A Review of Image Data Clustering Techniques

Ashwini Gulhane, Prashant L. Paikrao, D. S. Chaudhari

Abstract

In order to the find the close association between the density of data points, in the given data set of pixels of an image, clustering provides an easy analysis and proper validation. In this paper various clustering techniques along with some clustering algorithms are described. Further k-means algorithm, its limitations and a new approach of clustering called as M-step clustering that may overcomes these limitations of k-means is included.

Keywords: M-step clustering, k-means clustering.

I. INTRODUCTION

In 1854, in London during cholera outbreak, John Snow had plotted the diseased reported cases using a special map. After the creation of the map it was observed that there was close association between the density of disease cases and a single well located at a central street. With the above map it is very simple to identify association between phenomena, which is very difficult to analyze in actual. This was the first known application of clustering analysis for many researchers [6].

Since then cluster analysis has been widely used in many fields such as statics, marketing, engineering, medical and other social sciences, for clustering large data sets into natural groups, the number of clustering algorithms had been proposed for performing clustering task. k-means is the most popular and commonly used clustering algorithm. k-means was first proposed over 50 years ago. It was independently discovered in many scientific fields, in 1956 by Steinhaus, in 1957 by Lloyd, by Ball and Hall in 1965 and by Macqueen in 1967. A simple step for executing k-means algorithm was given by Jain and Dubes in 1988. Elkan's algorithm gives an efficient method for high-dimensional k-means clustering. Then Greg Hamerly gives another modified and simplified Elkans's algorithm [3]. k-means is most widely used clustering algorithm because it is easy to implement and simple to understand, but its output depends on the selection of the initial partition(s).In one of the study a new method for determining number of clusters in k-means clustering is proposed [8].

This method overcomes the limitation of having to indicate the number of cluster by using validity measure based on the intra-cluster distance measurement, but the additional time overhead associated with calculation of number of clusters and their cluster centers is present.

However the prior knowledge of number of clusters is must for k-means clustering. It also involves random initialization of cluster centers at the beginning of algorithm; therefore the convergence of the algorithm to same output is not guaranteed on every run. The proposed approach of clustering called as M-Steps clustering may be overcomes both the drawback of k-means clustering i.e. need of apriori of number of clusters and accuracy to produce same output for same input at various instances.

The next section of this paper details the steps of clustering, gives the broad classification of clustering techniques and also describes some clustering algorithms, and in the last section conclusion remarks are given regarding existing techniques.

II. AN OVERVIEW OF IMAGE DATA CLUSTERING

Clustering is the task of assigning set of objects into groups called as clusters so that the objects in one cluster are more similar than the objects in other cluster. Clustering itself is not one specific algorithm but it is task which can be performed by various algorithms that differs from each other in their methods of computing/finding the cluster. Clustering is process of grouping similar image pixels according to some property into one cluster so that the resulting output cluster shows high intra-cluster similarities and low inter-cluster similarities. Clustering process is an unsupervised classification of data points into groups or clusters [1] [4].

A) Clustering task steps

Clustering is a task which involves number of stages. Typical clustering process used the following steps [1] [6]:

i) Characteristics Representation: In this step the type, dimensions and features of available data are checked. For

Manuscript Details received on February 29, 2012

Ashwini U. Gulhane, Eectronics and Telecommunication Dept. Amravati University, GCOE Amravati, India. (email: ashwini.gulhane3113@gmail.com)

Prashant L. Paikrao, B.E. degree in Industrial Electronics from Dr. BAM University, Aurangabad and M. Tech. degree in Electronics from SGGSIE&T Nanded, India. (email: plpaikrao@gmail.com)

Devendra S. Chaudhari, Electronics and Telecommunication Dept, BE, ME, from Marathwada University, Aurangabad and PhD from Indian Institute of Technology, Bombay, Mumbai, India. (email:chaudhari.devendra@gcoea.ac.in)

that it involves processes such as feature selection or feature extraction.

