International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-2, May 2015
ISSN: 2395-3470
www.ijseas.com

GENERALIZED METHOD FOR IMAGE DATA CLUSTERING

Ms. Ashwini Gulhane[1], Mr. Shreyas Deshmukh[2],

JDIT, Yavatmal, India [1]

RSCOE, Pune, India [2]

Abstract

Clustering is considered aninteresting approach for finding similaritiesin data and putting similar data into groups.Clustering partitions a data set into several groupssuch that the similarity within a group is largerthan that among groups. In this paper various clustering approaches are discussed. Further the typical steps for clustering and methodology used for the execution of clustering steps are included.

1. INTRODUCTION

Clustering is the process of grouping the similar type of objects or data points in one group while the objects in different groups are less similar. There are number of algorithms available for clustering which differs in there process of finding the clusters. Popular methods of clustering include groups with low distancesamong the cluster members, dense areas of the data space, intervals or particular statistical distributions.Clustering can therefore be formulated as a multi-objective optimization problem. The major problems occurs in clustering are,

- Distance function to be used for calculating the similarity among the data points.
- Selection of optimal threshold for clustering
- Expected number of clusters
- As the data sets are constantly becoming larger, it prevents easy analysis and validation of results.

Cluster analysis is widely spread in number of fields which requires the analysis of variety of data. In 2007, a search via Google Scholar results 1660 entries with the words dataclustering. This vast literature shows the importance and emerging growth of the clustering in various fields. One of the scientific field that uses clustering techniques in image segmentation. Image segmentation is an important technology for image processing. There aremany applications such as synthesis of the objects or computer graphic images which require precise segmentation. Image segmentation is a multiple objective problem. It involves several processes such as pattern representation, feature selection, feature extraction and pattern proximity. Considering all these objectives is a difficult problem, causing a gap between nature of images. It is animportant problem in computer vision, In 1996 Jain & Flynn,in 1999 Frigui& Krishnapuram and in 2000 Shi & Malik formulated it as a clustering problem. In recent years, the dramatic rise in the use of the web and the improvement in communications in general have transformed our society into one that strongly depends on information technology. The huge and amount of data that is generated by this communication process contains important information that accumulates daily in databases and is not easy to extract. So it requires more sophisticated way or

method to partition the data in the database into groups or cluster. Documents can beclustered [Iwayama & Tokunaga, 1995] to generate topic wise hierarchies for efficientinformation access [Sahami, 1998] or retrieval [Bhatia & Deogun, 1998].In the field of marketing, clustering isalso used to group customers into different types for efficient marketing [Arabie &Hubert, 1994], to group services delivery engagements for workforce management andplanning [Hu et al. , 2007] .Also it is useful to study genome data [Baldi & Hatfield, 2002] inbiology[5].

Data clustering has been used for the following three main purposes.

- Feature archiving: to gain insight into data, generate hypotheses, detect anomalies, and identify salient features.
- Data classification: to identify the degree of similarity among data points or patterns and group them based on similarity.
- Data Compression: as a method for organizing the data and summarizing it through cluster prototypes.

2. CLUSTERING APPROACHES

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]. There are number of clustering algorithms which are based on different approaches.Different approaches to clustering data are given as follow.

2.1 Agglomerative vs. divisive

The agglomerative approach of clustering is also known as bottom up approach. In this method each data points are considered as separate cluster and on every iteration they are merged successively based on a certain criteria, until stopping criteria is reached. This method builds the hierarchy from the individual elements by successively merging clusters. The first step is to determine which elements to merge in a cluster. Usually, it takes the two closest elements, according to the chosen distance criteria.In divisive approach of clustering which is called as top down approach, all the data points are considered as a single cluster and in every iteration they are splitted into number of different clusters based on certain criteria.

In order to decide which clusters should be merged in agglomerative approach or where a cluster should be split for divisive approach, a measure of dissimilarity between the image data points is considered. In most of the methods of hierarchical clustering, this is done by use of an appropriate metric that is a measure of distance between data points. 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 [1].

2.2 Monothetic vs. polythetic

This approach gives the idea of the sequential or simultaneous use of features in the clustering process. A simple Monothetic algorithm considers features sequentially to divide the given set of patterns. A Monothetic divisive clustering method utilizes only a single variable for a division in a particular step. This is illustrated in the figure 2.1. The given data set is firstly divided into two groups, using feature $X_{2,}$ shown by the horizontal line H. Then By

using feature X_1 each of this group or cluster is again divided independently into two cluster shown by vertical line V_1 and V_2.The main drawback of this method is that it generates 2d clusters where dis the dimensionality of the data points. So for the large value of d it generates large number of clusters which ultimately divide the data set into very small and fragmented clusters.

