International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 13 (2015) pp 33143-33147

An Improved k-means Clustering Algorithm for Image Segmentation

Suresh Kurumalla[1], P Srinivasa Rao [2]
[1]Research Scholar in CSE Department, JNTUK Kakinada
[2] Professor, CSE Department, Andhra University,Visakhapatnam,AP, India
Mail id:kurumallasuresh@gmail.com

Abstract- Clustering is defined as the process of grouping similar objects and also an essential procedure within the arena of Data Mining. Since, the volume of information for numerous applications continues to rise day by day in terms of its size and dimensionality, it is essential to have effective and enhanced clustering approaches. K-means is the most prevalent clustering procedure that adopts the greedy approach to generate a group of K-clusters with accompanied centers of mass, and uses a squared error distortion measure to specify the convergence. The traditional K-means algorithm is not very efficient due to some of the reasons like, it is computationally expensive, and the group of obtained clusters intensely based on the selection of initial centroids, selection of the size of the cluster and data items for clustering. In order to address the problem of computationally expensive in k-means clustering algorithm, this paper proposed a novel clustering algorithm known as an Improved K-means by minimizing the size of the data objects. The experimental results are conducted on the image datasets for proposed Improved k-means Clustering Algorithm to generate clusters with better accuracy and less computational time, thus improving the efficiency of G-means clustering algorithm..

1. Introduction

In today's world, due to the substantial growth of digital information, the data is being doubled for every 20 months. The hasty improvements and extensive usage of information systems for the last 30 years had led to emerging of varieties of databases in magnitudes, personals and administrations. The computerization of business events generates ever-increasing streams of information because even simple communications, such as a telephone call, the usage of a credit card, or a medicated are usually recorded in a processor. With the widespread use of the Databases and the volatile progression of individual users and organizations are facing with the difficulty of usage of the huge data.

Numerous conventional methodologies for analysis and visualization has been introduced to face the rational usage of data, but due to increases in the volume of these data exponentially, traditional techniques had become inadequate. Data mining is the evolving method to mine models, outlines or knowledge of importance from enormous data stores [1]. From the last twenty years, data mining is being developed in an intense promptness. The novel methodologies initiated from the approaches from statistics, pattern recognition, databases, and artificial intelligence etc. [2]. Recent days, these are up surging to a multi-disciplinary study. Data mining is a non-trivial process that recognizes the active, unidentified, hypothetically beneficial and eventually the apprehensible design from data stores.

Data mining and KDD are swiftly growing areas of exploration which are at the juncture of numerous domains with high performance and parallel computing. The knowledge discovery practice in industry is accomplished essentially by analysis whose primary exercise and proficient duties are in static and data analysis. Data mining encompass finding models to or define patterns from perceived data. The more common data mining functions are association rule discovery, sequence analysis, classification rule, clustering analysis and deviation detection.

1.1. Clustering

Clustering is a significant process in data mining that partitions data objects into classes without categorized objects. The illustrations appropriate to identical classes are homogenous and the samples among the classes are heterogeneous. The procedure of assemblinga group of physical or abstract entities into classes of identical entity is called clustering. A cluster is a group of data items that are identical to each other with in the similar cluster and are dissimilar to the item in other clusters. A group of object entities can be canned jointly as one cluster in several solicitations. Cluster analysis is extensively employed in various applications, comprising of pattern recognition, data exploration, image processing and market research.

The clustering technique classifies thick and thin regions and consequently, determines complete distribution patterns and fascinating relationships amongst attributes. As a data mining task, clustering is employed as a stand-alone tool to achieve intuitions inthe distribution of information, to perceive the appearances of every cluster and to emphasize on specific group of clusters for further study. Furthermore, the clustering technique also function as a preprocessing stage for other algorithm, such as characterization and categorization that functions on the identified clusters. Clustering is an inspiring domain of study where it's potential solicitations pure its individual exceptional necessities.

Clustering attempts to assemble a group of items and discover whether there is certain association between those items. Clustering technique is employed in numerous areas such as engineering, biology, medicine and data mining [7]. It partitions the entities into clusters by diminishing the squared distances amongst the entities and the centroid of the clusters.

