Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i2.2473

A Survey of Clustering Algorithms for High-Dimensional Data Mining

Anil Kukreti
Faculty, School of Computing, Graphic Era Hill University,
Dehradun, Uttarakhand India 248002

Article Info
Page Number:1800 - 1810
Publication Issue:
Vol 70 No. 2 (2021)

Abstract

The increasing complexity and dimensionality of data in numerous domains, including bioinformatics, text mining, multimedia, and social network analysis, pose significant challenges to traditional clustering algorithms. This paper provides a comprehensive review of various clustering algorithms suitable for high-dimensional data mining. We begin by elaborating on the challenges and difficulties intrinsic to high-dimensional spaces, such as the curse of dimensionality and the concentration of measure. Following that, we investigate a variety of different approaches to clustering, such as partitioning techniques, grid-based methods, hierarchical methods, density-based methods, and model-based methods. These algorithms are examined with a particular emphasis on how well they deal with high-dimensional data and how well they deal with data noise, as well as how well they can scale and how easily they can be interpreted. The most current developments in subspace and correlation clustering, as well as embedding approaches and the application of deep learning to the clustering of high-dimensional data, are also included in this overview. With the help of this in-depth analysis, our goals are to give insights into the benefits and drawbacks of each algorithm and to aid academics and practitioners in picking the optimal approach for the high-dimensional data mining projects they are working on.

Article History
Article Received*: 05 September 2021*
Revised*: 09 October 2021*
Accepted*: 22 November 2021*
Publication*: 26 December 2021*

1. Introduction

As we step further into the data-driven age, the volume, velocity, and variety of data continue to grow exponentially, ushering in opportunities and challenges in equal measure. Among the more profound challenges is the increasing dimensionality of the datasets, a phenomenon that has elevated the field of high-dimensional data mining into a critical research frontier.

Historically, data mining emerged from the intersection of machine learning, statistics, and databases in the 1960s. The initial focus was on relatively low-dimensional data, with simple linear relationships. However, the advent of the internet, social networks, multimedia content, biotechnology, and other advanced technologies in the late 20th and early 21st centuries has radically transformed the data landscape. Today's datasets not only contain millions of data points but each data point can have hundreds or thousands of features or dimensions, leading to the so-called "high-dimensional data."

High-dimensional data mining is extensively applied across various domains. In bioinformatics, it's used to uncover genetic patterns and interactions across thousands of genes and genetic markers. In text and web mining, it helps understand high-dimensional representations of text data, including word frequencies, semantics, and more. Multimedia applications often involve processing high-dimensional data, such as image pixels or audio spectrograms. In the field of finance, high-dimensional data mining can uncover complex relationships between different financial indicators. The foundation of data mining lies in discovering patterns within these vast, high-dimensional datasets, and this is where clustering algorithms come into play. Clustering is a fundamental data mining task and a form of unsupervised learning. It groups similar instances based on certain similarity or distance measures. Traditional clustering algorithms, such as K-means, hierarchical clustering, and DBSCAN, were not originally designed for high-dimensional data, and they often falter when applied to such datasets due to the "curse of dimensionality."

To tackle high-dimensional data, numerous advanced clustering techniques have been proposed, including subspace clustering, correlation clustering, spectral clustering, and more recently, clustering algorithms based on deep learning. Each algorithm has its strengths and weaknesses, depending on factors such as the data's structure, noise level, and the application at hand.

This paper provides a comprehensive survey of these clustering algorithms, shedding light on their applicability and efficacy in high-dimensional data mining, and ultimately assisting researchers and practitioners in navigating this complex landscape.

2. Literature Review

The investigation of clustering algorithms in the literature has been extensive and wide-ranging, focusing on different methods and their application in various contexts. Fahad et al. (2014) conducted a comprehensive survey on clustering algorithms designed specifically for big data. The paper provides a useful taxonomy of algorithms and also presents an empirical analysis that compares their performance on large datasets. This paper is a valuable resource for understanding the state of the art in big data clustering.

