International Journal of Engineering & Scientific Research
Vol. 5 Issue 7, July 2017,
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://www.esrjournal.com Email: esrjeditor@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A Survey of Clustering Techniques: Foundations and Practical Applications

Jahnvi V Doshi, Assistant Professor in Computer Engineering, Government Engineering College - Rajkot

Komal K Shah, Assistant Professor in Computer Engineering, Government Engineering College - Rajkot

Abstract

Clustering is a fundamental method in unsupervised learning for finding innate groupings in datasets. This survey examines various clustering methods like Partitioning, hierarchical, density-based, grid-based and model-based clustering. It also shows the comparative analysis of these methods with their advantages and disadvantages. Additionally, it also discusses the clustering algorithms and real-world uses in various domains. All clustering methods aim to minimize inter-cluster similarity and maximize intra-cluster similarity, even though their variations in constructing approach. The paper concludes with current challenges in clustering and suggests potential future research directions. It also emphasizes the need for scalable, interpretable and adaptive clustering techniques.

1. Introduction

Clustering is one of the most fundamental tasks in unsupervised learning. It involves organizing data into groups or clusters [1] such that items within the same group are more similar to each other than to those in different groups. It plays a pivotal role in fields [2] ranging from bioinformatics and image analysis to market research and social network analysis.

Unlike supervised learning, clustering does not rely on pre-labeled [2] training data. Instead, it uncovers hidden patterns or intrinsic groupings in data. Various clustering methods have been developed, each suitable for different data distributions, dimensionality, scalability and application domains.

This survey explores clustering methods and algorithms, discusses their real-world applications and presents a comparative analysis to evaluate their strengths and limitations.

2. Clustering Methods and Algorithms

Clustering methods [3] can be broadly categorized based on the strategies they use to detect and form groups within data. These strategies influence how clusters are shaped, the scalability of the algorithm, its sensitivity to noise and its ability to deal with complex or high-dimensional data. The five main categories are partitioning, hierarchical, density-based, grid-based and model-based [4] clustering methods. Each offers a distinct perspective on grouping similar data points and the appropriate method is chosen depending on the structure and nature of the data as well as the analytical goals.

2.1 Partitioning Methods

Partitioning methods [5] are the most widely used in clustering. These techniques divide a dataset into a predetermined number of non-overlapping clusters. The goal is to assign each data point to exactly one cluster in such a way that intra-cluster similarity is maximized while inter-cluster similarity is minimized [5].

One of the most popular algorithms in this category is K-Means. The K-Means algorithm [6] begins by selecting k initial cluster centroids, often at random, where k is a user-defined number. Each data point is then assigned to the nearest centroid based on a distance metric, typically Euclidean distance. The centroids are updated iteratively by computing the mean of the points within each cluster until convergence is achieved. Although K-Means is simple and computationally efficient, it has some limitations [6], such as sensitivity to the initial choice of centroids, difficulty in identifying non-spherical clusters and poor handling of noise and outliers.

To address these limitations, the K-Medoids algorithm [7] , also known as PAM - Partitioning Around Medoids, was introduced. Unlike K-Means, K-Medoids selects actual data points as cluster centres known as medoids, which makes the algorithm more robust to outliers and suitable for datasets with arbitrary distance measures. However, it is computationally more intensive than K-Means.

For large datasets, algorithms such as CLARA (Clustering Large Applications) and CLARANS (Clustering Large Applications based on RANdomized Search) [8] offer scalable solutions. CLARA improves efficiency by running the PAM algorithm on multiple small samples of the dataset, while CLARANS uses a randomized approach to search the space of medoid configurations more effectively. These modifications allow partitioning techniques to maintain clustering quality while managing bigger and more complicated datasets.

2.2 Hierarchical Methods

Hierarchical clustering [9] methods create a nested sequence of clusters, represented in a tree-like structure known as a dendrogram. This approach does not require a predefined number of clusters, which makes it useful for exploratory data analysis. There are two primary types of hierarchical clustering: agglomerative (bottom-up) [10] and divisive (top-down). Agglomerative clustering starts with each data point as a single cluster. In each iteration, the two closest clusters are merged based on a linkage criterion such as single-linkage (minimum distance between cluster members), complete-linkage (maximum distance), or average-linkage (mean distance). This process continues until all data points are merged into a single cluster or until a stopping criterion is met.

