Quest Journals
Journal of Research in Environmental and Earth Sciences
Volume 10 ~ Issue 5 (2024) pp: 15-18
ISSN(Online) :2348-2532
www.questjournals.org

Research Paper

Satellite data Characterization using K Means Clustering Techniques – A Review

Binay Kumar Singh*[1] and Surendra Yadav[2]
[1] *Department of Remote Sensing, BIT Mesra, Off-Campus Jaipur*
[2] *Department of Computer Science, Vivekananda Global University, Jaipur*
** Corresponding author*

ABSTRACT: *This study provides an extensive analysis of the various methods of k-means clustering algorithms that have been developed over the years. Sorting the pixels that need to be analyzed into different groups is the aim of the k means method. Among all the k means algorithms, the traditional k-means algorithm is the most widely used. There are various algorithms available, such as the standard, basic, and classic k-means. Each pixel is assigned to its nearest centroids by these algorithms using the Euclidean distance and the minimum distance rule approach. Data clustering is used to group pixels in a collection that are more similar to one another than to pixels in other class. Since the scatterometer data is available in 2 days interval and resolution is 2 km, volume of data will be very large. Objective is to use the scatterometer data for various land applications for different users of remote sensing. Derived parameters from Ku band SAR (non-imaging scatterometer data) is to characterize the spatial and temporal signatures of regional land surfaces in Rajasthan and other parts of India. Scatterometers were designed to map wind speed and wind direction over the oceans. A rule-based classifier uses an array of if-then rules to categorize data from the Normalised Difference Vegetation Index (NDVI).*

Keywords: *Scatterometer, Cluster Analysis, unsupervised, K-means clustering, Euclidean distance*

R*eceived 07 May, 2024;* ***R****evised 17 May, 2024;* ***A****ccepted 19 May, 2024 © The author(s) 2024.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION

Data from ScatSAT sensors is mostly utilized for oceanography and the identification of oceanic disasters. The process of classifying data into appropriate groups in order to use, store, and protect it more efficiently is known as data classification. Data categorization and feature identification for land-use to detect features, identify changes in characteristics like agricultural phenology, forest degradation, and disaster management (change as a result of any disaster), and to increase the utility of ScatSAT image for land-use purposes.

ScatSAT data are classified using unsupervised classification method, which can divide the input data into a certain number of classes (say, "k") called K-means clustering. After cluster analysis, the data validation is being done using NDVI Data (Bhuvan portal-ISRO). Here, we employ it to identify a potential set of classes on the ScatSAT Image data, where each class might stand for a distinct geographical feature. Research on image processing, data mining, pattern recognition, data clustering etc. has significance because of its various applications. [30] Clustering can be done on multiple variable data sets, single variable datasets based on distances and similarities. [7]. It is a frequently used tool for statistical data analysis and is one of the best approaches for multiple variable data sets.

This led to the publications of a wide range of books by different authors Fisher [35], Tryon and Bailey [34], Hartigan [31], Spath [25], Aldenderfer [28], Romesburg [6] and others which in turn sparked research on clustering techniques globally.

K-means clustering [4] is the most often used technique for optimizing the predicted similarity between data sets and the cluster centroids they are associated with.

II. CLUSTER ANALYSIS

An unsupervised classification method is known as cluster analysis. Clustering divides a set of data/pixels that is typically complex, into clusters(groups). This clustering is done on the basis of the similarities of the data/pixel values [24 and 29].

Two primary forms of cluster analysis are Centroid based and hierarchical/non-hierarchical based clustering. A technique that puts related items together is called hierarchical clustering, often known as hierarchical cluster analysis. The clusters are gradually divided using a similarity metric.

Every step of the hierarchical process is depicted in the dendogram along with the connections formed between groups based on similarities and differences. The data collection is divided using techniques so that each pair of data clusters is unique. First cluster partition is adjusted until a locally optimal partition is found. [29].