ii) Similarity Measurement: In this step similarity among data points is measured. Generally various distance measurement methods viz. Euclidian Distance, Mean Square etc. are used to measure similarity between different data points.

iii) Collecting the data points: In this step the data points are grouped together into clusters, based on similarity measures obtained from the previous step.

iv) Data abstraction: In this process data is represented with compact description of individual cluster and further more the cluster prototype i.e. the centroid of the cluster is calculated and used as final representation [1].

v) Output Validation: This is an important stage in clustering process. In this step the resulting output clusters are observed to determine whether the resulted output is meaningful or not. It can be done by various methods; either it compares resulted output with apriori structure, or check whether the structure is intrinsically appropriate for data sets or not, or compares two derived outputs with each other and measure their exclusive merits.

Clustering is useful in number of application as it clusters the raw data and find out the hidden features / patterns in the database [5]. So it is widely used in image processing, data mining, image retrieval, pattern recognition, image segmentation and so on.

B) Classification of Clustering Techniques

Clustering algorithms are broadly classified as i) Hierarchical clustering ii) Partitional clustering [2].The following figure shows the different techniques of clustering,

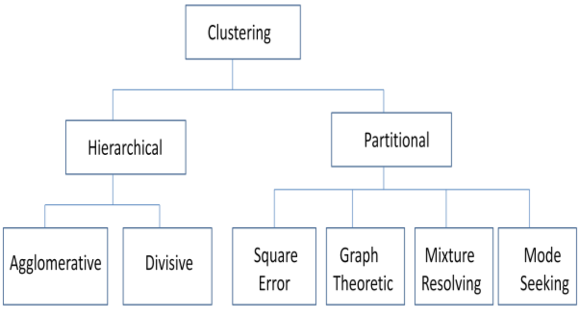

Figure 1 Classification of clustering techniques

A) Hierarchical Clustering

Hierarchical clustering, clusters the given image, based on the concept of pixel being more closely related to nearby pixels than the pixels which are farther i.e. these algorithms groups the pixels into cluster based on their distances. Hierarchical clustering represents data into a tree structure form. In which the whole data set is represented by root node and the individual data points are represented by leaf node. The intermediate nodes in a tree structure represent the similarity among the pixel data points. At different stages in structure different cluster will form thus expected clustering will obtained by cutting cluster structure at respective stages [3]. The Figure.2 shows the hierarchical tree structure form. In this clustering technique numbers of algorithms are proposed based on the method by which distances are computed. Hierarchical clustering mainly classified as agglomerative hierarchical clustering, starts with single elements in distinct patterns and merges it until the stopping criterion reached. The cluster divisive hierarchical clustering starts with considering complete data set as single cluster and splitting it into clusters. Along with the distance function the linkage criterion is also an important factor in hierarchical clustering. Since, in this method cluster consist of multiple elements, so multiple elements are involved in computing distances. The most commonly used linkage criterions are single-linkage and complete-linkage.

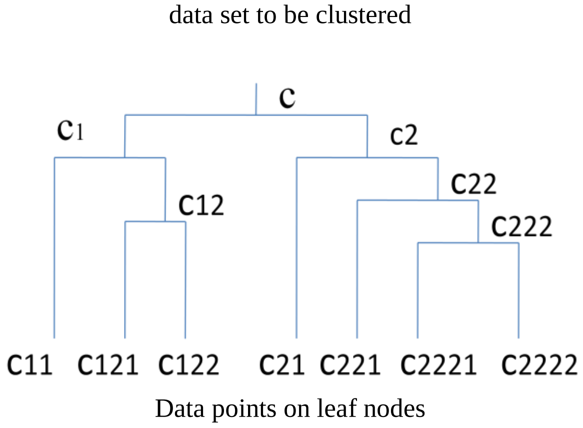

Figure 2 Hierarchical tree structure of data set

B) Partitional clustering

Partitional Clustering algorithms separate the pixels or data points into number of partitions. These partitions are referred as clusters. The partitional clustering organizes data into single partition instead of representing data into nested structure like hierarchical clustering. Partitional clustering is more useful for large data set in which it is difficult to represent data in tree structure. The partitional clustering is classified as square error clustering, Graph theoretic clustering, mixture resolving clustering and mode seeking clustering [1].

