International Journal of Engineering Sciences & Emerging Technologies, Oct. 2023.
ISSN: 22316604 Volume 11, Issue 2, pp: 62-70 ©IJESET

CLUSTERING IN DATA MINING: TECHNIQUES, ADVANTAGES, APPLICATIONS, AND CHALLENGES

Gautam Kudale, Sandeep Singh Rajpoot

Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.
gaukudale@gmail.com
sandeepraj413@gmail.com

ABSTRACT

Clustering is a technique that groups similar data points together for analysis and pattern discovery across various fields like machine learning, data mining, and image analysis. Its main purpose is to group similar objects together based on a defined distance measure. Essentially, clustering involves partitioning a data set into subsets, with each subset containing data points that are similar to each other. This research paper aims to provide a comprehensive understanding of clustering in data mining. It discusses the concept of clustering, its advantages and disadvantages, applications, various types of clustering techniques, different algorithms, challenges involved, and methods for obtaining the optimal number of clusters.

KEYWORDS: *Data mining, Clustering, Hierarchical clustering, Partition-based clustering, k-means clustering*

1. INTRODUCTION

In data mining, clustering involves grouping data points together based on their similarities, without prior knowledge of group labels [6, 23, 28]. It is an unsupervised technique [18] that discovers patterns [7] and structures in the data. By analyzing distance or similarity measures, clustering algorithms create subsets where data points within each subset are more similar to each other than to those in other subsets [23, 28]. The purpose is to uncover valuable insights, group similar data [20], and gain a deeper understanding of the underlying patterns and characteristics in the dataset. Overall, clustering is a vital technique in data mining, helping researchers extract meaningful information, understand complex data structures, detect anomalies, and facilitate decision-making processes. It empowers data analysts to explore, discover, and gain insights from large datasets in various domains and applications. The below diagram explains the working of the Clustering Algorithm.

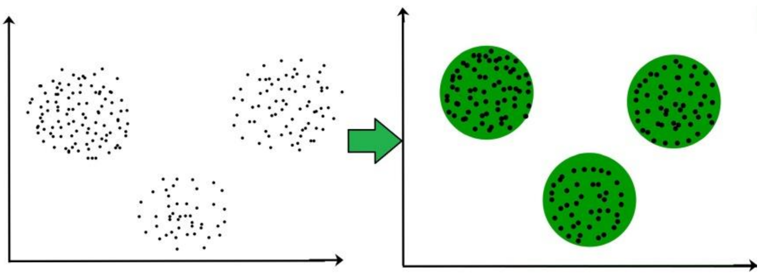

Figure 1. clustering

DOI: 10.5281/zenodo.10434263

Clustering plays a crucial role in analyzing large datasets for several reasons:

Data Exploration: Clustering helps in understanding the structure of complex and massive datasets by identifying inherent patterns and relationships among data points. It provides a way to summarize and explore the data, enabling researchers to gain insights into its underlying characteristics.

Pattern Discovery: Clustering is a valuable tool for revealing hidden patterns and structures that might not be immediately obvious. It accomplishes this by grouping similar data points together, enabling the identification of meaningful subsets within the data. Through clustering algorithms, researchers can uncover trends, associations, and correlations that might otherwise go unnoticed. Essentially, clustering helps unveil valuable insights by organizing data into meaningful clusters based on similarities [7, 10, 18].

Anomaly Detection: Clustering can assist in identifying outliers and anomalies within the dataset. These are data points that deviate significantly from the norm or exhibit unusual behavior. Detecting anomalies is valuable in various domains such as fraud detection, network intrusion detection, and detecting abnormal medical conditions [21].

Scalability: With the exponential growth of data, clustering provides a scalable approach to analyze large datasets efficiently. It allows researchers to partition the data into smaller subsets, making it more manageable for subsequent analysis tasks.

Data Preprocessing: Clustering serves as a preliminary step in data mining, preparing the data for subsequent analysis. By grouping similar data points together, it can reduce the dimensionality of the dataset and simplify subsequent analysis processes.