The polythetic algorithms consider the features simultaneously for clustering the given data set. Most of the clustering algorithms are polythetic, which uses all features simultaneously into the computation of distances between patterns and clustering is based on those distances [4].

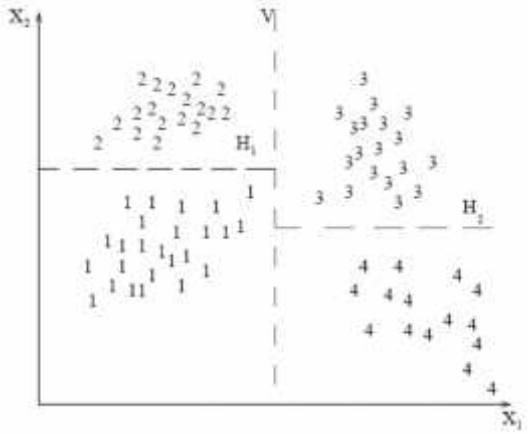

Fig. 2.1 Monothetic Partitional Clustering of Data Points

2.3 Hard *vs.* fuzzy

In hard clustering algorithm each pattern or data points belongs to a single cluster. In hard clustering data is partitioned into separate cluster, a single pattern is not belongs to more than one cluster. A hard partition can be obtained from a fuzzy partition clustering by thresholding the membership value. However fuzzy clustering algorithms allow the objects or patterns to belong more than one cluster simultaneously. It assigns different membership value to each pattern for several clusters. The fuzzy clustering can be converted to hard clustering by assigning each pattern to the cluster with large membership value. Many times fuzzy clustering gives more natural clustering than the hard clustering. For example for the objects on the boundaries between the several classes are not forced to fully belong to one of the class, in such case fuzzy clustering assigns membership value to belong object to more than one class. So the output clusters of this clustering algorithm are not disjoint [4][7].

The figure 2.2 shows the clusters obtained by the hard clustering and fuzzy clustering. The two square shows the completely disjoint hard clusters H_1= {3,2,8,1,4} and H_2 = {9,7,5,6} and the clusters obtained by the fuzzy clustering are shown by the two ellipse F_1 and F_2. From which it is clear that some patterns are associated with both the clusters, and does not gives disjoint clusters. Thus the Fuzzy clustering allows the overlapping of the clusters in which patterns partially belongs to all cluster with membership value in [0, 1] for each cluster.

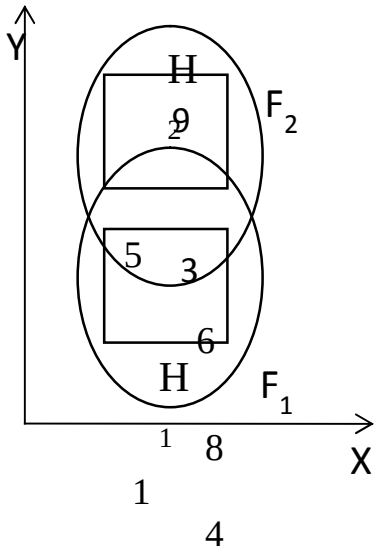

Fig. 2.2 Hard and Fuzzy Clustering of Data Points

3.CLUSTERING METHODOLOGY

Clustering as such is not an automatic task, but it is an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure.

Clustering is a task which involves number of stages. Typical clustering process used the following steps[1][2].

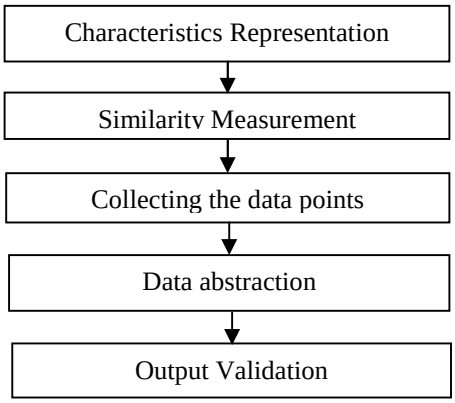

Fig. 3.1 Flow Of Clustering

Clustering is a task which can be performed by various algorithms that differs from each other in their methods of computing or finding the clusters. Thus the clustering results may depend on various parameters such as mean calculation formula, distance measures, threshold selection criteria and cluster models used etc in clustering algorithms. Typical clustering process used the following methods to be follow [3][4].

A. Graphical Representation

In this step the image is represented as a two dimensional intensity matrix. Here the intensities are represented in terms of 8 bit gray levels that are in the range of 0 to 255. RGB, HSI, HSV image models for graphical representation are also in existence.