III. K-MEANS CLUSTERING

K-means technique divides N data values into K unique clusters. Most researchers frequently used partitioning based clustering method that uses centroids for cluster imaging [3] is k-means clustering. Overall within cluster squared error criteria is used to evaluate the quality of k-means clustering. [10, 12 and 33]

Finding the number of clusters in the data is necessary before utilizing any of the k-means algorithm options. It might take multiple attempts to determine the optimal number of clusters. Using k-means approach it reduces the k-means problem. Since the ability to produce a global optimum depends on the properties of the data and its size.

The two iteration phases of k-means clustering methods are the centroid update phase and the initialization phase. In the first instance, cluster centroids are updated based on the partition obtained by the previous phase, while in latter, each data is assigned to its nearest centroid using Euclidean metric. Either no data changes clusters or the predetermined maximum number of iterations is achieved, marking the conclusion of the iterative process. [4]

The standard k-means approach was first presented by Forgy [21]. It is a batch algorithm that minimizes the average squared Euclidean distance between the data points and the cluster's centroid, a geometrical feature that is essentially the center and a generalization of the mean. Selecting the number of clusters, k, which match the cluster centers is the first step in the Forgy algorithm. Then it designates a pixel from the set to the cluster with the nearest centroid. Finally in order to update new centroids for each cluster, it averages the data points or objects associated to the cluster. If the cluster center remains unchanged, iteration comes to the end.

Unlike Forgy's method, which treats data distribution as continuous, Lloyd's technique [27] treats data distribution as a unique example. Lloyd also presented the classic k-means approach, a batch algorithm.

The foundation of the k-means algorithm was created by MacQueen [33]. This is an online algorithm that is different in updating the process yet similar to Lloyd's and Forgy's algorithms in terms of setup. Every time the centroid changes, the points are recalculated as part of the MacQueen algorithm refresh to update the centroids. The process ends when all the points are allocated to the cluster that has the closest centroid.

Conventional k-means approach was also proposed by Hartigan [29] and Wong [28]. This algorithm, which is not Lloyd or Forgy, adjusts cluster centers at each point in the dataset, rather than simply after each iteration. With this method, a point that is now in the subspace of the closest centroid might be assigned to a new subspace; it looks for the data space partition that has a locally best within cluster sum of squares of errors (SSE). For every data point contained in another cluster, the within cluster sum of squares is computed if the centroid of each included data point has been updated.

The K-means clustering algorithm is widely used for clustering datasets from various areas and is renowned for its simplicity. Few implementation-related issues severely limit its performance. Consequently, a great deal of study has been done to enhance the algorithm's overall performance [2].

Shuyu Miao [1] proposed a new technique called K-means Clustering Based Feature Consistency Alignment (KCFCA), designed to deal with different dataset distribution shifts. Labeled training sets and unlabeled test sets are clustered using the K-means algorithm by KCFCA, which then aligns the cluster centers with feature consistency. Then create a dynamic regression model in order to represent the connection between the distribution shifts and model accuracy.

Preeti Tahlani [03] mapped rice crop phenology using SCATSAT 1 Ku-Band data. Multi-temporal SCATSAT 1 was used to derive the rice crop phenology. Using multi-temporal Sentinel1 data gathered, a rice crop map created under the FASAL project was used to identify the pixels of the rice crop in the SCATSAT image. Extensive field verification conducted during the growth phase confirms that the categorization accuracy is above 90%. The SCATSAT1 pixel's rice cropped region was chosen using the resulting rice map as a guide. Large, connected paddy fields make up the majority of the farmed land in Punjab and Haryana. Thus mapping rice crop phenology with SCATSAT 1 Ku-Band data was done.

Basic correlation studies conducted over a thorough examination of nine urban locations in the United States, that the DSM-processed spaceborne radar data had a strong correlation with the airborne lidar data. The trended DSM and the high linear correlations between product and the trended lidar product show that, in the lack of extensive lidar acquisitions, the DSM approach is a reliable and feasible way to estimate urban volumes. These findings demonstrate the value of using radar backscatter from satellites to measure urban expansion in three dimensions over an extended period of time for cities around the globe [3].