Decision-Making Support: Clustering helps in making informed decisions by providing a comprehensive overview of the dataset. It can assist in segmenting customers, identifying market segments, optimizing resource allocation, and developing personalized recommendations.

Overall, clustering is essential in analyzing large datasets as it enables researchers to extract meaningful information, understand complex data structures, detect anomalies, and facilitate decision-making processes. It serves as a crucial tool for data exploration, pattern discovery, and gaining insights from vast amounts of data.

2. ADVANTAGES AND DISADVANTAGES OF CLUSTERING IN DATA MINING

2.1 Advangates

- Identifying hidden patterns and structures in data
- Data exploration and understanding
- Anomaly detection and outlier analysis
- Scalability and efficiency in handling large datasets

2.2 Disadvatages

- Sensitivity to input data and distance measures
- Difficulty in choosing appropriate clustering algorithms
- Subjectivity in interpreting clustering results

3. APPLICATIONS OF CLUSTERING IN DATA MINING

Clustering is a fundamental technique in data mining that has various applications across different domains. Some common applications of clustering in data mining include [21]:

- Customer segmentation and market analysis

DOI: 10.5281/zenodo.10434263

- Image and pattern recognition
- Anomaly Detection
- Bioinformatics and gene expression analysis
- Social network analysis
- Document Clustering
- Image and Video Segmentation
- Land use
- Earth quake study
- Market Segmentation
- Image segmentation
- Biological Data Analysis

These are just a few examples of how clustering is applied in data mining. Clustering techniques can be adapted and applied to various domains and datasets to uncover patterns, structure, and insights from complex data.

4. TYPES OF CLUSTERING TECHNIQUES

To address specific practical issues, researchers have categorized clustering analysis into several types: hierarchical, partition-based, density-based, grid-based, and model-based clustering which are shown in table-1. These categories serve different purposes and tackle distinct challenges in the clustering process. [4, 6, 13, 14, 15, 33].

Table – 1 Techniques adopted in clustering

Hierarchical	Partition-based	Density-based	Grid-based	Model-based
CURE CHAMELEON BIRCH COBWEB Probabilistic Agglomerative Vs Divisive	Fuzzy K-Means Clustering Expectation-Maximization Farthest First K-MEDOIDS K-Means CLARANS Filtered Cluster	DBSCAN DENCLUE Make Density Based Clustering Algorithm OPTICS	STING CLIQUE	Gaussian Mixture Models

4.1 Hierarchical clustering (e.g., Agglomerative, Divisive)

Hierarchical clustering is a technique that creates a cluster hierarchy by iteratively merging or splitting clusters based on similarity or dissimilarity. It doesn't require specifying the number of clusters in advance and provides dendrogram visualization. Agglomerative clustering merges clusters bottom-up, while divisive clustering splits clusters top-down. It allows for exploration at different granularity levels but can be computationally intensive and sensitive to distance metric and linkage criterion choices [10, 15].

4.2 Partition-based clustering (e.g., K-means, K-medoids)

Partition-based clustering aims to divide a dataset into non-overlapping clusters. K-means is a widely used algorithm in this category. It assigns data points to the nearest centroid iteratively. Other techniques include Fuzzy C-means and Gaussian Mixture Models. The number of clusters, K, must be specified beforehand. Partition-based clustering is efficient but assumes convex clusters and is sensitive to outliers [6, 13, 14, 15].

DOI: 10.5281/zenodo.10434263

4.3 Density-based clustering (e.g., DBSCAN, OPTICS)

Density-based clustering groups data points based on their density in the feature space. DBSCAN is a popular algorithm in this category. It defines clusters as regions with a minimum number of nearby points. It can discover clusters of arbitrary shape, handles noise well, and doesn't require specifying the number of clusters. Other methods include OPTICS and HDBSCAN. Density-based clustering is useful for complex datasets and spatial data analysis [15].