B. Calculating The Mean

There are various measures available for calculating the mean of a given data samples. In mathematics, an average, or measure of central tendency of a data set is a measure of the "middle" value of the data set. Generally, the data set is a array of numbers. The average of a array of numbers is a single number represent the numbers in the array. If all the numbers in the array are the same, then this number should be used. If the numbers are not the same, the average is calculated by combining the numbers from the array in a specific way and computing a single number as being the average of the array. The most commonly used way is arithmetic mean to calculate the mean of data set but depending on the nature of the data other types of measures may be more appropriate. Some of the measures are given as follows.

- Arithmetic mean (AM)

The arithmetic mean is the "standard" average, usually simply called as "mean". It is calculated using following equation,

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

The mean is the arithmetic average of a set of values, or distribution.

- Geometric mean (GM)

The geometric mean of n non-negative numbers is obtained by multiplying them all together and then taking the nth root. In algebraic terms, the geometric mean of a_1, a_2, ..., a_n is defined as,

$$\mathrm{GM} = \sqrt[n]{\prod_{i=1}^{n} a_i} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

- Harmonic mean

Harmonic mean for a non-empty collection of numbers a_1, a_2, ..., a_n, all different from 0, is defined as the reciprocal of the arithmetic mean of the reciprocals of the a_i's is defined as,

$$HM = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_i}} = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}}.$$

C. Choosing Optimum Threshold

The key parameter in clustering process is the selection of the threshold value. There are number of different methods available for choosing threshold. User can manually select threshold value for the convergence of algorithm or a thresholding algorithm can also used which

compute the threshold value automatically, which is known as automatic thresholding.

A simple method will choose the mean or median value as threshold, the reason behind that is if the object pixels are brighter than the background, then they should also be brighter than the average. The mean or median can be used as threshold for the noiseless image with uniform background, but it is not always possible.

Another method of choosing threshold is to create histogram of the image pixel intensities and use the valley points in the histogram as the threshold. The histogram approach assumes that there is some average values for both the background and object pixels, but that the actual pixel values have some variation around these average values. However, this may be computationally expensive, and image histograms may not have clearly defined valley points, so it makes difficult to select an accurate threshold. Thresholding can be of the following types,

Global Thresholding: When a single threshold is used for the entire image then it is called as global thresholding.

Adaptive thresholding: When different thresholds are used suitable for different regions in the image then it is called as adaptive thresholding. It is also known as local or dynamic thresholding [6].

D. Similarity Measures

To compute the similarity among data points various distance measures are available such as Euclidian distance,Mahalanobis distance, Minkowski distance, Cosine distance etc.

- **Euclidean Distance**

The Euclidean distance or Euclidean metric is the "ordinary" distance between two points. For N number of data points where each point is denoted as P_i , P_j and so on. k denotes the number of cluster and d denotes the dimension . An N x N matrix Me is calculated. For points with d dimensions, the Euclidean distance Me(Pi, Pj) between two points Pi and Pj is defined as follows[6]:

$$M_e(P_i, P_j) = \sqrt{\sum_{x=1}^{d} (P_{i_x} - P_{j_x})^2}$$

where P_{ix} and P_{jx} represent the xth dimension values of Pi and Pj respectively. Also, Me is a symmetric matrix.

- **Mahalanobis distance**

Mahalanobis distance is a distance based on correlations between variables by which different patterns can be identified and analyzed. It measures similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant.

The Mahalanobis distance of a multivariate vector $x=(x_1,x_2,x_3,...x_N)^T$ from a group of values with mean$\mu=(\mu_1, \mu_2, \mu_3,....\mu_N)^T$ and covariance matrix S is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}.$$

- **Minkowski distance**

For higher dimensional data, a popular measure is the Minkowski metric, It is defined as follows

$$d_p(x_i, x_j) = \left(\sum_{K=1}^{d} |x_{i,k} - x_{j,k}|^p \right)^{\frac{1}{p}}$$

where *d* is the dimensionality of the data. The *Euclidean* distance is a special case where *p*=2, while *Manhattan* metric has *p*=1. However, there are no general theoretical guidelines for selecting a measure for any given application.

- **Cosine distance**

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. The cosine of 0 is 1, and less than 1 for any other angle; the lowest value of the cosine is -1. The cosine of the angle between two vectors thus determines whether two vectors are pointing in roughly the same

direction. This is often used to compare documents in text mining. In addition, it is used to measure similarity within clusters in the field of data mining. For given two vectors , A and B, the cosine similarity, θ, is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

E. Assigning Data Points To Cluster

Distance measure plays an important role in clustering data points. Choosing the right distance measure for a given dataset is an important issue in clustering. The similarity between various objects is defined by a distance measure. The distance measure plays an important role in obtaining correct clusters. Different formulas lead to different clustering. If the data points of a given data set are all in same physical units then simple Euclidean distance measure are employed to group the similar data together.

