Analysis of Different Clustering Algorithms on Image Databases

Stuti Mehla and Ashok Kajal

1M.Tech Research Scholar, 2Assistant Professor
Ambala College of Engineering & Applied Research, Devsthali, India
E-mail:stuti21mehla@gmail.com

Abstract
When we apply Image Retrieval techniques to large image Databases. It provides restriction of search space to provide adequate response time. This restriction can be done minimized by using Clustering technique to partition the image dataset into subspaces of similar elements. In this article we will apply different clustering algorithms on large image database and then evaluate and analyse the performance of these algorithms to determine which algorithm is best for image retrieval.

Keywords: Clustering, Image Retrieval, Database.

Introduction
What is Clustering?
Clustering is the process in which set of observations is divided into subsets called clusters, so that observations in same cluster are similar in some sense. Clustering is method of unsupervised learning, used in many fields, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics.

Cluster analysis itself is not an algorithm but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space and multivariate normal distributions. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery that knowledge that fires.

Result of cluster analysis shown in three different squares.

The Goals of Clustering
The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

Image Clustering
The goal of image clustering is to find out a mapping of query image into classes called clusters such that these set of clusters provide same information about query image as the entire set collection. Because searching large databases of images is a challenging task. We calculate the similarity between the query image and all the images in the database and rank the images by sorting their similarities. One problem with this exhaustive search approach is that it does not scale up for large databases. The retrieval time for exhaustive search is the sum of two terms: Tsim and Tsort. Tsim is the time to calculate the similarity between the query image and every image in the database, and Tsort is the time to rank all the images in the database according to their similarity to the query.

\[ T_{exhaustive} = nT_{sim} + O(n\log n) \]

Where \( n \) is the number of images in the database, \( T_{sim} \) is the time to calculate the similarity between two images, and \( O(n\log n) \) is the time to sort \( n \) elements. When the images in the database are clustered, the retrieval time is the sum of three terms, the time to calculate the similarity between the query and the cluster centers, the time to calculate the similarity between the query and the images in the nearest clusters and the time to rank these images. Therefore the total search time is:

\[ T_{cluster} = kT_{sim} + lT_{sim} + O(l\log l) \]

Here \( k \) is the number of clusters, \( l \) is the number of images in the clusters nearest to the query. Since \( k<<n \) and
Image clustering is important for efficient search and retrieval in large image databases. [1]

Possible applications
Clustering algorithms can be applied in many fields, for instance:

Marketing: finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records;

Biology: classification of plants and animals given their features.

Libraries: book ordering;

Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;

City-planning: identifying groups of houses according to their house type, value and geographical location;

Earthquake studies: clustering observed earthquake epicentres to identify dangerous zones;

WWW: document classification; clustering weblog data to discover groups of similar access patterns.

Steps in clustering:
Feature Extraction:
Feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from original ones. Both are very crucial to effectiveness of clustering applications. [9]

Clustering algorithm design or selection: The step is usually combined with the selection of a corresponding proximity measure and the construction of a Criterion function. Patterns are grouped according to whether they resemble each other. Obviously, the proximity measure directly affects the formation of resulting clusters. Almost all clustering algorithms are explicitly or implicitly connected to some definition of proximity measure [9].

Cluster validation: Given a data set, each clustering algorithm can always generate a division, no matter whether the structure exists or not. Moreover, different approaches usually lead to different clusters, and even for the same algorithm, parameter identification or presentation order of input patterns may affect the final result. Therefore effective evaluation standards and criteria are important to provide users with a degree of confidence for the clustering results derived from the used algorithms [9].

Results interpretation. The ultimate goal of clustering is to provide users with meaningful insights from the original data, so that they can effectively solve the problems encountered [9]

Literature Survey
Different approaches have been used in literature for clustering of database.

