WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comparative Analysis on Different Data Clustering Algorithms

Richa. S[1], Shanmugapriya[2]

[1,2]B. Sc AIML, Sri Krishna Arts and Science College Coimbatore

A B S T R A C T

Clustering is a popular technique used in data mining and machine learning to group similar data points together based on certain criteria. In this paper, we conduct a comparative study of various clustering algorithms for journal paper clustering. Journal paper clustering is a crucial task in the field of academic research, where large volumes of scholarly articles are published every year, making it difficult for researchers to stay up-to-date with the latest developments in their respective fields. We evaluate several clustering algorithms, including K-means, hierarchical clustering, DBSCAN, and spectral clustering, on a dataset of journal papers from various academic domains. We compare their performance based on clustering quality metrics, such as silhouette score, Davies-Bouldin index, and Calinski-Harabasz index .Our experimental results show that spectral clustering outperforms other clustering algorithms in terms of clustering quality metrics, while hierarchical clustering performs better in terms of computational efficiency. We also investigate the impact of different text preprocessing techniques, such as stemming, stop-word removal, and TF-IDF normalization, on the performance of clustering algorithms. The results of our study provide insights into the selection of suitable clustering algorithms and text preprocessing techniques for journal paper clustering, which can aid researchers in exploring large volumes of scholarly articles efficiently.

INTRODUCTION:

Clustering is a widely used technique in data mining and machine learning, with applications in various fields such as marketing, biology, image processing, and many more. The main goal of clustering is to group similar objects together in order to discover underlying patterns in the data. Clustering algorithms are an essential part of this process, and there are numerous algorithms available that can be applied to different types of data and problems.The aim of this paper is to provide an analysis of various clustering algorithms used in data mining. We will begin by introducing the concept of clustering and its importance in data mining. Then, we will discuss different types of clustering algorithms, including partition-based clustering, hierarchical clustering, density-based clustering, and model-based clustering. For each type of clustering, we will provide an overview of the algorithm, its strengths and weaknesses, and its suitability for different types of data .We will also discuss various evaluation metrics that can be used to measure the quality of clustering results, such as silhouette score, Davies-Bouldin index, and Rand index. Additionally, we will provide a comparative analysis of different clustering algorithms based on their performance on various datasets and evaluation metrics.Overall, this paper aims to provide a comprehensive analysis of clustering algorithms in data mining, which can be useful for researchers and practitioners in the field. By understanding the strengths and weaknesses of different clustering algorithms, researchers can choose the most suitable algorithm for their specific data and problem, while practitioners can apply these algorithms to their data to discover meaningful patterns and insights

Clustering and it's types:

Clustering is stipulated as the sorting of indistinguishable text report into clusters that is reports within the clusters have high parallelism when compared to other but heterogeneous to reporoidsts in other clusters [4]. Since all the information have been added to the World Wide Web it becomes very consequential to graze or probe the pertinent information dramatically. The identification of appropriate algorithms for clustering induces the optimal clustering techniques, and inclines imperative to possess the contraption for differentiating the consequences of clustering techniques. Several heterogeneous clustering process to retain directive to decode the issue from distinct approach,that is,

- **Partitioned clustering**
- **Density based clustering**
- **Hierarchical clustering**

Partitioned clustering:

Partitioning methods breaks down the data into a set of different clusters. Give no objects, this method produces k clusters of data where k

Hierarchical clustering:

Hierarchical Methods is fester a given data object set for several level, so it will size a grouping tree. According to the ordered breakdown is based on bottom-up or top-down principle, it can be further divided into overflowing together and division. The former to each entity as a divide class and it process with the data comparison, then a suitably large data progressively merged into larger categories; the latter equally, the entire set as a group, and then regularly divided into small dissimilar types. In order to make up for no definite deficiency of breakdown or polymerization, hierarchical clustering technique often mingle some other methods,such as circular positioning. This method creates a hierarchical break of the given set of data objects. We canclassify hierarchical methods on the basis of how the hierarchical break of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical disintegration is formed.They are two types of approaches here Agglomerative approach and divisive approach.

Agglomerative Approach:

Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

Divisive approach:

Also known as a top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.

Density based clustering:

Density-based Methods, it see cluster as a high density area what is splitting a hole by low density section. The basic idea is following: as long as the adjoining room section of the dot concentration (number of data points) beyond a certain entrance, will maintain to clustering, till the field must contain at least a certain number of points. This process is based on the idea of density. The basic idea is to continue rising the given cluster as long as the solidity in the district exceeds some entrance, i.e., for each data point within a given cluster as long as the the solidity in the district exceeds some entrance, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a least amount number of points.

Application of clustering:

Clustering in data mining:

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. On the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.There are several different ways to implement this partitioning, based on distinct models. Distinct algorithms are applied to each model, differentiating its properties and results. These models are distinguished by their organization and type of relationship between them.

Application of clustering in text mining:

Text mining, also referred to as text data mining, approximately consistent to text analytics, refers to the procedure of originates first-rate information from text. Excellent in initiate is typically copied through the plan of patterns and developments through means such as arithmetical pattern learning. Text mining frequently encompasses the process of structuring the input text, deriving patterns within the ordered data, and after all evaluation and analysis of the output. ‘High quality’ in text mining usually refers to some combination of application, novelty, and interestingness. Typical passage removal tasks contain copy category, text clustering ,object extraction, construction of rough taxonomy, opinion analysis, text summarization, and object relative model Text mining consists of removal in order from hidden patterns in large text-data collections.

Conclusion:

Taking due consideration of the relative advantages and disadvantages offered by the usage of Hierarchical and Partitional methods, the application of either techniques onto a problem domain varies according to the size of the dataset involved and the cluster quality desired. The practical implementation may even incorporate both the techniques to get the best of both words.

Reference:

✓ Analysis On Algorithm And Application Of Cluster In Data Mining ,Yuhua Feng Information Engineering School Of Nanchang University, Nanchang 330031, Jia

✓ Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber Second Edition.

✓ Mythili, S., & Madhiya, E. (2014). An Analysis on Clustering Algorithms in Data Mining Abstract :,3(1) 334-340

✓ Thouheed Ahmed S., Sandhya M. “Real-Time Biomedical Recursive Images Detection Algorithm for Indian Telemedicine Environment”. In: Mallick P., Balas V., Bhoi A., Zobaa A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore 2019 Rashi Chauhan,Pooja Batna, Sarika Chaudhary “A survey of density based clustering

✓ algorithm” International journal of computer science & technology (IJCST) , vol 5.