4.4 Grid-based clustering (e.g., STING, CLIQUE)

Grid-based clustering partitions the data space into a grid structure and assign data points to grid cells. It offers efficiency for large datasets and is suitable for spatial data analysis. STING is a commonly used algorithm in this category. Grid-based clustering is efficient but assumes uniform grid cell sizes and may face challenges with high-dimensional data [4, 6, 15, 33].

4.5 Model-based clustering (e.g., Gaussian Mixture Models)

Model-based clustering assumes data points are generated from probabilistic models. Gaussian Mixture Model (GMM) is a popular algorithm in this category. It estimates model parameters and assigns data points to the most likely cluster. Model-based clustering captures various cluster shapes, handles overlapping clusters, and allows for soft assignments. However, it can be computationally demanding and assumes distribution assumptions [36].

5. Popular Clustering Algorithms

Following are few popular clustering algorithms [10]

- K-means and K-medoids
- Hierarchical Agglomerative Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Expectation-Maximization (EM) algorithm for Gaussian Mixture Models

6. Challenges in Clustering

These are some of the challenges in clustering

- Determining the optimal number of clusters [31]
- Handling high-dimensional data
- Dealing with noisy and incomplete data
- Scalability and efficiency in clustering large datasets

7. Methods Used for Obtaining the Optimal Number of Clusters

There are several methods commonly used for determining the optimal number of clusters in unsupervised clustering analysis. Elbow Method, Silhouette Analysis, Gap Statistic, Information Criteria, Hierarchical Clustering, Domain Knowledge and Interpretability, Model Selection criteria are few popular methods [4, 28, 31, 34, 35].

7.1 Elbow method

The elbow method is a technique used to find the optimal number of clusters in a dataset. It involves plotting the within-cluster sum of squares (WCSS) against the number of clusters. As the number of clusters increases, the WCSS generally decreases since each cluster becomes smaller. However, there is a point where the rate of improvement slows down significantly, resulting in a bend or "elbow" in the plot. This bend represents a trade-off between a lower WCSS and a simpler, more interpretable

DOI: 10.5281/zenodo.10434263

solution. The number of clusters corresponding to this elbow point is often chosen as the optimal number for subsequent analysis [31, 34, 35].

7.2 Silhouette analysis

Silhouette analysis assesses the quality of clustering by measuring how well each data point fits into its assigned cluster. It calculates an average silhouette coefficient for varying numbers of clusters. The silhouette coefficient ranges from -1 to 1, with higher values indicating better clustering. The optimal number of clusters is typically determined by selecting the number that yields the highest average silhouette coefficient [4, 28, 35].

It's important to note that no single method is universally applicable, and the choice of the optimal number of clusters depends on the specific dataset and the underlying problem. It is often recommended to combine multiple methods and perform robustness checks to ensure the stability of the clustering results.

Importance and future directions of clustering research

The importance of clustering research lies in its wide range of applications and the potential for advancing data analysis techniques. Here are some key reasons why clustering research is significant:

Improved Data Understanding: Clustering helps researchers gain a deeper understanding of complex datasets by revealing underlying patterns and structures. This understanding can lead to better decision-making, improved resource allocation, and enhanced problem-solving in various domains.

Enhanced Data Exploration: Clustering enables data exploration by summarizing and organizing large datasets. It provides a starting point for further analysis, helping researchers identify areas of interest, potential relationships, and areas requiring further investigation.

Pattern Discovery and Anomaly Detection: Clustering facilitates the discovery of hidden patterns and associations within data. It also aids in identifying anomalies or outliers that may be indicative of important events or abnormalities, enabling effective anomaly detection and outlier analysis.

Scalability and Efficiency: With the exponential growth of data, clustering research plays a crucial role in developing scalable and efficient algorithms. Efficient clustering algorithms enable the analysis of large datasets in a reasonable amount of time, ensuring that data analysis processes keep pace with the growing data volumes.

