dimensional data. *Information Sciences*, *623*, 434–454.

Elgarhy, I., Badr, M. M., Mahmoud, M., Fouda, M. M., Alsabaan, M., & Kholidy, H. A. (2023). Clustering and Ensemble Based Approach For Securing Electricity Theft Detectors Against Evasion Attacks. *IEEE Access*.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022a). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022b). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279.

Fakir, Y., & El Iklil, J. (2021). Clustering techniques for big data mining. *International Conference on Business Intelligence*, 183–200.

Faroughi, A., Boostani, R., Tajalizadeh, H., & Javidan, R. (2023). ARD-Stream: An adaptive radius density-based stream clustering. *Future Generation Computer Systems*, *149*, 416–431.

Fu, X., Yuan, Y., Qiu, H., Suo, H., Song, Y., Li, A., Zhang, Y., Xiao, C., Li, Y., & Dou, L. (2024). AGF-PPIS: A protein–protein interaction site predictor based on an attention mechanism and graph convolutional networks. *Methods*.

Gao, L., Song, J., Liu, X., Shao, J., Liu, J., & Shao, J. (2017). Learning in high-dimensional multimedia data: the state of the art. *Multimedia Systems*, *23*, 303–313.

Ghazal, T. M. (2021). Performances of K-means clustering algorithm with different distance metrics. *Intelligent Automation & Soft Computing*, *30*(2), 735–742.

Ghosal, A., Nandy, A., Das, A. K., Goswami, S., & Panday, M. (2020). A short review on different clustering techniques and their applications. *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, 69–83.

Gu, B., & Sheng, V. S. (2013). Feasibility and finite convergence analysis for accurate on-line $\nu $-Support vector machine. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(8), 1304–1315.

Guo, T., Yu, K., Aloqaily, M., & Wan, S. (2022). Constructing a prior-dependent graph for data clustering and dimension reduction in the edge of AIoT. *Future Generation Computer Systems*, *128*, 381–394.

Han, X., Quan, L., Xiong, X., Almeter, M., Xiang, J., & Lan, Y. (2017). A novel data clustering algorithm based on modified gravitational search algorithm. *Engineering Applications of Artificial Intelligence*, *61*, 1–7.

Hao, Z., Lu, Z., Li, G., Nie, F., Wang, R., & Li, X. (2023). Ensemble clustering with attentional representation. *IEEE Transactions on Knowledge and Data Engineering*.

Haris, M., Yusoff, Y., Zain, A. M., Khattak, A. S., & Hussain, S. F. (2024). Breaking down multi-view clustering: A comprehensive review of multi-view approaches for complex data structures. *Engineering Applications of Artificial Intelligence*, *132*, 107857.

Hassan, Z. F., Al-Shareefi, F., & Gheni, H. Q. (2023). A Coloured Image Watermarking Based on Genetic K-Means Clustering Methodology. *Journal of Advances in Information Technology*, *14*(2).

He, G., Jiang, W., Peng, R., Yin, M., & Han, M. (2022). Soft Subspace Based Ensemble Clustering for Multivariate Time Series Data. *IEEE Transactions on Neural Networks and Learning Systems*.

He, M., & Chen, H. (2024). Anomaly Detection in Species Distribution Patterns: A Spatio-Temporal Approach for Biodiversity Conservation. *Journal of Biobased Materials and Bioenergy*, *18*(1), 39–50.

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(2), 521–526.

Huang, Q., Gao, R., & Akhavan, H. (2023). An ensemble hierarchical clustering algorithm based on merits at cluster and partition levels. *Pattern Recognition*, *136*, 109255.

Iam-On, N., & Boongoen, T. (2015). Diversity-driven generation of link-based cluster ensemble and application to data classification. *Expert Systems with Applications*, *42*(21), 8259–8273.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023a). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023b). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323.

Jain, P. K., & Pamula, R. (2019). Two-step anomaly detection approach using clustering algorithm. *International Conference on Advanced Computing Networking and Informatics: ICANI-2018*, 513–520.