Hierarchical Clustering algorithm: In this paper I have studied the hierarchical clustering algorithm which is used for image retrieval. In Hieratical clustering we create a hierarchy of clusters which may be represented in a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. Algorithms for hierarchical clustering are generally either agglomerative starts at the leaves and successively merges clusters together; or divisive in which one starts at the root and recursively splits the clusters [1].

K-Means Clustering: In this paper I have studied k-means algorithm and a modified version have also discussed. In this paper distance matrices are improved. In k-means procedure follows a simple and easy way to classify a given data set through a certain number of clusters is to define k centroids, one for each cluster. These centroids should be placed in a way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. [3]

In this paper I have studied comparative analysis of Hierarchial, K-means, PAM, CLARA algorithm. Basics about PAM and CLARA are described below PAM: The pam is a clustering algorithm related to the means algorithm and the medoidshift algorithm. Both the k-means and PAM algorithms are partitioned (breaking the dataset up into groups). In contrast to the k-means algorithm, k-medics chooses data points as centres. PAM is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. A useful tool for determining k is the
silhouette. It is more robust to noise and outliers as compared
to k-means. A medoid can be defined as the object of a cluster,
whose average dissimilarity to all the objects in the cluster is
minimal. CLARA (Clustering large applications) improves
time complexity of PAM by using samples of dataset. The
basic idea is that it applies PAM to a sample of underlying
database and then uses medics found as the medoids for the
complete clustering. Each item from complete database is then
assigned to the cluster with medoid to which it is closest. Because, of sampling CLARA is more efficient then
PAM. [6]

In this paper improved fuzzy c-means algo is discussed. Fuzzy
c-means algorithm is developed because in Hard
clustering every data element is related to only one cluster. If
we want that a data element should related to more than one
cluster than we use fuzzy logic. [7]

In this paper Image clustering and compression technique
is applied. Color base Image clustering is done. [8]

Problem Formulation
We will analyze the performance of clustering algorithms
described above.

\{I1, I2, I3,………In\} is an array of images in image
database.

\{A1, A2, A3,………Am\} is an array of clustering
algorithms.

After applying clustering algorithms a\in A

\{i1,……ik\}\in I we will analyse the performance of clustering
algorithms by applying them on image databases.

Proposed Model
In the proposed model we will apply clustering algorithms on
image database which are discussed above and then graph is
plotted with the help of MATLAB and after this performance
is evaluated. The steps are pictorially represented

Conclusion & Future Scope
We are concerning with performance analysis of clustering
algorithms on image database.

In future work we can analyze the algorithms on
relational database, image and pattern recognition.

We can also improve efficiency of any clustering
algorithm.

References

[1] “Hierarchical clustering algorithm for fast image retrieval” Santhana Krishnamachari Mohamed Abdel-
Mottaleb Philips Research 345 Scarborough Road
Briarcliff Manor, NY 10510 {Sgk, msa}@philabs.research.philips.com2

Man, Cybern., vol. 30, no. 6, pp 835–845, 2000

means algorithm with A distance based on cluster


K-means clustering For relational databases,” IEEE
Trans. Know. Data Eng., vol. 16, no. 8, pp. 909–921,
Aug. 2004

Jain * Department of Computer Science and
Engineering, Michigan State University, East Lansing,
Michigan 48824, USA Department of Brain and
Cognitive Engineering, Korea University, Anam-dong,
Seoul, 136-713, Korea 0167-8655/$ - see front matter
2009

Categorization in Image Retrieval” Thomas Kaster,
Volker Wendt, and Gerhard Sagerer, Springer-Verlag
Berlin Heidelberg, 2003

means clustering algorithms,” IEEE Trans. Fuzzy

[8] An Algorithm for Image Clustering and Compression,
Mertin KAYA, Turk Demirdokum Fabrika A.S
Bozuyuk Bieleck-TURKEY, VOL13, NO-1, 2005

[9] Survey of Clustering Algorithms Rui Xu, Student
Member, IEEE and Donald Wunsch II, Fellow, IEEE
TRANSACTIONS ON NEURAL NETWORKS,
VOL. 16, NO. 3, MAY 2005