A) Square error Clustering

In partitional clustering the data points are randomly initialize and assign it to the predefined number of clusters depend on similarity between data points and cluster center until convergence criterion has been reached. The most frequently used convergence criterion in partitional criterion is squared error algorithms. The main advantage of square error algorithms is that it works well with isolated and compact clusters [1].

B) Graph theoretic clustering

Graph Theoretic Clustering algorithm is divisive clustering algorithm. This algorithm first forms the minimal spanning tree (MST) for the given data. Then it obtained clusters by deleting edges of largest length from the minimal spanning tree structure [1].

C) Mixture-Resolving Clustering

Clustering can be understood by studying density distribution functions. In the mixture resolving algorithm the parametric distribution function like Gaussian distribution are used and the vectors of component density are form. These vectors are grouped together iteratively based on maximum likelihood estimation to form the clusters [1].

D) Mode-Seeking Clustering

In non-parametric technique the algorithms are developed inspired by the Parzen window approach. It forms clusters by creating bins with large counts in multidimensional histogram of the input mixture patterns. This approach considered as mode seeking approach [1].

C) Clustering algorithms

Based on the above methods there are number of clustering algorithm available. The clustering algorithms included in this paper are k-means clustering, N-cut clustering, Mean shift clustering etc

A) k-means Clustering Algorithm

k-means is commonly used simplest algorithm which employs the square error criterion. In this algorithm the number of partitions is initially defined. The cluster centers are randomly initialized for predefined number of clusters. Each data point is then assigned to one of the nearest cluster. The cluster centers are then re-estimated which and new centroid is calculated. This process is repeated until the convergence has been reached or until no significant change occurs in cluster center [3]. k-means is easy to implement and its time complexity is less. But the output of k-means algorithm depends on selection of predefined clusters. If the initial number of clusters is not properly chosen then the output of algorithm may converge to false cluster locations and completely different clustering result [1] [7].

B) N-cut Clustering Algorithm

N-cut method is a hierarchical divisive clustering process that represents the clusters in the tree structure form. This method organizes the nodes of tree into groups or cluster, such that the similarity between nodes in one cluster is high and similarity between the group is low. The output of this method results in more than two clusters, so this method is useful for the application where the data should be required to cluster into multiple clusters. In each step the subgraph with maximum number of nodes is further divided into clusters until stopping criterion has been reached [3].

C) Mean Shift Clustering Algorithm

The Mean Shift algorithm clusters the given data set by associating each point with a peak of the data set’s probability density. For each point, Mean Shift computes its associated peak by defining a spherical window of radius ‘r’ at the data point and then compute the mean of the points that lie within the window. The algorithm then shifts the window to the mean and repeats until convergence, i.e., until the shift is less than the threshold. After every iteration the window will shift to a more densely populated portion of the data set until a peak is reached, where the data is equally distributed in the window [9].

III. CONCLUSION

In this paper various techniques and parameter are discussed for clustering methods. The shortcoming of k-means clustering algorithm to find optimal k value and initial centroid for each cluster is discussed and validity of N-cut algorithm, mean shift algorithm and M-step algorithm is discussed for overcoming the shortcoming of k-means algorithm.

References

[1] S. Anitha Elavarasi, Dr. J. Akilandeswari, Dr. B. Sathiyabhama,” A survey on partition clustering algorithms”, International Journal of Enterprise Computing and Business SystemInternational Systems, vol. 1, pp. 1-13, 2011.

[2] Monika Jain, Dr. S.K.Singh,” A Survey On: Content Based Image Retrieval Systems Using Clustering Techniques For Large Data sets”, International Journal of Managing Information Technology (IJMIT) Vol.3, No.4, November 2011, pp. 23-39.

[3] Juntao Wang, Xiaolong Su,” An improved K-Means clustering algorithm”, IEEE proceeding, pp. 44-46, 2011.