JayaLakshmi, A. N. M., & Kishore, K. V. K. (2022). Performance evaluation of DNN with other machine learning techniques in a cluster using Apache Spark and MLlib. *Journal of King Saud University-Computer and Information Sciences*, *34*(1), 1311–1319.

Jeong, S., Park, J., & Lim, S. (2023). mr2vec: Multiple role-based social network embedding. *Pattern Recognition Letters*, *176*, 140–146.

Kadiravan, G., Sujatha, P., Asvany, T., Punithavathi, R., Elhoseny, M., Pustokhina, I. V, Pustokhin, D. A., & Shankar, K. (2021). Metaheuristic Clustering Protocol for Healthcare Data Collection in Mobile Wireless Multimedia Sensor Networks. *Computers, Materials & Continua*, *66*(3).

Kannout, E., Grodzki, M., & Grzegorowski, M. (2023). Towards addressing item cold-start problem in collaborative filtering by embedding agglomerative clustering and FP-growth into the recommendation system. *Computer Science and Information Systems*, *00*, 52.

Karthikeyan, B., George, D. J., Manikandan, G., & Thomas, T. (2020). A comparative study on k-means clustering and agglomerative hierarchical clustering. *International Journal of Emerging Trends in Engineering Research*, *8*(5).

Kaya, M.-F., & Schoop, M. (2022). Analytical comparison of clustering techniques for the recognition of communication patterns. *Group Decision and Negotiation*, *31*(3), 555–589.

Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, *18*(7), 723–732.

Kim, S., Cha, J., Kim, D., & Park, E. (2023). Understanding Mental Health Issues in Different Subdomains of Social Networking Services: Computational Analysis of Text-Based Reddit Posts. *Journal of Medical Internet Research*, *25*, e49074.

Krishnaswamy, R., Subramaniam, K., Nandini, V., Vijayalakshmi, K., Kadry, S., & Nam, Y. (2023). Metaheuristic Based Clustering with Deep Learning Model for Big Data Classification. *Comput. Syst. Sci. Eng.*, *44*(1), 391–406.

Kuo, R. J., Chang, C. K., Nguyen, T. P. Q., & Liao, T. W. (2021). Application of genetic algorithm-based intuitionistic fuzzy weighted c-ordered-means algorithm to cluster analysis. *Knowledge and Information Systems*, *63*, 1935–1959.

Kuwil, F. H., Shaar, F., Topcu, A. E., & Murtagh, F. (2019). A new data clustering algorithm based on critical distance methodology. *Expert Systems with Applications*, *129*, 296–310.

lahmood HAMEED, F., & DAKKAK, O. (2022). Brain Tumor Detection and Classification Using Convolutional Neural Network (CNN). *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1–7.

Laohakiat, S., & Sa-Ing, V. (2021). An incremental density-based clustering framework using fuzzy local clustering. *Information Sciences*, *547*, 404–426.

Lee, Y., Park, C., & Kang, S. (2022). Deep Embedded Clustering Framework for Mixed Data. *IEEE Access*, *11*, 33–40.

Li, X., Chen, X., & Rezaeipanah, A. (2023). Automatic breast cancer diagnosis based on hybrid dimensionality reduction technique and ensemble classification. *Journal of Cancer Research and Clinical Oncology*, 1–19.

Liu, C., Nie, F., Wang, R., & Li, X. (2022). Scalable fuzzy clustering with anchor graph. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, H., Yang, J., Ye, M., James, S. C., Tang, Z., Dong, J., & Xing, T. (2021). Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data. *Journal of Hydrology*, *597*, 126146.

Liu, R., Ren, R., Liu, J., & Liu, J. (2020). A clustering and dimensionality reduction based evolutionary algorithm for large-scale multi-objective problems. *Applied Soft Computing*, *89*, 106120.

Lv, Y., Ma, T., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2016). An efficient and scalable density-based clustering algorithm for datasets with complex structures. *Neurocomputing*, *171*, 9–22.

Lydia, E. L., Moses, G. J., Varadarajan, V., Nonyelu, F., Maseleno, A., Perumal, E., & Shankar, K. (2020). Clustering and indexing of multiple documents using feature extraction through apache hadoop on big data. *Malaysian Journal of Computer Science*, 108–123.