Conversely, divisive clustering starts with all data points in one cluster and then splits them into smaller clusters. Although it is used less often because it requires more computing power, divisive clustering can offer valuable insights when looking at the overall structure of the data from a top-down view.

Several hierarchical clustering algorithms have been created to improve efficiency and reliability. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [11] is

designed for very large datasets. It builds a compact summary of the data using a Clustering Feature (CF) tree and performs clustering step by step. CURE (Clustering Using Representatives) [12] improves on traditional hierarchical methods by using multiple representative points for each cluster and moving them toward the cluster center. This lets the algorithm find clusters of any shape and deal with outliers. Chameleon [13] is another method that adjusts to the data's features by looking at both how points are connected and how close they are to each other. This makes it suitable for datasets with complex cluster structures.

2.3 Density-Based Methods

Density-based clustering methods [14] identify clusters as dense regions of data points separated by regions of lower density. These methods work well to detect clusters of different shapes and sizes. They can also spot noise or outliers in the dataset.

The most well-known algorithm in this category is DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN [15] defines a cluster as the largest set of density-connected points. It relies on two parameters: ε (epsilon), which specifies the radius around a point and MinPts, which determines the minimum number of points needed to create a dense region. Points that do not meet the density requirement are considered noise. DBSCAN is efficient and easy to understand, but it can struggle when clusters have very different densities. Its performance is also sensitive to the chosen values of ε and MinPts.

To address DBSCAN's limitations, OPTICS (Ordering Points To Identify Clustering Structure) [16] was developed. OPTICS does not directly create a clustering but instead generates an ordering of the dataset that reveals its natural clustering structure. This method allows for identifying clusters at different density levels without needing a single global density threshold. Another key density-based method is Mean Shift [17] . It uses a kernel density estimation technique to find the densest areas in the feature space. Each data point is gradually moved toward the area of highest data density (the mode). The final positions of the points indicate the clusters. Mean Shift does not require specifying the number of clusters beforehand and can detect clusters of any shape. However, it can be costly in terms of computation, mainly for high-dimensional data.

2.4 Grid-Based Methods

Grid-based clustering methods [18] convert the data space into a finite number of non-overlapping cells that form a grid. Clustering is then performed on these cells instead of on the actual data points. This discretization of the data space reduces computational complexity and makes grid-based methods suitable for large and high-dimensional datasets.

One algorithm using a grid-based method is STING (Statistical Information Grid-based Clustering) [19], which organizes the data space into a hierarchical grid structure and computes statistical parameters, such as mean, count and standard deviation, for each cell. These statistics are then used to find dense regions and form clusters. STING is efficient and scalable but may suffer from accuracy loss due to its dependence on fixed grid resolution.

International Journal of Engineering & Scientific Research
Vol. 5 Issue 7, July 2017,
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://www.esrjournal.com Email: esrjeditor@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Another algorithm is CLIQUE (Clustering In QUEst) [20], which is designed for high-dimensional data. CLIQUE combines grid-based and density-based methods by identifying dense units in subspaces of the data. It begins by dividing the space into a uniform grid and then detects regions where the density exceeds a given threshold. By analysing combinations of dense units across different dimensions, CLIQUE can discover clusters in subspaces that traditional methods might overlook. The grid-based methods can be sensitive to the choice of grid granularity.

2.5 Model-Based and Fuzzy Methods

Model-based clustering assumes that data comes from a mix of known probability distributions, usually Gaussian. The most common method is the Gaussian Mixture Model (GMM) [21], which uses the Expectation-Maximization (EM) algorithm to find the best fit. Unlike K-means, GMM allows soft clustering, where each point can partly belong to multiple clusters. This makes it more flexible but also more sensitive to errors like poor initialization or too many clusters.

Fuzzy clustering also allows partial membership. The most popular method, Fuzzy C-Means (FCM) [22], gives each point a degree of belonging to each cluster. This is helpful when cluster boundaries are unclear or overlapping. FCM updates both cluster centres and membership values to group the data well. However, it can be slower and more affected by noise than simpler methods like K-means.

Table - 1 shows the comparative analysis of various clustering methods along with advantages and limitations.

Table - 1 Comparative analysis of clustering methods.

