International Journal of Multidisciplinary Research and Development

IJMRD 2015; 2(3): 801-805
www.allsubjectjournal.com
Received: 12-03-2015
Accepted: 30-03-2015
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

Tamije Selvy P
CSE, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

Ramya S
CSE, Sri Krishna College Of Technology, Coimbatore, Tamil Nadu, India

Correspondence:
Ramya S
CSE, Sri Krishna College Of Technology, Coimbatore, Tamil Nadu, India

Evaluation of clustering methods for mining duplicate image groups

Tamije Selvy P, Ramya S

Abstract
The main task of clustering method is grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in another group. A general technique for data analysis are used in several fields, including Pattern recognition, machine learning, image analysis, information retrieval. Various Clustering methods used in Data Mining for collecting near duplicate images. The goal is to identify the duplicate images groups using different clustering techniques. The chapter begins by providing distance measures and criteria that are used for determining whether two image groups are similar or dissimilar. Then the clustering methods are presented, divided into: K-means Clustering, hierarchical, Centroid based clustering, a distribution based clustering. Social media is one of the main application, many users share their personal photos with friends in social media sharing websites such as Flickrs ,Face book and Picasa .People like to follow and forward their favorite images which are the main sources of near duplicate images.

Keywords: Clustering, Duplicate image groups, Hierarchical clustering method, K-means clustering method, Distribution Based Method, DBSCAN.

1. Introduction

Clustering technique is most fundamental tasks in Data Mining. Clustering is used for unsupervised learning and the goal of clustering method is descriptive and new groups are interest in themselves. Groups of data are clustered into subsets in a manner that similar images are grouped together, while different images belong to different groups. In clustering method, social networks are used for clustering the image group's [1] .Social networks are very well-liked for young people to gain information, especially with the development of Smartphone and internet. People tend to forward, share and follow what they are interested in. Favorite images sharing websites as flickr. The whole images are more than 6 billion. Face book has gathered about one billion users, and uploaded about 0.25 billion images per day. The major problem is to manage the big image data is very challenging for effective indexing and retrieval. Most of the user share large amount of images as their places of interest. For the moment, some of the images are modified, forwarded and copied by other users before being shared in social communities.

Phil bin model images as nodes of the graph and image-to-image similarity as edge between the corresponding nodes. Clustering based approach is adopted to divide the group into smaller groups containing near duplicate image groups. In graph based NDIG detection, the weights of edges are merely measuring by the concurrence of visual words which neglects the context information between images. Moreover, the image-to-image similarity computing is too time-consuming for a large scale dataset.

Hu *et al.* proposed a coherent phrase model for near-duplicate image retrieval. Different from the standard BoW, their model represents every local region using multiple descriptors and enforces the coherency across multiple descriptors [2].Spatial coherent phrase and Feature coherent phrase are designed to represent spatial and feature coherency. They mentioned that near duplicate images retrieval approach was hard to achieve the task of near duplicate image group's detection.

Gao et.al simultaneously utilizes both textual and visual information to estimate images relevance, which is determined with a hyper graph learning approach [4]. They propose an interactive 3-D object retrieval scheme, additionally. Wang et.al obtained relevant images by exploring the image content and the associated tags .A Greedy ordering algorithm which optimizes average diverse precision as the ranking method.

In the present paper a study of various clustering techniques have been made. Section 2 deals

with a study on various clustering method, section 2.1 deals with the study on K-means clustering algorithm, section 2.2 deals with study on Centroid based clustering, section 2.3 deals with the study on hierarchical clustering method, section 2.4 deals with the study on Centroid based clustering, section 2.5 deals with the study on Distribution based clustering methods. Section 2.6 deals with study on DBSCAN, Section 3 deals with results and discussion. Finally the last section 4 concludes the paper.

2. Various Clustering Methods

2.1 K-Means Clustering Method

K-means is an algorithm to classify or group objects based on features/attributes into K number of group. Here,K is positive integer number. Minimizing the sum of squares of distances between cluster centroid and the corresponding cluster data. The main purpose of K-mean clustering is to classify the data. Image Classification has a major role in the field of mining analysis. Cluster analysis is a useful for image classification and object diagnosis. In image processing, various clustering algorithm are used for image classification. K-means algorithm split the given image into different clusters of pixels in the feature space, each of them defined by its center. In the image each pixels is allocated to the nearest cluster. Then the new centers are compute with the original clusters. Repeat the steps until convergence. Mostly we have to to find out the number of clusters K first. Then assumed the centroid for these clusters and also assume random objects as the initial centroids or the first K objects in sequence could also serve as the initial centroids. The K means algorithm in a logical representation: Execute the below steps until convergence. Do the following steps while no object move group [3].

a) First, centroid coordinate is determined (Random assignment).

b) Calculate the distance of each Object pixel to the Centroids.

c) Based on minimum distance we group the objects with the Centroid

Steps:

1. If the number of data < number of cluster then assign each data as centroid
2. Each centroid have a cluster number
3. If the number of data> number of cluster then for each data we have to calculate minimum distance measure of all centroid for each cluster.
4. Location of centroid is not determined correctly, we need to adjust the centroid location based on current updated data.
5. Then assign all data to new centroid. Repeat this steps until no data is moving to another cluster.