Carrizosa et al. (2013) approached clustering from a network perspective. The authors proposed a variable neighbourhood search algorithm for minimum sum-of-squares clustering on networks. This work demonstrates the potential of metaheuristic search strategies in addressing complex clustering problems.

In an industrial context, Benabdellah et al. (2019) offered a survey of clustering algorithms, emphasizing the practical aspects of clustering such as scalability, robustness, and ease of interpretation. The paper also discussed various cluster validation techniques, which are critical for assessing the quality of clustering results.

A different perspective was presented by Anter et al. (2019), where the authors combined a fast fuzzy c-means algorithm with a crow search optimization algorithm for crop identification in agriculture. This work underscores the potential of hybrid approaches that leverage the strengths of multiple algorithms.

Cura (2012) introduced a particle swarm optimization (PSO) approach to clustering. PSO is a bio-inspired optimization technique that mimics the social behavior of bird flocks. The author's approach demonstrated that PSO can effectively navigate the solution space of clustering problems and find high-quality solutions.

Bandaru et al. (2017) provided a survey of data mining methods for knowledge discovery in multi-objective optimization. Although not focused on clustering per se, this paper offers valuable insights into how data mining techniques, including clustering, can be used to analyze the results of multi-objective optimization problems.

The application of clustering in traffic flow forecasting was explored by Gavali et al. (2014). They used the MapReduce framework for distributed processing of large traffic data, showing the effectiveness of data-driven clustering methods in predicting traffic flow patterns.

Mane et al. (2020) introduced the Modified Quick Fuzzy Hypersphere Neural Network (MQFHSNN) for pattern classification using supervised clustering. Their method was shown to be effective in various applications, suggesting that neural network-based clustering algorithms hold great potential for high-dimensional data.

Lastly, Kosmidis and Karlis (2016) proposed a model-based clustering approach using copulas. Their method can handle various data types and dependencies, and it illustrates the flexibility and power of model-based clustering.

The literature covers a wide range of clustering algorithms and their applications, demonstrating the diversity and adaptability of clustering methods in dealing with complex, high-dimensional data.

3. Clustering Methodologies

Clustering algorithms can be broadly categorized into several classes based on their underlying principles and the type of clusters they can form. Below, we provide an overview of this taxonomy and a brief description of each category of clustering algorithm.

1. Partitioning Methods:

These algorithms partition the data into a set of k clusters, such that each cluster contains at least one data point and each data point belongs to exactly one cluster. The most representative algorithms in this category are K-means and K-medoids.

a. K-means: In order to organise the data into clusters, this technique begins with k initial centroids and then allocates each data point to the centroid that is geographically closest to it. It then continues the procedure until the centroids no longer shift or until a maximum number of iterations has been achieved, whichever comes first. The updated centroids are based on the existing cluster assignments.

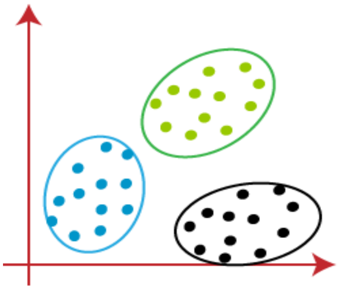

Figure 1. K-Means Clustering

b. K-medoids: This approach is a modification of K-means that employs real data points as cluster centres (medoids) rather than centroids. As a result, it is more resistant to noise and outliers than the original.

2. Hierarchical Methods:

The creation of a tree-like representation of the data using hierarchical clustering algorithms makes it possible to investigate the clusters on a variety of various levels of granularity. These may be agglomerative (coming from the bottom up) or divisive (coming from the top down).

a. Agglomerative Clustering: This approach begins with each data point being treated as its own cluster. Subsequently, the procedure combines the clusters that are geographically nearest to one another until there is only one cluster left, or a certain number of clusters.

b. Divisive Clustering: This method starts with all data points in a single cluster and successively splits the most heterogeneous cluster until each data point is in its own cluster (or a predefined number of clusters is reached).