Clustering Method	**Example Algorithms**	**Cluster Shape Support**	**Noise Handling**	**Scalability**	**Key Parameters**	**Advantages**	**Limitations**
Partitioning	K-Means, K-Medoids, CLARA	Spherical	Poor	High	Number of clusters (k)	Simple, efficient, fast for large datasets	Requires k, sensitive to outliers and initialization
Hierarchical	Agglomerative, BIRCH, CURE	Arbitrary (depends on linkage)	Moderate	Moderate to Low	Linkage type, threshold	Produces dendrogram, no need for k	High computational cost, sensitive to noise

Density-Based	DBSCAN, OPTICS, Mean Shift	Arbitrary	Excellent	Moderate	Epsilon (ε), MinPts	Detects noise/outliers, no need for k	Struggles with varying density, sensitive to parameters
Grid-Based	STING, CLIQUE	Rectangular, subspace-specific	Low to Moderate	High	Grid size, density threshold	Fast, suitable for high-dimensional data	Accuracy depends on grid resolution, potential quantization error
Model-Based	Gaussian Mixture Models	Elliptical	Low to Moderate	Moderate	Number of components, distribution type	Soft clustering, handles cluster overlap	Assumes distribution, risk of overfitting, initialization sensitivity
Fuzzy Clustering	Fuzzy C-Means	Spherical (soft boundaries)	Low	Moderate	Number of clusters, fuzziness coefficient	Allows overlap, good for ambiguous boundaries	Sensitive to noise and outliers, higher computation than K-means

3. Applications of Clustering

Clustering is important [2] in many fields because it helps uncover hidden patterns and structures in unlabelled data. Its unsupervised nature makes it a strong tool for exploring data and discovering knowledge in various practical situations.

One well-known application of clustering is market segmentation [23]. Businesses and marketers group customers based on buying behaviour, demographics, browsing history, or preferences. By identifying these customer segments, organizations can create targeted

marketing campaigns, personalize services and boost customer satisfaction and loyalty. In biology and bioinformatics [24], clustering techniques classify genes with similar expression patterns, group organisms by genetic traits and identify disease subtypes. Clustering is also key in image processing [25] and computer vision. It is used for tasks like image segmentation, object recognition and pattern detection. Techniques like K-means can be used to segment images by grouping pixels based on colour intensity or texture, which simplifies image representation and improves object detection. In social network analysis [26], clustering helps find communities within a network. These communities usually consist of individuals who interact more often with each other than with others outside the group. This analysis plays an important role in understanding social dynamics, influence spread and recommendation systems.

Another significant area is anomaly detection [27], which is crucial in fraud detection, cybersecurity and fault diagnosis. Clustering helps spot patterns that stray from the norm, flagging potential outliers or suspicious activity. For instance, in banking, clustering algorithms can identify unusual transaction patterns that may indicate fraud. In cybersecurity, they can detect network intrusions or malware activity.

Clustering is also fundamental in document [28] and text clustering [29] . It helps manage large collections of unstructured text data. Algorithms like hierarchical clustering and density-based methods group similar documents, making information retrieval, topic modelling and summarization more efficient. In search engines and recommendation systems, clustering improves the personalization of results based on user behaviour and content similarity.

Overall, clustering is a useful and flexible tool used in many fields. It helps find patterns in data and supports better, smarter decisions based on that information.

5. Challenges and Future Directions

Despite significant advancements, clustering techniques still face several key challenges [30] that limit their effectiveness in real-world applications. One major issue is figuring out the right number of clusters. Algorithms like K-means [6] require users to set the number of clusters (k) beforehand, but choosing the best value is not always easy.

Another important challenge occurs when working with high-dimensional [31] data. As the number of dimensions increases, traditional distance-based similarity measures become less reliable, a problem known as the "curse of dimensionality.".

Interpretability [32] is also a concern. It can be hard to understand and explain why data points are grouped together in certain ways with soft or overlapping clusters. Additionally, most standard clustering algorithms are made for static datasets and don't work well in dynamic or streaming data scenarios. This has increased interest in incremental and online clustering methods that can respond to real-time data changes.

Evaluating results of clustering results is also a challenge due to the lack of universal performance metrics. Handling these challenges is essential for developing more robust, interpretable and adaptive clustering methods in the future.