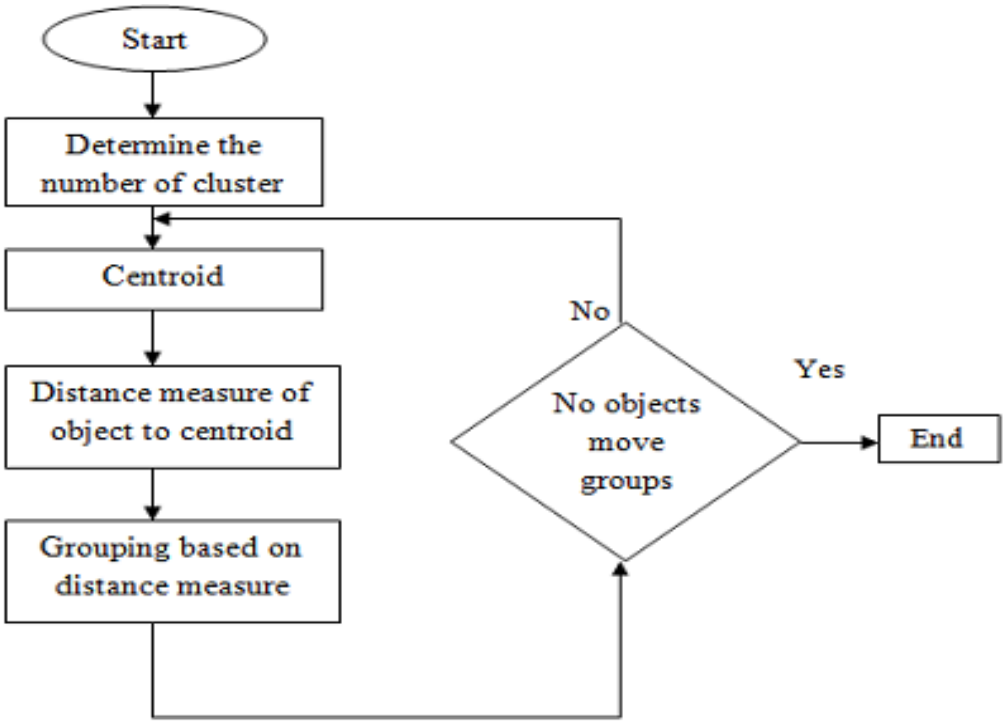

Fig 1: A simple iteration for Numerical example

2.2 Centroid Based Clustering

Clusters are represented by a central vector in centroid based clustering which may not essentially be a member of the data set. When the number of clusters is fixed to k. For optimization problem k-means clustering gives formal definition: find the k cluster centers and allocate the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

$$\mathbf{E} = = \sum_{i=1}^{k} \sum_{p \in ci}^{k} \mathbf{dist}\ (\mathbf{p}, \mathbf{c_i})^2$$

Where **E** is the sum of the squared error for all objects in the data sets, **p** is the point in space representing of an object, c_i is the centroid of cluster $c_{i.}$ The optimization problem is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximate method referred to as "k-means algorithm". Its find a local optimum, and is commonly run multiple times with different random initializations. Some variation of k-means such that optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set k-medoids, choosing medians that is k-medians clustering, choosing the initial centers less randomly K-means++ or allowing a fuzzy cluster assignment is referred as Fuzzy c-means[4-5].

One of the biggest drawbacks of these k-means algorithms is require the number of clusters Specified in advance.

Additionally, the k-means algorithms prefer clusters of approximately similar size, and always assign an object to the nearest centroid. This algorithm optimized cluster centers, not cluster borders.

2.3 Hierarchical Clustering Method

Hierarchical clustering is a method of cluster analysis which is to build a hierarchy of clusters. There are two types of hierarchical method such as,

- **Agglomerative**: This method is a "bottom up" approach: each and every observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This method is a "top down" approach: in one cluster all observation are started, and splits are performed recursively as one moves down the hierarchy.