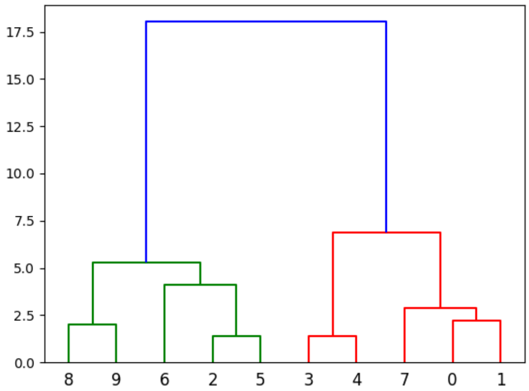

Figure 2. Hierarchical Clustering

3. Density-Based Methods:

Density-based clustering algorithms group together data points in high-density regions and separate data points in low-density regions. They are particularly useful for discovering clusters of arbitrary shape and for handling noise and outliers. The most well-known algorithm in this category is DBSCAN.

a. DBSCAN: Outliers are data points that are isolated in low-density zones and are labelled as such by DBSCAN. Data points that are clustered tightly together are grouped together by DBSCAN.

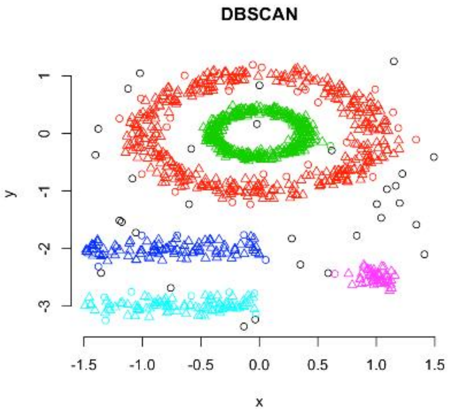

Figure 3. Density-Based Clustering

4. Grid-Based Methods:

Grid-based clustering algorithms divide the data space into a finite number of cells forming a grid structure, and then perform all clustering operations on this grid structure, typically resulting in faster processing times. The representative algorithm of this category is STING.

a. STING (STatistical INformation Grid) STING divides the spatial area into rectangular cells with varying sizes and stores statistical information about the data points in each cell. Clustering is then performed based on these statistical values.

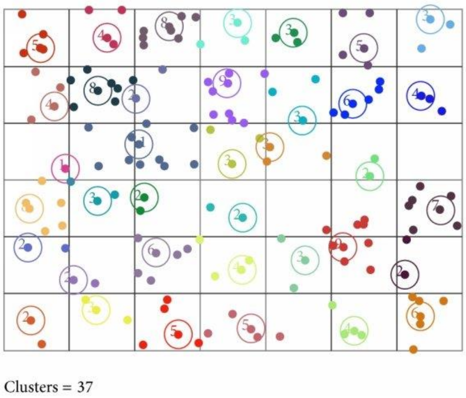

Figure 4. Grid-Based Clustering

5. Model-Based Methods:

Model-based clustering algorithms begin with the presumption that each of the clusters has a corresponding model and then search for the data that provides the best possible match to the models. The Expectation-Maximization (EM) clustering method, which makes use of Gaussian Mixture Models, is one of the most well-known algorithms in this area.

a. Gaussian Mixture Models (GMM): The GMM is a probabilistic model that makes the assumption that all of the data points are produced from a mixture of a limited number of Gaussian distributions with unknown parameters. This makes the GMM a mixture model. The EM method is used in order to learn the GMM's parameter settings.

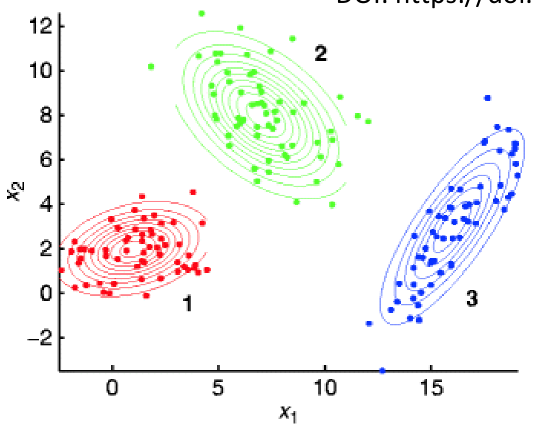

Figure 5. Model-Based Clustering

6. Subspace and Correlation Clustering:

These methods are more recent additions to the clustering taxonomy, particularly developed for high-dimensional data. They focus on finding clusters in subspaces of the data, i.e., they look for clusters that exist in some dimensions (subspaces) of the data.