1.2. Motivation

The conventional clustering techniques like k-means algorithm have complications in supervising the challenges postured by pool of regular statistics that are frequently ambiguous and unclear. The K-Means clustering is

simple, but it has high time complexity, so it is not appropriate for enormous data set. Numerous approaches have been suggested to enhance the efficiency of k-means clustering technique as to discover healthier initial centroids in addition to further accurate clusters with less computational time. The regions obtained by K-means algorithm are also not close to human perception regions and has high computational time.

The time complexity of the K-means algorithm is O (n k d t) that depends on the variables like n which is the size of the data elements in the database, k is the number of clusters of data, d is the dimensionality of the data objects and t is time required to execute the k-means algorithm. Apart from several proposed methodology, this paper proposed a novel technique as to decrease the computational time by minimizing the size of the data objects n. This methodology is known as G-Means clustering algorithm.

1.3. The Organization of Paper

A brief discussion is of data mining, clustering technique and K-means clustering technique is given in this section. Section 2 briefly gives the literature survey of different clustering techniques and other variants of K-Means algorithm. The suggested improved K-means clustering technique discussed in section 3 briefly. Section 4 analysis the experimental results of the proposed methodology on the image dataset and section 5 concludes the proposed Improved k-means Clustering Algorithm with high performance and less computational time.

2. Literature Survey

Clustering problems ascend in numerous diverse applications, such as data mining and knowledge data discovery [12], information compression and vector quantization [13], and pattern recognition and pattern categorization [10]. The concept of a worthy cluster hinge on the application and there are various techniques for defining clusters subject to various standards, both ad hoc and efficient. These entails methodologies grounded on splitting and integration such as ISODATA [14], randomized methodologies such as CLARA [16], CLARANS [19], and procedures depending on neural networks [17], and approaches intended to measure enormous databases, including DBSCAN [11], BIRCH [20], and Scale KM [8]. For additional facts regarding clustering and clustering algorithms, see [16], [9], [15], [14]. An asymptotically proficient estimation for the k-means clustering problem has been suggested by Matousek [18], but the huge continuous features recommend that it is not a good applicant for practical application. Bradley et al. [8] have presented how to measure k-means clustering to tremendous data sets by means of sampling and pruning. Note that Lloyd's algorithm does not identify the primary assignment of centers.

A. M. Fahimetal [23] proposed an improved procedure to place data points in to the appropriate cluster. In [23], the author proposed an approach that requires less computational time compared to the traditional k-means which is computationally expensive. But initial centroids are still selected arbitrarily in this method. Chen Zhang et al.in [25] suggested a new algorithm for selection of initial centroids in k-means clustering that evaded the arbitrary selection of cluster seeds and also improve performance of k-means. A. Bhattacharya et al. [26] presented a heuristic clustering algorithm which was named Divisive Correlation Clustering Algorithm (DCCA) for grouping data objects. The method generates a cluster of dataset that is deprived to consider the initial centroids and the significant number of clusters k. The time complexity of this procedure is too high. Numerous approaches are recommended to augment the effectiveness and accuracy of k-means clustering algorithm and advised that enhanced initial centroid estimation can lead to accurate and effectual cluster groups with improved time complexity.

Kathiresan V. et al.in [27] recommended an procedure to choose healthier initial centroids depending on the Z-Score Ranking technique. In this procedure, the author suggested to estimate the Z-Score of every point in the dataset and then sorting the points based on Z-Score values. Madhu et al. suggested an enriched technique to obtain improved initial centroids [21]. This procedure primarily examines if the data point comprises of negative information. If it encompasses, then are transformed to positive data by means of traditional adaptation. The data points are sorted depending on the distance from the source and separated into k number of chosen clusters or equivalent groups. K. A. Abdul Nazeer et al recommended an iterative process to select initial centroids [22].The system mainly works in two sequential phases. In first phase k initial centroids are chosen depending on the comparative distance so for every data points. In the second phase, the clusters are computed depending on the distance of every point from the initial centroids.

In [5], researchers have introduced an enhanced K-Means algorithm to improve the time complexity using uniform data. They make clusters in two phases. In phase one, they find initial clusters on similarity basis, while in the second phase, they finalize the clusters. Similarly, in [6] all the requirements, advantages and disadvantages of K-Means clustering are discussed. Rather than numerous runs with random selections, Bradley and Fayyad [4] proposed an alternative procedure that employed numerous primary K-Means iterations to offer the initial points for subsequent runs of the algorithm. This augments computational overhead to the complete execution, however it offers an enhanced minimized error. In [28] a novel technique is used to obtain a weighted average score of dataset. An identical process is employed to obtain the rank score by averaging the attribute of every data point that produces initial centroids and monitor the data dissemination of the specified group.