Given a set of N objects to be clustered, and N*N distance matrix, the basic process of hierarchical clustering is this:

1. Start by assigning each object to a cluster, so that if it contain N objects, now you have N clusters, each containing just one objects. Let the distances between the clusters the same as the distances between the objects they contain.
2. Find the most similar pair of clusters and merge them into a single cluster, so that you have one cluster less.
3. Compute similarities between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

2.4 Agglomerative Hierarchical Clustering:

Grouping the data one by one on the basis of the nearest distance measure. All the pair wise distance between the data point are grouped based on the nearest distance measure. Distance between the data point is recalculated but don't know which distance has to consider when the groups has been formed? Several methods are available for this. Some of them are:

1. Single linkage or single-nearest distance
2. Complete linkage or complete farthest distance
3. Average linkage or average-average distance
4. Centroid distance.
5. Sum of squared Euclidean distance is minimized- ward's Method.

In this way we grouping the data until one cluster is formed. we can calculate how many numbers of clusters should be actually present on the basis of dendogram graph [6].

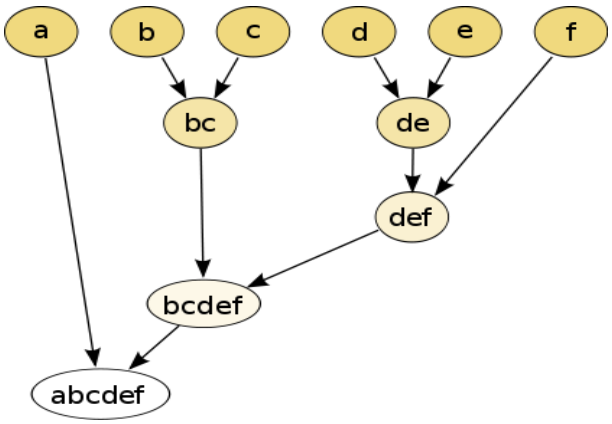

Fig 2: Agglomerative clustering on data objects

- Single-link clustering is also called nearest neighbor method or minimum method. Consider the distance between two clusters to be equal to the shortest distance from one cluster to another cluster. The similarity between a pair of clusters is considered to be equal to the greatest similarity from one cluster to another cluster
- Complete-link clustering is also called furthest neighbor method or maximum - methods that consider the distance between two clusters to be equal to the longest distance from one cluster to other cluster.
- Average-link clustering is also called minimum variance method. Consider the distance between two clusters to be equal to the average distance from one cluster to other cluster.
- Hierarchical methods are characterized with the following Advantages:

Versatility - The single-link methods, for example, maintain good performance on data sets containing separated, chain-like and concentric clusters.

Multiple partitions - hierarchical methods produce multiple nested partitions not only one partitions, which allow different partitions can be chosen by different users, according to the preferred similarity level.The main disadvantages of the hierarchical methods are:

Inability to scale well - The time complexity is at least O (m2), where m is the total number of instances by using a hierarchical algorithm is to clustering a large number of objects is characterize by huge Input/output costs.There is no back-tracking capability in hierarchical Method.

2.5 Distribution-Based Clustering

In distribution models, the clustering model is closely related to statistics. Clusters can easily be defined as objects most likely to the same distribution. A suitable property of this approach is that this closely resembles the way artificial data sets are generated by sampling random objects from a distribution. Although the theoretical foundation of these distribution methods is excellent, they go through from one key problem known as over fitting, if not constraint are put on the model complexity. A complex model will easily can able to explain the data better, it makes choosing the suitable model complexity logically difficult [7].

Gaussian mixture models are an prominent method by using the expectation-maximization algorithm. To avoid over fitting problem, number of Gaussian distributions are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This will join to a local optimum, a outcome multiple runs may produce different results. For hard clustering, objects are assigned to the Gaussian distribution. For soft clustering is not necessary. Distribution-based clustering produces complex models for clusters that can capture dependence and correlation between attributes.

Spatial Database Systems (SDBS) [9] are used for the management of spatial data such as points and polygons. The mission of spatial database in clustering, the main problem is to detecting the clusters of points which are distributed as Poisson point. This type of distribution is referred as random distribution or uniform distribution.

The function to large spatial databases raises the following requirements for clustering algorithms:

1. In many applications, approximate values are not know so minimal number of input parameters are taken.
2. We usually do not know the density, shape and number of cluster in the application of detecting minefield.

3. Shape of clusters in spatial databases may be spherical, elongated, drawn-out etc.discover of cluster based on arbitrary shape.
4. Provides a good efficiency on large databases.

The problem of detecting surface-laid minefields on the basis of an image from a investigation aircraft. After processing, an image is reduced to a set of points, some of images may be mines, and some of images may be noise, such as rocks or other metal objects. The main aim of the analysis is to find out whether minefields are present or not.