6. Conclusion

Clustering remains a cornerstone of unsupervised learning. With a wide array of algorithms catering to different data types and applications, no single method is universally best. The choice of algorithm depends on the specific dataset characteristics, application goals and performance constraints.

Future research is expected to focus on scalable, explainable and flexible clustering methods. More domain-specific evaluation strategies and hybrid models are also emerging to address complex data representations.

References

[1] C. C. Aggarwal and C. K. Reddy, "Data clustering," *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*, 2014.

[2] L. V. Bijuraj, "Clustering and its Applications," in *Proceedings of National Conference on New Horizons in IT-NCNHIT*, 2013.

[3] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117,* 2012.

[4] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of data science,* vol. 2, p. 165–193, 2015.

[5] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley, 1990.

[6] L. Morissette and S. Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica," *Tutorials in Quantitative Methods for Psychology,* vol. 9, p. 15–24, 2013.

[7] T. S. Madhulatha, "Comparison between k-means and k-medoids clustering algorithms," in *International Conference on Advances in Computing and Information Technology*, 2011.

[8] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, p. 1003–1016, 2002.

[9] S. Patel, S. Sihmar and A. Jatain, "A study of hierarchical clustering algorithms," in *2015 2nd international conference on computing for sustainable global development (INDIACom)*, 2015.

[10] M. R. Ackermann, J. Blömer, D. Kuntze and C. Sohler, "Analysis of agglomerative clustering," *Algorithmica*, vol. 69, p. 184–215, 2014.

[11] T. Zhang, R. Ramakrishnan and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1996.

[12] S. Guha, R. Rastogi and K. Shim, "CURE: An efficient clustering algorithm for large databases," *Information Systems*, vol. 26, p. 35–58, 2001.

[13] G. Karypis, E. H. Han and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, p. 68–75, 1999.

[14] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[15] K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. Sarasvady, "DBSCAN: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, 2014.

[16] M. Ankerst, M. M. Breunig, H.-P. Kriegel and J. Sander, "OPTICS: Ordering points to identify the clustering structure," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1999.

[17] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, p. 790–799, 1995.

[18] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, p. 86–97, 2012.

[19] W. Wang, J. Yang and R. Muntz, "STING: A statistical information grid approach to spatial data mining," in *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)*, 1997.

[20] D. Duan, Y. Li, R. Li and Z. Lu, "Incremental K-clique clustering in dynamic social networks," *Artificial Intelligence Review*, vol. 38, p. 129–147, 2012.

[21] J. Liu, D. Cai and X. He, "Gaussian mixture model with local consistency," in *Proceedings of the AAAI conference on artificial intelligence*, 2010.

[22] J. Nayak, B. Naik and H. Behera, "Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014," in *Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014*, 2014.

[23] K. R. Kashwan and C. M. Velu, "Customer segmentation using clustering and data mining techniques," *International Journal of Computer Theory and Engineering*, vol. 5, p. 856, 2013.

[24] R. Nugent and M. Meila, "An overview of clustering applied to molecular biology," *Statistical methods in molecular biology*, p. 369–404, 2010.

[25] N. Dhanachandra, K. Manglem and Y. J. Chanu, "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm," *Procedia Computer Science*, vol. 54, p. 764–771, 2015.

[26] M. C. Pham, Y. Cao, R. Klamma and M. Jarke, "A clustering approach for collaborative filtering recommendation using social network analysis.," *J. Univers. Comput. Sci.*, vol. 17, p. 583–604, 2011.

[27] I. Syarif, A. Prugel-Bennett and G. Wills, "Unsupervised clustering approach for network anomaly detection," in *Networked Digital Technologies: 4th International Conference, NDT 2012, Dubai, UAE, April 24-26, 2012. Proceedings, Part I 4*, 2012.

[28] D. Cai, X. He and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, p. 902–913, 2010.

[29] C. C. Aggarwal and C. Zhai, "A survey of text clustering algorithms," *Mining text data*, p. 77–128, 2012.

[30] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, p. 651–666, 2010.

[31] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1998.

[32] M. C. N. Barioni, H. Razente, A. M. R. Marcelino, A. J. M. Traina and C. Traina Jr, "Open issues for partitioning clustering methods: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, p. 161–177, 2014.