Integration with Other Data Mining Techniques: Clustering research contributes to the integration of clustering with other data mining techniques, such as classification, regression, and association rule mining. This integration enhances the capabilities of data analysis and leads to more comprehensive and accurate insights.

Real-World Applications: Clustering has diverse applications in various fields, including marketing, healthcare, finance, bioinformatics, and social network analysis. Continued research in clustering opens up new possibilities for applying this technique to solve complex problems and improve decision-making in these domains.

Future directions in clustering research include:

Scalable Algorithms: Developing clustering algorithms that can handle even larger datasets efficiently, ensuring scalability and reducing computational requirements.

Handling Complex Data Types: Extending clustering techniques to handle complex data types such as text, images, time series, and graphs, enabling effective analysis in diverse data domains.

DOI: 10.5281/zenodo.10434263

Hybrid Approaches: Exploring hybrid clustering approaches that combine multiple clustering algorithms or integrate clustering with other data mining techniques to improve clustering accuracy and performance.

Interpretable Clustering: Advancing interpretability of clustering results by providing meaningful explanations and insights into the underlying patterns and structures identified by clustering algorithms.

Handling High-Dimensional Data: Developing clustering techniques that can effectively handle high-dimensional data by addressing the curse of dimensionality and discovering meaningful clusters in high-dimensional spaces.

Stream Data Clustering: Investigating clustering algorithms that can handle streaming data, enabling real-time analysis and clustering of continuous data streams.

By focusing on these future directions, clustering research can further enhance its impact and enable more accurate, efficient, and scalable data analysis in various domains.

8. CONCLUSION

By addressing the above aspects, this research paper will provide

1) A comprehensive overview of clustering in data mining, enabling readers to understand its significance, advantages, limitations, applications, and challenges.

2) Additionally, it will offer insights into determining the optimal number of clusters, facilitating effective clustering analysis in various domains.

3) Future research focuses on scalable algorithms, handling complex data, hybrid approaches, interpretability, high-dimensional data analysis, and stream data clustering. Overall, clustering enhances data analysis and decision-making across various domains.

ACKNOWLEDGEMENT

Authors would like to thank Dr. Ranjit Patil, Principal, Dr. D. Y. Patil Arts, Science and Commerce College, Pimpri, Pune (MS) for helping to write this research paper and useful discussions. I also like to express my sincere thanks to Dr. Bharat Shinde, Principal, Vidya Pratishthan's Arts, Science and Commerce College, M.I.D.C., Baramati, Pune (MS). I would like to thank Mr. Gajanan Joshi, Head Department of Computer Science, Vidya Pratishthan's Arts, Science and Commerce College, M.I.D.C., Baramati, Pune (MS) for giving me valuable guidelines and suggestions regarding this work.

REFERENCES

[1] Data Mining Introductory and Advanced Topics, Margaret H. Dunhan, Pearson

[2] Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Ian H.witten, Eibe Frank, Mark A. Hall

[3] Mining of Massive Datasets, 2nd Edition, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman

[4] Data Mining, Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei

[5] Prof. Prashant Sahai Saxena, Prof. M. C. Govil, "Prediction of Student's Academic Performance using Clustering," Special Conference Issue: National Conference on Cloud Computing & Big Data

[6] Bindiya M Varghese, Jose Tomy J, Unnikrishnan A, Poulose Jacob K, "Clustering student data to characterize performance patterns," (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence,

DOI: 10.5281/zenodo.10434263

[7] Md. Hedayetul Islam Shovon, Mahfuza Haque, “An Approach of Improving Student’s Academic Performance by using K-means clustering algorithm and Decision tree,” (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.3, No. 8, 2012

[8] Oyelade, O. J, Oladipupo, O. O., Obagbuwa, I. C., “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance,” (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010

[9] Rakesh Kumar Arora, Dr. Dharmendra Badal, “Evaluating Student’s Performance Using k-Means Clustering,” International Journal of Computer Science And Technology, IJCST Vol. 4, Issue 2, April - June 2013, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