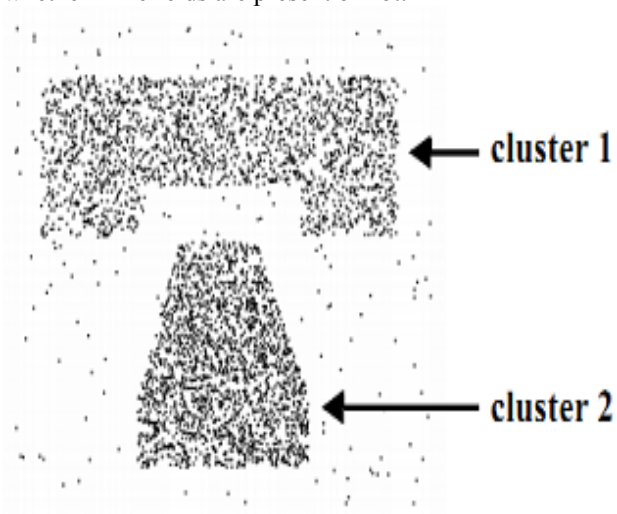

Fig 3: minefield database

2.6 Density-Based Algorithms for Discovering Clusters in Large Spatial Databases with Noise (Dbscan)

DBSCAN is a density based algorithm which discovers clusters with minimal number of input parameters and with the arbitrary shape. The input parameters required radius of the cluster (Eps) and minimum points required inside the cluster (Minpts). The basic idea behind this DBSCAN [8] algorithm is as follows,

Definition 1: NEps(p) is defined by
NEps(p) = {p D| dist(p,q)≤ Eps} where Eps is neighborhood point of p.
There are two types of points in the cluster, the points which is inside the cluster is denoted as core points and points on the border of the cluster is denoted as border points[8-9].

Definition 2: A point p is density-reachable from a point q wrt. A chain of points in Eps and Minpts is denoted as p1, pn where p1 = q, pn = p such that pi+1 is directly density-reachable from pi.

Definition 3: A point p is density-connected to a point q wrt. If there is a point o in Eps and Minpts both, p and q is density-reachable point from o wrt. Eps and MinPts.

Definition 4: Let D be a database of points. A cluster C wrt. Eps and MinPts is a non-empty subset of D satisfying the following conditions:

1. p, q: if p C and q is density-reachable from p wrt.Eps and MinPts, then q C is Maximality
2. p, q C: p is density-connected to q wrt. Eps and MinPts is Connectivity .

Description of the Algorithm

In the algorithm, Density Based Spatial Clustering of Applications [DBSCAN] is designed to discover the spatial data clusters with noise.

The steps involved in this algorithm are as follows,

1. Choose an arbitrary point p
2. Eps and Minpts are retrieving all the points density reachable from p.
3. A cluster is formed, when p is a core point
4. DBSCAN visits the next point of the database when p is a border points and there is no points are density reachable from p
5. Keep on the process until all the points have been processed.

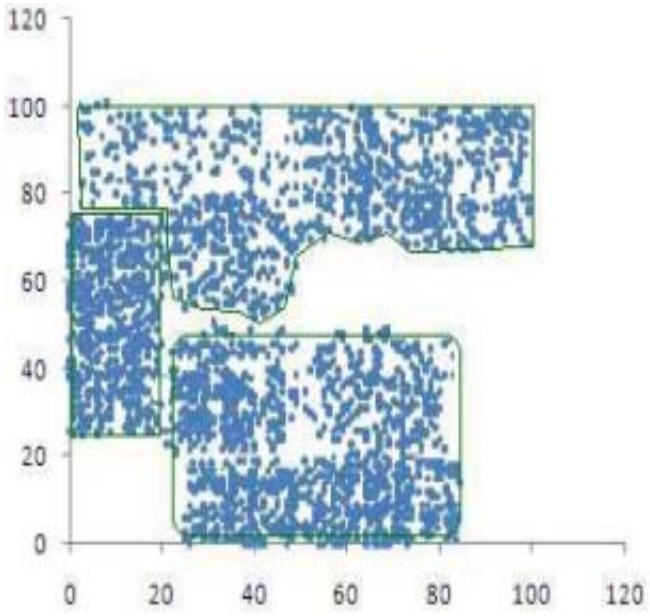

Fig 4: Cluster Generated by DBSCAN Algorithm

3. Result And Discussion

Based on the application of detecting near duplicate images the various clustering techniques are compared. The K-means clustering method results in 0.8 of Precision, 0.4 of Recall and 0.3 of F-measure are predicted in the performance.