[4] Harikrishna Narasimhan, Purushothaman Ramraj” Contribution-Based Clustering Algorithm for Content-Based Image Retrieval”, 2010 5th International Conference on Industrial and Information Systems, ICIIS 2010, pp. 442-447.

[5] Shi Na, Liu Xumin, Guan yong,” Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm”, Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 63-67, 2010.

[6] Wenbing Tao, Hai Jin and Yimin Zhang,” Color Image Segmentation Based on Mean Shift and Normalized Cuts”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 37, NO. 5, OCTOBER 2007, pp. 1382-1389.

[7] Periklis Andritsos,” Data Clustering Techniques”, March 11, 2002.

[8] A. Jain , M. Murty, and P. Flynn " Data clustering: a review.," ACM Computing Surveys, vol. 31,pp. 264-323,1999.

[9] Siddheswar Ray, Rose H. Turi,” Determination of Number of Clusters in K-Means Clustering andApplication in Colour Image Segmentation”, IEEE proceeding.

Ashwini U. Gulhane received the B.E. degree in Electronics Engineering from the Rashtrasant Tukadoji Maharaj Nagpur University in 2010, and she is currently pursuing the M. Tech. degree in Electronic System and Communication (ESC) at Government College of Engineering, Amravati.

Prashant L. Paikrao received the B.E. degree in Industrial Electronics from Dr. BAM University, Aurangabad in 2003 and the M. Tech. degree in Electronics from SGGSIE&T, Nanded in 2006. He is working as Assistant Professor, Electronics and Telecommunication Engineering Department, Government College of Engineering Amravati. He has attended An International Workshop on Global ICT Standardization Forum for India (AICTE Delhi & CTIF Denmark) at Sinhgadh Institute of Technology, Lonawala, Pune and a workshop on ECG Analysis and Interpretation conducted by Prof. P. W. Macfarlane, Glasgow, Scotland. He has recently published the papers in conference on 'Filtering Audio Signal by using Blackfin BF533EZ kit lite evaluation board and visual DSP++' and 'Project Aura: Towards Acquiescent Pervasive Computing' in National Level Technical Colloquium "Technozest-2K11", at AVCOE, Sangamner on February 23rd, 2011. He is a member of the ISTE and the IETE.

Devendra S. Chaudhari obtained BE, ME, from Marathwada University, Aurangabad and PhD from Indian Institute of Technology, Bombay, Mumbai. He has been engaged in teaching, research for period of about 25 years and worked on DST-SERC sponsored Fast Track Project for Young Scientists. He has worked as Head Electronics and Telecommunication, Instrumentation, Electrical, Research and incharge Principal at Government Engineering Colleges. Presently he is working as Head, Department of Electronics and Telecommunication Engineering at Government College of Engineering, Amravati.

Dr. Chaudhari published research papers and presented papers in international conferences abroad at Seattle, USA and Austria, Europe. He worked as Chairman / Expert Member on different committees of All India Council for Technical Education, Directorate of Technical Education for Approval, Graduation, Inspection, Variation of Intake of diploma and degree Engineering Institutions. As a university recognized PhD research supervisor in Electronics and Computer Science Engineering he has been supervising research work since 2001. One research scholar received PhD under his supervision.

He has worked as Chairman / Member on different university and college level committees like Examination, Academic, Senate, Board of Studies, etc. he chaired one of the Technical sessions of International Conference held at Nagpur. He is fellow of IE, IETE and life member of ISTE, BMESI and member of IEEE (2007). He is recipient of Best Engineering College Teacher Award of ISTE, New Delhi, Gold Medal Award of IETE, New Delhi, Engineering Achievement Award of IE (I), Nashik. He has organized various Continuing Education Programmes and delivered Expert Lectures on research at different places. He has also worked as ISTE Visiting Professor and visiting faculty member at Asian Institute of Technology, Bangkok, Thailand. His present research and teaching interests are in the field of Biomedical Engineering, Digital Signal Processing and Analogue Integrated Circuits.