[10] Sharmila, R.C Mishra, “Performance Evaluation of Clustering Algorithms,” International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013, ISSN: 2231-5381

[11] Ramjeet Singh Yadav, P. Ahmed, A. K. Soni and Saurabh Pal, “Academic performance evaluation using soft computing techniques,” CURRENT SCIENCE, VOL. 106, NO. 11, 10 JUNE 2014

[12] Harwatia, Ardita Permata Alfiania, Febriana Ayu Wulandaria, “Mapping student’s performance based on data mining approach (a case study),” The 2014 International Conference on Agro-industry (ICoA): Competitive and sustainable Agro industry for Human Welfare, Agriculture and Agricultural Science Procedia 3 (2015) 173 – 177

[13] Patel, J. and Yadav, R.S. (2015) “Applications of Clustering Algorithms in Academic Performance Evaluation.” Open Access Library Journal, 2: August 2015 | Volume 2 | e1623

[14] Jyotirmay Patel, Ramjeet Singh Yadav, “Applications of clustering algorithms in academic performance evaluation”

[15] Atul Prakash Prajapati, Sanjeev Kr. Sharma, Manish Kr. Sharma, “Student’s performance analysis using machine learning tools,” International Journal of Scientific & Engineering Research Volume 8, Issue 10, October-2017 ISSN 2229-5518

[16] E.Venkatesan, S.Selvaragini, “Prediction of students academic performance using classification and clustering algorithms,” International Journal of Pure and Applied Mathematics Volume 116 No. 16 2017, 327-333 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)

[17] Snehal Bhogan , Kedar Sawant , Purva Naik , Rubana Shaikh , Odelia Diukar , Saylee Dessai, “Predicting student performance based on clustering and classification,” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN:2278-8727, Volume 19, Issue 3, Ver. V (May-June 2017), PP 49-52

[18] Mr. Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, “Evaluating student’s performance using k-means clustering,” International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 6 Issue 05, May – 2017

[19] Mrs .Mary vidya john, Akshata police patil, Anjali mishra, Bindhu reddy G, Jamuna N, “Clustering technique for student performance,” International Research Journal of Computer Science (IRJCS), Issue 06, Volume 6 (June 2019), ISSN: 2393-9842

[20] Noel Varela , Edgardo Sánchez Montero , Carmen Vásquez , Jesús García Guiliany , Carlos Vargas Mercado , Nataly Orellano Llinas , Karina Batista Zea , and Pablo Palencia, “Student performance assessment using clustering techniques,” © Springer Nature Singapore Pte Ltd. 2019 Y. Tan and Y. Shi (Eds.): DMBD 2019, CCIS 1071, pp. 179–188, 2019. https://doi.org/10.1007/978-981-32-9563-6_19

[21] N.Valarmathy, S.Krishnaveni, “Performance evaluation and comparison of clustering algorithms used in educational data mining,” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S5, April 2019

[22] Lubna Mahmoud Abu Zohair, “Prediction of Student’s performance by modelling small dataset size,” Abu Zohair International Journal of Educational Technology in Higher Education (2019) 16:27 https://doi.org/10.1186/s41239-019-0160-3

[23] Mrs. Bhawna Janghel, Dr. Asha Ambhaikar, “Performance of student academics by k-mean clustering algorithm,” International J. Technology. January – June, 2020; Vol. 10: Issue 1, ISSN 2231-3907 (Print), ISSN 2231-3915 (Online)

DOI: 10.5281/zenodo.10434263

[24] Marzieh Babaie, Mahdi Shevidi Noushabadi, "A review of the methods of predicting students' performance using machine learning algorithms," Archives of Pharmacy Practice ¦ Volume 11 ¦ Issue S1 ¦ January-March 2020

[25] Dr. G. Rajitha Devi, "Prediction of student academic performance using clustering," International Journal of Current Research in Multidisciplinary (IJCRM) ISSN: 2456-0979 Vol. 5, No. 6, (June'20), pp. 01-05