IV. RELATED LITERATURE

A version of Forgy's et. al. k-means technique (Anderberg. [30]) was proposed by Jancey [34]. It is anticipated to result in less favorable local minima and speed up convergence. The old center is updated by reflecting the new cluster mean through the old center. The new cluster is not the mean of the old and added locations. Many evolving algorithm based techniques have been developed to prevent using unsatisfactory local solutions [19, 21].

Likas [5] created the global k-means clustering algorithm, an efficient and iterative global optimization technique [6]. Moreover, the k-means algorithm is utilized as a local search method instead of the costly and complete global k-means strategy. Moreover it does not depend on any prior conditions.

V. CONCLUSION

This paper reviewed the state of the art k-means clustering algorithms. According to this paper, there have been numerous k-means algorithm variations from 1960's to the present that have addressed various k-means algorithmic shortcomings. In order to analyze the accuracy and efficiency of various k-means clustering algorithm for the characterization of satellite data analysis, by exploring computational complexity.

The K-means clustering algorithm is widely used for clustering datasets from multiple domains and is recognized for its simplicity. Despite this benefit, a few implementation-related issues seriously impair its performance. Consequently, a great deal of study has been done to enhance the algorithm's overall performance. The multiple versions created to address the stated issues and the standard algorithm's varied limitations have been uncovered.

ACKNOWLEDGEMENT

The author is thankful to team of Space Application Centre (SAC) Ahmedabad for sponsoring the project. Thanks to BIT Mesra who provided the required assistances during the execution of the project.

REFERENCES

[1]. Miao, Shuyu, Zheng, Lin, Liu, Jingjing (2023). K-means Clustering Based Feature Consistency Alignment for Label-free Model Evaluation; Accepted by CVPR 2023 workshop

[2]. Abiodun M. Ikotun, A. E. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Journal of Information Sciences, Vol 622, 178-210.

[3]. Preeti Tahlani, P. G. (2019). Identification of Rice crop phenology using SCATSAT-1 KU-BAND Scattermoter in Punjab and Haryana. ISPRS-GEOGLAM-ISRS Joint Int. Workshop on Earth Observations for Agricultural Monitoring (pp. 549-555). New Delhi: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

[4]. N. Slonim, E. A. (2013). Hartigan's K-Means Versus Lloyd's K-Means-Is It Time for a Change? . Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 1677-1684.

[5]. Likas, N. V. (2003). The global k-means clustering algorithm. Pattern Recognition by Elsevier Science Direct, Volume 36, Issue 2, Pages 451-461

[6]. Maulik, S. B. (2002). An evolutionary technique based on K-Means algorithm for optimal clustering in Rn. Information Science, Volume 146, Issues 1–4, Pages 221-237.

[7]. Yang, C. Y. (2019). Research on k-value selection method of k-means clustering algorithm. Multidisciplinary Scientific Journal, 2(2), 226-235.

[8]. B. Everitt, S. L. (2011). Cluster Analysis, 5th ed. John Wiley and Sons.

[9]. T. Haste, R. T. (2009). the Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed. Springer Series.

[10]. Maulik, S. B. (2002). An evolutionary technique based on K-Means algorithm for optimal clustering in Rn. Information Science, Volume 146, Issues 1–4, Pages 221-237.

[11]. Murty, K. K. (1999). Genetic K-Means algorithm. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 29(3): 433-439.

[12]. Spath, . (1989). Cluster Analysis Algorithms. Chichester.

[13]. Dubes, R. a. (1988). Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall.

[14]. Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transaction on Information Theory, 28(2), 129-137.

[15]. Wong, J. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), Journal of the Royal Statistical Society. Series C (Applied Statistics).

[16]. Hartigan, J. A. (1975). Clustering Algorithms. New York: John Wiley & Sons. Inc.

[17]. Anderberg, M. R. (1973). Cluster Analysis for Applications. Academic Press.

[18]. Bailey, R. C. (1970). Cluster Analysis. New York: McGraw Hill.

[19]. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Fifth Berkeley Symposium on Mathematical Statistics and Probability, (pp. vol 1, 281-297).

[20]. Jancey, R. C. (1966). Multidimensional group analysis. Australian Journal of Botany, 14(1), 127-130.

[21]. Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics, vol. 21, 768-769.