3. Proposed Methodology

The main intuition behind the proposed approach is to organize the patterns in a suitable data structure such that one can discover all the patterns, which are neighbor to specified prototypes. This proposed methodology as to minimize the computational time mainly focus on the size of the data objects that differs from the other alternatives of existing K-means clustering Algorithm. The performance efficiency of clustering technique for proposed Improved k-

means Clustering Algorithm is obtained with the help of some natural environmental image datasets with different cluster value i.e. k values. As in any image the number of distinct values will be a maximum of 256 values.

3.1. Existing K-means Algorithm

K-means is a most prevalent unsupervised learning procedure and segregating way for clustering. The segregating cluster technique defines complete clusters in a single shot. For partitioning technique, a universal principle is the widely used and its optimization motivated the complete process to generate a single stage dissection of information. The rudimentary notion of K-means technique is to categorize the data items D into k diverse clusters where D is the data set, k is the number of required clusters. The technique comprises of two elementary stages [7]. The initial stage is to choose the initial centroids for every cluster arbitrarily. The second and finishing stage is to consider every point in the database and allocate it to the adjacent centroids [7].

The Euclidean Distance is employed to find the distance amongst the data points. When a novel point is allotted to the cluster then the cluster mean is instantaneously updated by estimating the average of points in the cluster [3]. After all the points are encompassed in certain clusters initial grouping is done. Then every data item is allotted to a cluster grounded on intimacy with the cluster center where intimacy is specified using Euclidean distance. This technique of allotting a points to a cluster and updating their cluster centroids endures until the convergence principles is encountered or the centroids does not vary amongst two successive generations. Once this condition is encountered where centroids don't change any longer, the algorithm stops. With the suitable amount of initial cluster, K-means algorithm can professionally cluster the data. Further, the appropriate number of items means the system can resourcefully scrutinize information in few generations. Specifying the number of preferred clusters, the partitioning algorithm finds all the k clusters at once, such a way that the sum of distances over the objects to their cluster centers is negligible. Furthermore, for the clustering outcome to be precise, apart from the low intra-cluster distance and the high inter-cluster distance a well parted cluster is obtained. K-means is a distinctive partitioning technique that depends on the idea of cluster centers and a point in the data space that typically not surviving in the dataset themselves in a cluster.

The Pseudo code for k-means partitioning clustering algorithm is given below [3].

Input: D =d1, d2, ...dn //set of n data items, k =the number of desired clusters,

Output: a set of k clusters.

Steps:

1. Arbitrarily select k items from n as primary cluster centers.
2. Select distant and diverse centroids for each of the preferred group of K clusters.
3. Consider each object and assign to one of the k clusters of the specified group and associate its distance to all the centroids of the K clusters.
4. Grounded on the estimated distance, the new data point is added to the cluster whose centroid is adjacent to the data points.
5. Update, the cluster means by re calculating centroids after each assignment of the items for every cluster k.
6. Until no alteration in a cluster means /min error E is attained.
7. This is an iterative method and continuously updated.

Normally, the K-Means technique has the following characteristics [29]: (i) it is efficient in clustering huge data sets (giga bytes or even terabytes), (ii) it frequently stops at a local optimum, (iii) the cluster has spherical shapes, (iv) it is sensitive to noise, distance measures, (v) no recommendation on the initial location of cluster centroids and (vi) no criteria for selecting the number of clusters. This algorithm is sensitive to the selection of the initial clusters to be stated in advance. The disadvantages of K-means algorithm K has to be determined beforehand, the algorithm has the time complexity of O (n d k t), which is very large for large data n and regions obtained are not close to the human eye perception.

3.2. Proposed improved k-means clustering algorithm

A grey level image is considered in this methodology were the pixel varies from 0 to 255 in the spatial domain. Thus the size of the image object is reduced by removing the redundant pixels for the Improved k-means Clustering Algorithm. The main concept of the proposed approach is to maintain the frequency of pixels of the same value in the image. Similar values are stored in array as indices and the total count, before the main clustering process.