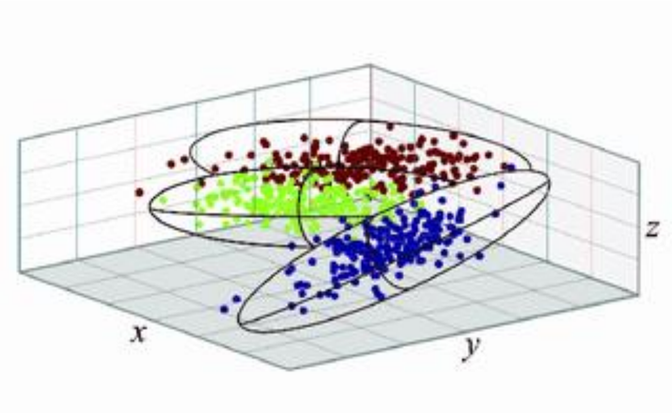

Figure 6. Subspace and Correlation Clustering

7. Spectral Clustering:

In order to conduct dimensionality reduction before clustering in less dimensions, spectral clustering approaches make use of the spectrum (eigenvalues) of the similarity matrix of the data. This is done before clustering in fewer dimensions. This approach of clustering is quite effective.

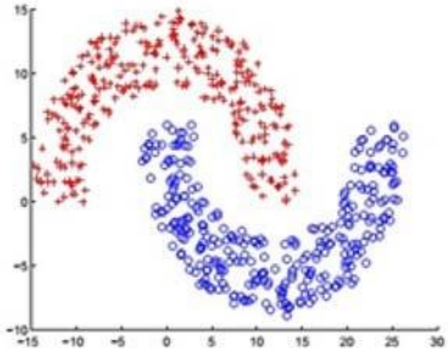

Figure 7. Spectral Clustering

4. Challenges In High Dimensional Clustering

High-dimensional data mining, especially with the application of clustering algorithms, poses a variety of challenges. These challenges largely arise from the inherent properties of high-dimensional spaces and the complexities of real-world data. Some of the most prominent challenges include:

1. Curse of Dimensionality: This word relates to high-dimensional data analysis and organisation phenomena that do not exist in low-dimensional ones. As dimensionality grows, space volume increases rapidly, sparsing data. This sparsity makes statistical significance difficult. A statistically valid and trustworthy conclusion requires exponentially more data with dimensionality. Dimensions also make data organisation and search harder.

2. Data Noise and Outliers: Real-world data is often noisy and contains outliers that deviate significantly from the rest of the data. Clustering high-dimensional data is particularly susceptible to noise and outliers because they can disproportionately affect the distance measures used to group similar instances, leading to erroneous cluster assignments.

3. Scalability: With the ever-increasing size of datasets, scalability remains a major challenge in high-dimensional data mining. Not only must a clustering algorithm handle data with many features, but it must also efficiently process a large number of instances. Many existing algorithms suffer from poor performance as the size of the dataset grows.

4. Interpretability: High-dimensional data often comes with complexities that make it hard to interpret the results of clustering. For instance, it can be difficult to visualize clusters or understand what makes one cluster different from another in high-dimensional space. This poses a challenge in applying these algorithms in practice, where interpretability is crucial for decision-making.

5. Feature Correlation: In high-dimensional data, one often finds that many features are correlated. This redundancy can lead to a dilution of relevant information and therefore impact the effectiveness of clustering.

6. Model Complexity and Overfitting: As we build more complex models to tackle high-dimensional data, there's an increased risk of overfitting the model to the training data, thereby failing to generalize well to new data.

7. Computational Costs: Higher-dimensional data significantly increases computational costs, making some algorithms impractical for real-world, time-sensitive applications.

Addressing these challenges requires the development of more robust, scalable, and interpretable clustering algorithms, as well as strategies for handling noise, outliers, and feature correlations in high-dimensional data.