Maia, J., Junior, C. A. S., Guimarães, F. G., de Castro, C. L., Lemos, A. P., Galindo, J. C. F., & Cohen, M. W. (2020). Evolving clustering algorithm based on mixture of typicalities for stream data mining. *Future Generation Computer Systems*, *106*, 672–684.

Marqués-Sánchez, P., Martínez-Fernández, M. C., Benítez-Andrades, J. A., Quiroga-Sánchez, E., García-Ordás, M. T., & Arias-Ramos, N. (2023). Adolescent relational behaviour and the obesity pandemic: A descriptive study applying social network analysis and machine learning techniques. *PloS One*, *18*(8), e0289553.

Mayanglambam, S. D., Horng, S.-J., & Pamula, R. (2023). PSO clustering and pruning-based KNN for outlier detection. *Soft Computing*, 1–17.

Mohammadi, M., Shokrollahi, A., Reisi, M., Abdollahpouri, A., & Moradi, P. (2023). *Scalable and robust big data clustering with adaptive local feature weighting based on the Map-Reduce and Hadoop*.

Mortensen, K. O., Zardbani, F., Haque, M. A., Agustsson, S. Y., Mottin, D., Hofmann, P., & Karras, P. (2023). Marigold: Efficient k-Means Clustering in High Dimensions. *Proceedings of the VLDB Endowment*, *16*(7), 1740–1748.

Mrukwa, G., & Polanska, J. (2022). DiviK: divisive intelligent K-means for hands-free unsupervised clustering in big biological data. *BMC Bioinformatics*, *23*(1), 1–24.

Mussabayev, R., Mladenovic, N., Jarboui, B., & Mussabayev, R. (2023). How to use K-means for big data clustering? *Pattern Recognition*, *137*, 109269.

Nie, X., Qin, D., Zhou, X., Duo, H., Hao, Y., Li, B., & Liang, G. (2023). Clustering ensemble in scRNA-seq data analysis: Methods, applications and challenges. *Computers in Biology and Medicine*, 106939.

Nozari, H., & Sadeghi, M. E. (2021). Artificial intelligence and Machine Learning for Real-world problems (A survey). *International Journal of Innovation in Engineering*, *1*(3), 38–47.

Ollagnier, A., Cabrio, E., & Villata, S. (2023). Unsupervised fine-grained hate speech target community detection and characterisation on social media. *Social Network Analysis and Mining*, *13*(1), 58.

Omar, N., Nazirun, N. N., Vijayam, B., Wahab, A. A., & Bahuri, H. A. (2023). Diabetes subtypes classification for personalized health care: A review. *Artificial Intelligence Review*, *56*(3), 2697–2721.

Ortakci, Y. (2017). Parallel particle swarm optimization in data clustering. *Int. J Soft Comput. Artif. Intell.(IJSCAI)*, *5*(1), 10–14.

Oskouei, A. G., Balafar, M. A., & Motamed, C. (2021). FKMAWCW: categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning. *Chaos, Solitons & Fractals*, *153*, 111494.

Pareek, J., & Jacob, J. (2021). Data compression and visualization using PCA and T-SNE. *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019*, 327–337.

Patel, D., Modi, R., & Sarvakar, K. (2014). A comparative study of clustering data mining: Techniques and research challenges. *International Journal of Latest Technology in Engineering, Management & Applied Science*, *3*(9), 67–70.

Pérez-Ortega, J., Rey-Figueroa, C. D., Roblero-Aguilar, S. S., Almanza-Ortega, N. N., Zavala-Díaz, C., García-Paredes, S., & Landero-Nájera, V. (2023). POFCM: A Parallel Fuzzy Clustering Algorithm for Large Datasets. *Mathematics*, *11*(8), 1920.

Pham, N. D., Le, T. D., Park, K., & Choo, H. (2010). SCCS: Spatiotemporal clustering and compressing schemes for efficient data collection applications in WSNs. *International Journal of Communication Systems*, *23*(11), 1311–1333.

Phan, H. T., & Nguyen, N. T. (2024). A Fuzzy Graph Convolutional Network Model for Sentence-Level Sentiment Analysis. *IEEE Transactions on Fuzzy Systems*.

Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Aspect-level sentiment analysis: A survey of graph convolutional network methods. *Information Fusion*, *91*, 149–172.

Price, M. A., McEwen, J. D., Cai, X., Kitching, T. D., Wallis, C. G. R., & Collaboration), L. D. E. S. (2021). Sparse Bayesian mass mapping with uncertainties: hypothesis testing of structure. *Monthly Notices of the Royal Astronomical Society*, *506*(3), 3678–3690.

Purwandari, K., Sigalingging, J. W. C., Fhadli, M., Arizky, S. N., & Pardamean, B. (2020). Data mining for predicting customer satisfaction using clustering techniques. *2020 International Conference on Information Management and Technology (ICIMTech)*, 223–227.

Qoku, A., & Buettner, F. (2023). Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity. *International Conference on Artificial Intelligence and Statistics*, 11545–11562.

Qu, W., Xiu, X., Chen, H., & Kong, L. (2023). A Survey on High-Dimensional Subspace Clustering. *Mathematics*, *11*(2), 436.

Rahayu, K., Novianti, L., & Kusnandar, M. (2020). Implementation data mining with K-Means algorithm for clustering distribution rabies case area in Palembang City. *Journal of Physics: Conference Series*, *1500*(1), 012121.

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, *56*(8), 8219–8264.

Ray, P., Reddy, S. S., & Banerjee, T. (2021). Various dimension reduction techniques for high dimensional data analysis: a review. *Artificial Intelligence Review*, *54*, 3473–3515.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *Ieee Access*, *8*, 54776–54788.

Rehman, M. U., & Khan, D. M. (2021). A novel density-based technique for outlier detection of high dimensional data utilizing full feature space. *Information Technology and Control*, *50*(1), 138–152.

Richards, J. A., & Richards, J. A. (2022). *Remote sensing digital image analysis* (Vol. 5). Springer.

Rubarth, K., Sattler, P., Zimmermann, H. G., & Konietschke, F. (2021). Estimation and testing of Wilcoxon–Mann–Whitney effects in factorial clustered data designs. *Symmetry*, *14*(2), 244.

Sabitha, A. S., & Bansal, A. (2017). Climate change analysis to study land surface temparature trends. *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, 1–8.

Sahoo, S. K., Pattanaik, P., Mohanty, M. N., & Mishra, D. K. (2023). Opposition Learning Based Improved Bee Colony Optimization (OLIBCO) Algorithm for Data Clustering. *International Journal of Advanced Computer Science and Applications*, *14*(4).

Saklani, R., Purohit, K., Vats, S., Sharma, V., Kukreja, V., & Yadav, S. P. (2023). Multicore Implementation of K-Means Clustering Algorithm. *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 171–175.

Samoilenko, S., & Osei-Bryson, K.-M. (2019). Representation matters: An exploration of the socio-economic impacts of ICT-enabled public value in the context of sub-Saharan economies. *International Journal of Information Management*, *49*, 69–85.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017a). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017b). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681.

Shah, N. H., Priamvada, A., & Shukla, B. P. (2023). Decoding spatial precipitation patterns using artificial intelligence. *Spatial Information Research*, 1–12.

Sharma, S., Agrawal, J., Agarwal, S., & Sharma, S. (2013). Machine learning techniques for data mining: A survey. *2013 IEEE International Conference on Computational Intelligence and Computing Research*, 1–6.

Sheng, G., Wang, Q., Pei, C., & Gao, Q. (2022). Contrastive deep embedded clustering. *Neurocomputing*, *514*, 13–20.

Shi, Y., Yang, K., Yu, Z., Chen, C. L. P., & Zeng, H. (2023). Adaptive Ensemble Clustering With Boosting BLS-Based Autoencoder. *IEEE Transactions on Knowledge and Data Engineering*.

Shrifan, N. H. M. M., Akbar, M. F., & Isa, N. A. M. (2022). An adaptive outlier removal aided k-means clustering algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 6365–6376.

Sinaga, K. P., Hussain, I., & Yang, M.-S. (2021). Entropy K-means clustering with feature reduction under unknown number of clusters. *IEEE Access*, *9*, 67736–67751.