1. In this methodology, at first only the distinct values of the image are considered from redundant pixels present in the image. Then the random numbers are generated as the initial cluster centroids.
2. In the initial stage each cluster contains only one item. As the process continues cluster centers change for each iteration. This process stops when cluster centers do not change.
3. Clustering is done only on the distinct pixels obtained from the image. Each value is assigned to the nearest cluster, which is determined by the Euclidean distance amongst the center and the input data.
4. Cluster centers are updated in each iteration in the following way:
    a. The sum of all the values in each cluster divided by the total number of pixels. While calculating the sum and the total number of pixels in each cluster, the frequency of each pixel in every cluster is considered.
5. The updated centroids are now considered as new centroids. The procedure is repeated until the centroids do not change or clusters does not change.
6. After clustering, the new image data obtained is generated the output image. The output image consists of k no of regions where each region is represented by a distinct color.

4. Experimental results and its Analysis

The Experimental Analysis of the proposed improved k-means Clustering Algorithm for image Segmentation is evaluated with the help of three different types of images in this paper. They are 1000 Corel database images used in this methodology. The performance evaluation of the proposed improved k-means Clustering Algorithm is compared with existing iterative based on conventional k-means algorithm.

Consider the images of flowers and airplane, as shown in Fig.1, Fig. 2 and Fig. 3 as the input image for the proposed methodology where the number of clusters considered for this method are k=2,4,8, and 16.

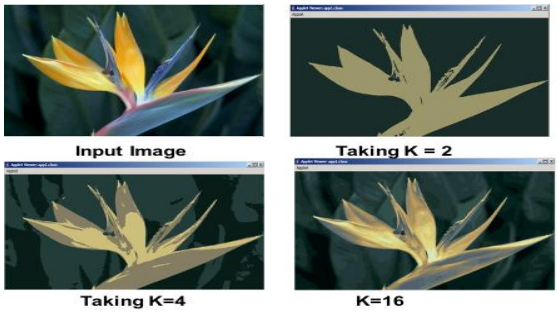

Fig 1: Results of Paradise flower image

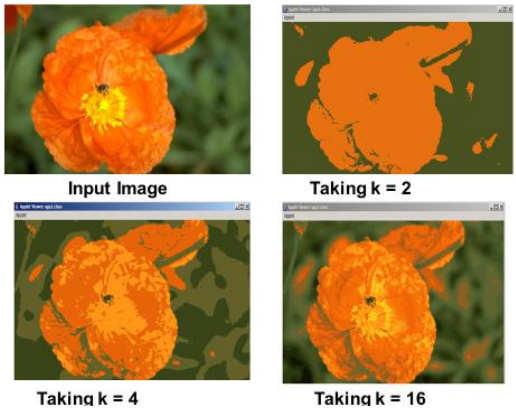

Fig 2: Results of Marigold Flower Image

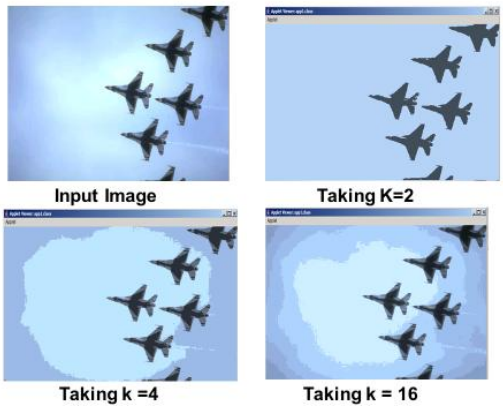

Fig 3: Results of airplane image

Thus, by maintaining the frequency of each distinct pixel in the image, we reduce the actual number of pixels to be clustered to a total number of distinct values, which is a maximum of 256 values. This number of pixels scanned to a large extent. For example, in an 800x600 image there are 480000 pixels. Intended of clustering 480000 pixels we cluster only distinct values of pixels in the image i.e. a maximum of 256 values. Thus significantly reducing the scanning time in each iteration.