Table 1: Shows the performance of the clustering methods.

S8	Clustering method	Precision	Recall	F-measure
1	Hierarchical Based Clustering	0.70	0.3	0.2
2	Distribution Based Clustering	0.5	0.2	0.1
3	K-means Clustering	0.8	0.4	0.3
4	DBSCAN	0.65	0.3	0.2

4. Conclusion

This paper deals with various clustering methods and study on each of them. Methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. Hence these clustering techniques show how a data can be clustered and grouped in large volume of dataset is available. As the number of cluster increases slowly, the time to form the clusters also increases. Based on the application of mining near duplicate image groups the K-means clustering methods predicts best method to mines duplicate image groups.

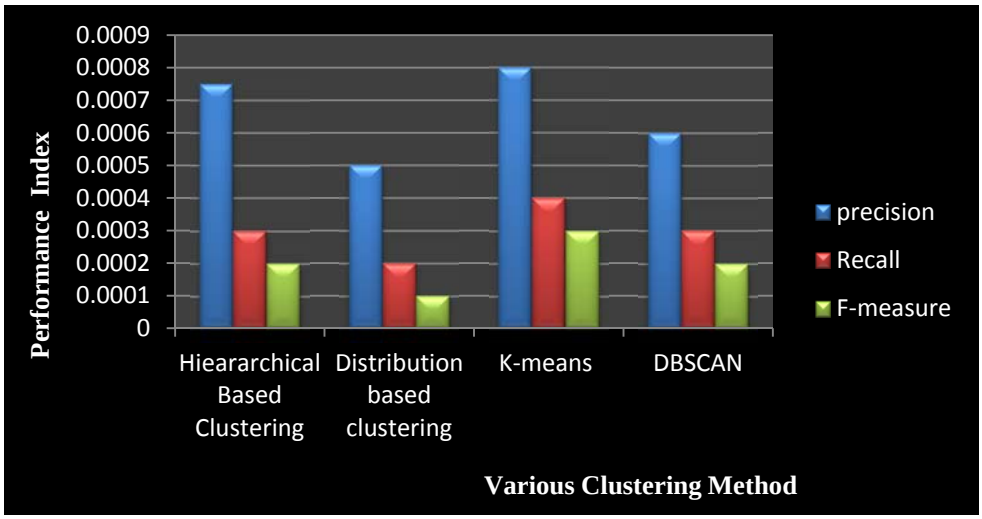

Fig 4: Shows the Comparison of the Clustering Methods

References

1. Rui Xu *et al.* "Survey of Clustering Algorithms", Student Member, IEEE and Donald Wunsch II, Fellow, IEEE.
2. Wang, B., Li, Z., Li, M., and Ma, W. "Large-scale duplicate detection for web image search," In Multimedia and Expo.
3. Pavel Berkhin *et al.* "Clustering Algorithm in Data Mining", Accrue Software, Inc.
4. Lior Rokach *et al.* "Various clustering methods" , Department of Industrial Engineering.
5. Tapas Kanungo *et al.* "An Efficient k-Means Clustering Algorithm", Senior Member, IEEE, David M. Mount, Member, IEEE.
6. Ryan Tibshirani *et al.* "Hierarchical clustering in data mining", School of Information Technology& Engineering.
7. Ester M,sander J *et al.* "A distribution based clustering algorithm for mining in large spatial database", University of Munich
8. Ester, Hans-Peter Kriegel, Jiirg Sander *et al.* "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Institute for Computer Science, University of Munich.
9. Martin *et al.* "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases"
10. Ujjwal Maulik *et al.* "Performance Evaluation of Some Clustering Algorithms and Validity Indices", Member, IEEE, and Sanghamitra Bandyopadhyay, Member, IEEE
11. Bangoria Bhoomi M*et al.* "Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Values", International Journal of Computer Science and Information Technologies.

Short Bio Data for Authors

Dr.P.TamijeSelvy received B.Tech (CSE), M.Tech (CSE) in 1996 and 1998 respectively from Pondicherry University. Since 1999, she has been working as faculty in reputed Engineering Colleges. At Present, she is working as Associate Professor in the department of Computer Science &Engg, Sri Krishna College of Technology, and Coimbatore. She has published more than 40 papers in reputed journals and conferences. Her Research interests include Image Processing, Data Mining, Pattern Recognition and Artificial Intelligence.

Ms.S.Ramya has received Bachelor of Engineering degree in Computer Science and Engineering under Anna University, Chennai in 2013. She is currently pursuing Master of Engineering degree in Computer Science and Engineering under Anna University, Chennai, India. Her areas of interest are image processing and data mining