[26] Dewi Ayu Nur Wulandari; Riski Annisa; Lestari Yusuf, Titin Prihatin, "An educational data mining for student academic prediction using k-means clustering and naïve bayes classifier," journal Pilar Nusa Mandiri Vol 16, No 2 September 2020

[27] Yann Ling Goh, Yeh Huann Goh, Chun-Chieh Yip, Chen Hunt Ting, Raymond Ling Leh Bin, Kah Pin Chen, "Prediction of students' academic performance by k-means clustering," Peer-review under responsibility of 4th Asia International Multidisciplinary Conference 2020 Scientific Committee

[28] Revathi Vankayalapati, Kalyani Balaso Ghutugade, Rekha Vannapuram, Bejjanki Pooja Sree Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," Revue d'Intelligence Artificielle Vol. 35, No. 1, February, 2021, pp. 99-104

[29] Rina Harimurti , Ekohariadi, Munoto , I. G. P Asto Buditjahjanto, "Integrating k-means clustering into automatic programming assessment tool for student performance analysis," Indonesian Journal of Electrical Engineering and Computer Science Vol. 22, No. 3, June 2021, pp. 1389~1395 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i3.pp1389-1395

[30] Rui Shang , Balqees Ara, Islam Zada, Shah Nazir , Zaid Ullah, and Shafi Ullah Khan, "Analysis of simple k-mean and parallel k-mean clustering for software products and organizational performance using education sector dataset," Hindawi Scientific Programming Volume 2021, Article ID 9988318, 20 pages https://doi.org/10.1155/2021/9988318

[31] Bao Chong, "K-means clustering algorithm: a brief review," Academic Journal of Computing & Information Science ISSN 2616-5775 Vol. 4, Issue 5: 37-40, DOI: 10.25236/AJCIS.2021.040506

[32] Said Abubakar Sheikh Ahmed, "Evaluating students' performance of social work department using k-means and two-step cluster "a case study of mogadishu university"," Mogadishu University Journal, Issue 7, 2021, ISSN 2519-9781

[33] Zhihui Wang, "Higher education management and student achievement assessment method based on clustering algorithm," Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 4703975, 10 pages https://doi.org/10.1155/2022/4703975

[34] Ahmad Fikri Mohamed Nafuri , Nor Samsiah Sani, Nur Fatin Aqilah Zainudin , Abdul Hadi Abd Rahman and Mohd Aliff, "Clustering analysis for classifying student academic performance in higher education," Appl. Sci. 2022, 12, 9467. https://doi.org/10.3390/app12199467.

[35] Trupti M. Kodinariya, Dr. Prashant R. Makwana "Review on determining number of Cluster in K-Means Clustering", International Journal of Advance Research in Computer Science and Management Studies Volume 1, Issue 6, November 2013 ISSN: 2321-7782 (Online).

[36] K.Renuka Devi "Evaluation of Partitional and Hierarchical Clustering Techniques", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 8, Issue. 11, November 2019, pg.48 – 54ISSN 2320–088X.

AUTHORS

Gautam A. Kudale- Received his Bachelor in Computer Science and Masters in Computer Science from Savitribai Phule Pune University, Pune, Maharashtra. He is a Ph.D. Candidate in Dr. APJ Abdul Kalam University, Indore, M.P. India. He is working as Assistant Professor, in Department of Computer Science, Vidya Pratishthan's Arts, Science and Commerce College, M.I.D.C., Baramati, Pune (MS). His research interests are in Data Mining, Digital Image Processing.

DOI: 10.5281/zenodo.10434263

Dr. Sandeep Singh Rajpoot- Received his Bachelor of computer Application from Dr Hari Singh Gour University, Sagar and Master of Computer Application from RGPV, Bhopal. Ph.D.(CSA) from Dr A.P.J. Abdul Kalam University, Indore. Currently working as an Associate Professor in the Department of Computer Application, College of Engineering, Dr. APJ Abdul Kalam University, Indore.

DOI: 10.5281/zenodo.10434263