4. CLUSTER MODELS

Once we set the threshold value, the data points within threshold are assign to cluster centre to form a cluster. The idea of how to form cluster varies between algorithms. Selecting the proper clustering algorithm for a particular problem is depends upon many factors. Basically the cluster is nothing but the group of a data points. However the cluster found by different algorithms vary significantly in their properties. For understanding the differences between the various algorithms, understanding of cluster model acts as key factor. Some typical cluster models are

- **Connectivity models:** In this model clusters are form based on the distance connectivity.
- **Centroid models:** In this model clusters are form by assigning data points to a mean point which is treated as centre. In this model clusters are represent by a single mean vector.
- **Distribution models:** In this model clusters are form using statistic distributions, such as multivariate normal distributions.
- **Density models:** In this model clusters are defines as connected dense regions in the data space.

Clustering algorithms can be categorized based on the above listed cluster model. Following are some most commonly used clustering algorithms based on these models:

- **Connectivity based clustering**

The connectivity based clustering clusters the given data set by using distance connectivity. It is based on the core concept that the near by objects or data points in the data set are more related to each other than the object which are farther away. These algorithms represent the clusters in the nested form, from which the different clusters will obtained at different distances. These algorithms do not represent clusters in a single partition but instead provide a hierarchy of clusters that merges with each other or divide at certain distances. In connectivity based clustering along with usual distance measures user also have to take into account the linkage criterion. The most commonly used linkage criterion are single-linkage criterion and complete-linkage criterion. Connectivity based clustering, also known as hierarchical clustering.

- **Centroid based clustering**

The centroid based clustering algorithms cluster the given data set by assigning data points to a mean point which is treated as centre. In these algorithms clusters are represented by a central vector which may not be member of the

data set. Various distance measure are used to assign the data points to a cluster centre. k-means is the most common centroid based clustering algorithm. In which the user have to define number of clusters initially. Most of the k-means type algorithms require the number of clustersk to be defined in advance, which is considered to be one of the biggest drawbacks of these algorithms.

- **Distribution-based clustering**

The distribution based clustering algorithms cluster the given data by comparing the data set with any standard distribution models. The clusters can then be defined as per the distribution model. The expectation-maximization algorithm is an example of distribution based clustering algorithms. This algorithm usually modeled data with fixed number of Gaussian distributions. First it randomly initialised the distribution model to the data set then its parameters are iteratively optimized to fit appropriately to data set. This iterative method converges to local optimum and produces different output on every runs.

The advantage of distribution based clustering is that it not only gives the number of clusters but also shows the nature of clusters. But it is quite difficult to get appropriate data models for every data type. Many times there may be no mathematical models available for many real data sets.

- **Density-based clustering**

In density based clustering algorithms the densest region of the data set as compare to remainder of the data set is consider as cluster. Density based algorithm continue to grow the given cluster as long as the density in the neighbourhoodexceeds certain threshold. This algorithm is suitable for handling noise in the dataset. The most popular density based clustering algorithm is DBSCAN, as compare to many new algorithms it has well defined feature called density-reachability. Similar to the linkage based clustering, it is based on the connecting points within certain distance threshold. In this algorithms a cluster consist of all the density connected objects which form a cluster of an arbitrary shape. Also the complexity of DBSCAN algorithm is considerably low. The density based clustering has following features,

- It forms cluster of an arbitrary shape.
- Handle noise.
- Needs density parameters to be initialized.

5. CONCLUSION :

In this paper various clustering approaches for clustering algorithms are discussed.The main steps of clustering along with different measures for each step are discussed. Every clustering algorithm should follow the above steps and use one of the methods. It is observed that the output of any clustering algorithm is depends on the similarity measures used, mean calculation formula, selected threshold and opted cluster models.

REFERENCES

1. S. Anitha, Akilandeswari.J and Sathiyabhama.B,'A survey on partition clustering algorithm', International Journal of Enterprise Computing and Business System International Systems, vol. 1, pp. 1-13, 2011.
2. Vimal A, Valluri. S. R, Karlapalem. K, 'An Experiment with Distance Measures for Clustering', International Conference on Management of Data COMAD 2008, Mumbai, India, pp. 17–19, December 2008.
3. Andritsos Periklis, Data Clustering Techniques, Department of Computer ScienceUniversity of Toronto March 11, 2002.
4. Jain. A, MURTY. M. N and FLYNN. P. J,Data clustering: a review, ACM Computing Surveys, vol. 31,pp. 264-323,1999.

5. Ray and Turi, Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation, School of Computer Science andSoftware Engineering Monash University Australia.
6. Gonzalez and Woods, Digital Image Processing,third edition Pearson education, pp. 1-976, 2009.
7. Volkan Uslan and Đhsan Ömür Bucak, Microarray image segmentation using clustering Methods,Department of Computer Engineering, Fatih University, 34500.