5. Application Of Machine Learning And Deep Learning In Clustering High-Dimensional Data

Machine Learning and Deep Learning have made significant strides in dealing with the challenges of high-dimensional data, including its complexity, noise, and non-linearity. They have been instrumental in developing robust, scalable, and efficient methods for clustering such high-dimensional datasets. Here are some notable applications:

1. Dimensionality Reduction:

Techniques for reducing the dimensionality of data are frequently used as a preprocessing step in order to minimise the dimensionality of the data and alleviate the problems caused by the curse of dimensionality. The Principal Component Analysis (PCA) is a tried-and-true approach for accomplishing this. However, non-linear methods that are based on machine learning, such as t-Distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), have also garnered a lot of attention in recent years. Because autoencoders, a form of deep neural network, are able to generate an unsupervised lower-dimensional representation of the input data, they are beneficial for dimensionality reduction in high-dimensional data clustering.

2. Deep Clustering:

Deep clustering methods integrate representation learning and clustering into a single model. They leverage deep neural networks to learn a compact representation of the data, with the aim of enhancing cluster separability in the learned low-dimensional space. Algorithms such as Deep Embedded Clustering (DEC) and Deep Clustering Network (DCN) fall under this category. These models have demonstrated superior performance on complex datasets compared to traditional clustering methods.

3. Convolutional Neural Networks (CNNs) for Image Clustering:

CNNs have shown remarkable results in image classification tasks due to their ability to capture spatial hierarchies of features. This ability can be extended to clustering high-dimensional image data. The approach typically involves using a pre-trained CNN to extract features from images, followed by clustering on these features.

4. Graph Neural Networks (GNNs) for Graph Clustering:

Graphs are a common type of high-dimensional data, where each node and edge can have multiple attributes. GNNs have emerged as powerful tools for learning on graph-structured data, including for clustering tasks. They are capable of capturing complex patterns in the graph structure and node attributes, leading to more accurate clustering results.

5. Reinforcement Learning for Clustering:

Reinforcement learning, particularly deep reinforcement learning, has been employed in clustering high-dimensional data. The approach involves framing the clustering task as a sequential decision-making problem, where an agent learns to assign data points to clusters so as to maximize a long-term reward.

6. Self-Supervised Learning for Clustering:

In self-supervised learning, auxiliary tasks are designed to exploit the structure of the data and learn powerful data representations without the need for explicit labels. These learned representations can then be used for clustering. This approach is particularly useful in scenarios where labelled data is scarce or expensive to obtain.

Despite these advancements, the field of clustering high-dimensional data with machine learning and deep learning is still ripe for exploration, with many challenges to be overcome and opportunities to be exploited.

6. Result And Discussion

Table 1 shows the overview of clustering techniques used by researchers, their advantages and disadvantages.

Table 1. Overview of clustering techniques

Author(s) and Year	Methodology Used	Algorithm Used	Advantages	Disadvantages
Fahad et al. (2014)	Survey and empirical analysis	Various big data clustering algorithms	Comprehensive, empirical comparison	Limited to big data scenarios
Carrizosa et al. (2013)	Metaheuristic search strategy	Variable neighborhood search	Robust to network structures	May not find global optimum
Benabdellah et al. (2019)	Survey	Various industrial clustering algorithms	Practical considerations discussed	Limited to industrial applications
Anter et al. (2019)	Optimization with fuzzy logic	Fast fuzzy c-means, Crow search	Resilient to noise, effective in agriculture	Specific to certain applications
Cura (2012)	Bio-inspired optimization	Particle swarm optimization	Efficient navigation of solution space	Can be trapped in local optima
Bandaru et al. (2017)	Survey	Various data mining methods	Detailed overview of methods	Not specific to clustering
Gavali et al. (2014)	Data-driven clustering	MapReduce-based clustering	Effective for large traffic data	Restricted to traffic data
Mane et al. (2020)	Neural network-based clustering	Modified Quick Fuzzy Hypersphere Neural Network	Powerful for pattern classification	Complexity of neural network models
Kosmidis & Karlis (2016)	Model-based clustering	Copula-based clustering	Handles various data types and dependencies	Complex and computationally intensive

The examined literature presents a diverse landscape of clustering methodologies and algorithms. It starts with a comprehensive survey and empirical comparison of various big data clustering algorithms. One study employs a metaheuristic search strategy that proves effective in network structures, although it may occasionally miss the global optimum.