5. Conclusion

The k-means technique is extensively employed for clustering a huge set of data points. The traditional k-means technique does not constantly ensure better outcomes as the accuracy of the ultimate clusters. The computational complexity of the standard k-means algorithm is high than our proposed k-means algorithm. This paper offers a modified k-means algorithm which reduces the size of the data objects by reducing the redundant elements in the database to assign data into clusters. The efficiency of this methodology is analyzed with the help of image objects. This algorithm overcomes the disadvantages of the known partitioned K-means algorithms and at the same time enhances the clustering performance. The output results, i.e., obtained regions obtained from this proposed methodology are closed to human perception regions (objects) and takes less computational time compared to existing clustering k-means algorithms significantly.

6. References

[1] Mehmed Kantardzic. Data Mining-Concepts, Models, Methods, and Algorithms. Copyright by IEEE Press, 2002, P.171-189.

[2] Hand, D., H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge: MA, 2001.

[3] Han J. Characteristic Rules. In [3] Margaret H. Dunham, Data Mining-Introductory and Advanced Concepts, Pearson Education, 2006.

[4] D. Foti, D. Lipari, C. Pizzuti, and D. Talia. Scalable parallel clustering for data mining on multi computers. Proceedings of the 15th IPDPS 2000 Workshops on Parallel and Distributed Processing, pages 390–398, 2000.

[5] Napoleon, D. and P.G. Lakshmi, 2010. "An Efficient K-means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points," in Trendz in Information Sciences and Computing (TlSC), Chennai.

[6] Master, C.P. and X.G. Professor, 2011. "A Brief Study on Clustering Methods Based on the K-means algorithm," in 2011 International Conference on E-Business and E-Government (ICEE), Shanghai, China ..

[7] J. Hanand, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, San Diego, 2001.

[8] P.S. Bradley, U. Fayyad, and C. Reina, "Scaling Clustering Algorithms to Large Databases", Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining, pp. 9-15, 1998.

[9] V. Capoyleas, G. Rote, and G. Woeginger, "Geometric Clustering's", J. Algorithms, vol. 12, pp. 341-356, 1991.

[10] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973.

[11] M. Ester, H. Kriegel, and X. Xu, "A Database Interface for Clustering in Large Spatial Databases,º Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD-95), pp. 94-99, 1995.

[12] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[13] A. Gersho and R.M. Gray, Vector Quantization and Signal Compression. Boston: Kluwer Academic, 1992.

[14] A.K. Jain, P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[15] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

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

[17] T. Kohonen, Self-Organization and Associative Memory, third ed. New York: Springer-Verlag, 1989.

[18] J. Matousek, "On Approximate Geometric k-clustering", Discrete and Computational Geometry, vol. 24, pp. 61-84, 2000.

[19] R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", Proc. 20th Int'l Conf. Very Large Databases, pp. 144-155, Sept. 1994.

[20] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: A New Data Clustering Algorithm and Its Applications", Data Mining and Knowledge Discovery, vol. 1, no. 2, pp. 141-182, 1997.

[21] Madhu Yedla, Srinivasa Rao Pathakota, T. M. Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 1(2), 2010, 121-125.

[22] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering 2009 Vol IWCE2009, July1-3, 2009, London, U.K.

[23] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):16261633, 2006.

[24] K. A. Abdul Nazeer, S. D. Madhu Kumar and M. P. Sebastian, "Enhancing the k-means clustering algorithm by using a O (nlogn) heuristic method for finding better initial centroids," Second International Conference on Emerging Applications of Information Technology, 2011.

[25] Chen Zhang and Shixiong Xia, "K-means clustering algorithm with improved initial centroids," Second International Workshop on Knowledge Discovery and Data Mining pp.790-792, 2009.

[26] A Bhattacharya and R. k. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," bioinformatics, Vol. 24, pp. 1359-1366, 2008.

[27] Kathiresan V. andDr. P Sumanthi, "An Efficient Clustering Algorithm based on Z-Score Ranking method," International Conference on Computer Communication and Informatics (ICCCI-2012), Jan.10-12, 2012.

[28] Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md. Nasim Akhtar, "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average", 2012 7th International Conference on Electrical and Computer Engineering, 20-22 December, 2012.

[29] N. KarthikeyaniVisalakshi and J. Suguna, "K-Means Clustering using Max-min Distance Measure", the 28th North American Fuzzy Information Processing Society Annual Conference, USA-June 14-17, 2009, IEEE.