Souiden, I., Omri, M. N., & Brahmi, Z. (2022). A survey of outlier detection in high dimensional data streams. *Computer Science Review*, *44*, 100463.

Sun, L., Zhang, J., Ding, W., & Xu, J. (2022). Feature reduction for imbalanced data classification using similarity-based feature clustering with adaptive weighted K-nearest neighbors. *Information Sciences*, *593*, 591–613.

Tejasree, S., & Chandra Mohan, B. (2023). An improved differential bond energy algorithm with fuzzy merging method to improve the document clustering for information mining. *Expert Systems*, e13261.

Thrun, M. C., & Ultsch, A. (2021). Using projection-based clustering to find distance-and density-based clusters in high-dimensional data. *Journal of Classification*, *38*, 280–312.

Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, *7*, 1–30.

Tiwari, A. (2021). Enhancing k-means algorithm clustering performance with improved time complexity. *National Conference on "Unprecedented and Advanced Concepts of Computer Vision" NCUACC*, *11*(12).

Ukey, N., Yang, Z., Li, B., Zhang, G., Hu, Y., & Zhang, W. (2023). Survey on exact knn queries over high-dimensional data space. *Sensors*, *23*(2), 629.

Utku, A., Can, U., & Aslan, S. (2023). Detection of hateful twitter users with graph convolutional network model. *Earth Science Informatics*, *16*(1), 329–343.

Vandhana, S., & Anuradha, J. (2021). Environmental air pollution clustering using enhanced ensemble clustering methodology. *Environmental Science and Pollution Research*, *28*, 40746–40755.

Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T., & Han, J. (2013). A phrase mining framework for recursive construction of a topical hierarchy. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 437–445.

Wang, F., Zheng, Z., Zhang, Y., Li, Y., Yang, K., & Zhu, C. (2023). To see further: Knowledge graph-aware deep graph convolutional network for recommender systems. *Information Sciences*, *647*, 119465.

Wang, L., Wang, Y., Deng, H., & Chen, H. (2023). Attention reweighted sparse subspace clustering. *Pattern Recognition*, *139*, 109438.

Wickramasinghe, C. S., Marino, D. L., & Manic, M. (2021). ResNet autoencoders for unsupervised feature learning from high-dimensional data: Deep models resistant to performance degradation. *IEEE Access*, *9*, 40511–40520.

Wright, J., & Ma, Y. (2022). *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press.

Xie, J., Xu, X., Lan, Y., Shi, X., Yong, Y., & Wu, D. (2023). Automatic velocity picking with restricted weighted k-means clustering using prior information. *Frontiers in Earth Science*, *10*, 1076999.

Xie, W.-B., Lee, Y.-L., Wang, C., Chen, D.-B., & Zhou, T. (2020). Hierarchical clustering supported by reciprocal nearest neighbors. *Information Sciences*, *527*, 279–292.

Xie, Z., Nie, M., & Wang, T. (2009). Clustering Based Compress Data Cube Algorithm. *2009 WRI World Congress on Software Engineering*, *4*, 429–433.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165–193.

Yedla, M., Pathakota, S. R., & Srinivasa, T. M. (2010). Enhancing K-means clustering algorithm with improved initial center. *International Journal of Computer Science and Information Technologies*, *1*(2), 121–125.

Yu, T.-T., Chen, C.-Y., Wu, T.-H., & Chang, Y.-C. (2023). Application of high-dimensional uniform manifold approximation and projection (UMAP) to cluster existing landfills on the basis of geographical and environmental features. *Science of The Total Environment*, *904*, 167013.

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J*, *2*(2), 226–235.

Yue, G., Deng, A., Qu, Y., Cui, H., & Liu, J. (n.d.). Fuzzy-Rough induced spectral ensemble clustering. *Journal of Intelligent & Fuzzy Systems*, *Preprint*, 1–18.

Zhong, L., Yang, J., Chen, Z., & Wang, S. (2023). Contrastive Graph Convolutional Networks With Generative Adjacency Matrix. *IEEE Transactions on Signal Processing*, *71*, 772–785.