Moving further, another study focuses on the practical considerations of clustering algorithms in an industrial context. There's also fascinating work on combining fuzzy logic with optimization techniques to improve the resilience to noise and effectiveness in agricultural applications.

One interesting approach uses a bio-inspired optimization method known as particle swarm optimization, which has demonstrated efficiency in navigating the solution space of clustering problems. Despite its effectiveness, there's a potential risk of this algorithm getting trapped in local optima.

Another noteworthy contribution is a survey of various data mining methods, providing a detailed overview of the methods. Still, it's not specifically focused on clustering. A unique data-driven clustering approach was effective for large traffic data, with its application domain primarily being traffic flow data.

The literature also showcases powerful neural network-based clustering algorithms, particularly effective for pattern classification. However, the complexity of neural network models is an undeniable challenge. Finally, model-based clustering approaches that can handle various data types and dependencies have also been explored. Although powerful, these methods can be complex and computationally intensive.

7. Conclusion

This paper offers an extensive review of clustering algorithms suitable for high-dimensional data mining. It brings to light the unique challenges inherent in high-dimensional spaces and highlights diverse methodologies that have been developed to tackle these issues. From partitioning to hierarchical methods, and density-based to model-based methods, each approach has been thoroughly evaluated on its merit in dealing with high-dimensional data. Furthermore, we delve into recent advancements in subspace and correlation clustering, embedding methods, and deep learning applications. This survey serves as a comprehensive guide, aiming to aid researchers and practitioners in the selection of the most appropriate clustering technique for their specific high-dimensional data mining tasks. By drawing attention to each algorithm's strengths and weaknesses, we hope to inspire further research and innovation in this critical area of data analysis.

Refrences

[1]. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A., ... & 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.

[2]. Carrizosa, E., Martín-Barragán, B., & Romero Morales, D. (2013). Variable neighborhood search for minimum sum-of-squares clustering on networks. European Journal of Operational Research, 230(1), 142-154.

[3]. Benabdellah, A.C., Ouardighi, F. E. & Yassa, S. (2019). A survey of clustering algorithms for an industrial context. Procedia Computer Science, 151, 1132-1139.

[4]. Anter, A. M., Hassanien, A. E., & Elhoseny, M. (2019). An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural. Expert Systems with Applications, 126, 245-258.

[5]. Cura, T. (2012). A particle swarm optimization approach to clustering. Expert Systems with Applications, 39(1), 1582-1585.

[6]. Bandaru, S., Ng, A. H. C., & Deb, K. (2017). Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey. Expert Systems with Applications, 70, 139-159.

[7]. Gavali, V. N., Mane, P., & Patil, R. R. (2014). Data driven traffic flow forecasting using MapReduce in distributed modelling. Networks, 1(2).

[8]. Mane, D.T., Kshirsagar, J.P., Kulkarni, U.V. (2020). Modified Quick Fuzzy Hypersphere Neural Network for Pattern Classification Using Supervised Clustering. In: Reddy, V., Prasad,

V., Wang, J., Reddy, K. (eds) Soft Computing and Signal Processing. ICSCSP 2019. Advances in Intelligent Systems and Computing, vol 1118. Springer, Singapore.

[9]. Kosmidis, I., & Karlis, D. (2016). Model-based clustering using copulas with applications. Statistical Computing, 26(5), 1079-1099.

[10]. Tejaswini Mallavarapu, Jie Hao, Youngsoon Kim, Jung Hun Oh, Mingon Kang, Pathway-based deep clustering for molecular subtyping of cancer, Methods, Volume 173, 2020, Pages 24-31, ISSN 1046-2023, https://doi.org/ 10.1016/j.ymeth.2019.06.017.