ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

JOURNAL OF CURRENT SCIENCE

CLUSTERING OF IMAGE DATA, IN GENERAL

K. Nishteswar
Research Scholar, Energy Technology, Shivaji University

Abstract

Clustering is an interesting technique for exploring data for commonalities and structuring that data into useful groups.The goal of cluster analysis is to partition a dataset into groups whose members have more characteristics in common with one another than they do with those in other clusters. Several distinct types of cluster analysis are investigated in this work. Instructions on how to do common clustering tasks are included as well.

1. INTRODUCTION

Data points or objects with similar characteristics are clustered together, while those with less similarities are separated. There are a number of different clustering algorithms available, each with its own unique method for identifying clusters. Methods for solving the clustering issue sometimes include grouping data points together into dense groups, creating intervals, or applying certain statistical distributions.Clustering may therefore be seen as a problem in which several objectives must be optimized. o The distance function that will be utilized to estimate how closely the data points are connected to one another, is the most typical difficulty that arises from clustering.o Identifying an appropriate cutoff point for categorizing dataA Prediction of Cluster DensityGrowing data collections provide challenges for effective analysis and verification of results.Cluster analysis is used extensively in a wide variety of fields that depend on a wide variety of data sets. If you search Google Scholar for "dataclustering" from that year, you'll find 1660 results. This massive collection of work illustrates the relevance and rapid growth of clustering in a variety of fields. Clustering techniques are used in several fields of study, including image segmentation. Image segmentation is an essential approach in the study of image processing. Many applications rely on accurate segmentation, such as object creation and computer graphics. The task of picture segmentation has several objectives. The technique involves a number of steps, including representation of the pattern, feature selection, feature extraction, and pattern proximity. Trying to achieve all three aims at once might be difficult due to the incongruity in the nature of the resulting images. It's a serious problem in computer vision research. Jain and Flynn (1996), Frigui and Krishnapuram (1999), and Shi and Malik (2000) were the first to frame it as a clustering problem. Over the last several years, our reliance on technology has increased dramatically due to the widespread use of the internet and the development of more efficient means of communication. Due to the enormous data volumes produced by this kind of interaction, databases quickly get bloated with irrelevant or irrelevant-but-valuable information. Therefore, a more complicated strategy is required to partition the database into portions. Using topical hierarchies [Sahami, 1998] derived from document clustering [Iwayama & Tokunaga, 1995], information may be quickly accessed [Bhatia & Deogun, 1998] and retrieved [Sahami, 1998].Cluster analysis is frequently used in the advertising and service provision industries to better target audiences with tailored messages [Arabie &Hubert, 1994; Hu et al., 2007].There are real-world biological uses for analyzing genomic data [Baldi & Hatfield, 2002][5].

Data clustering has been used for thefollowing three main purposes.

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

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

CLUSTERING APPROACHES

Clustering is the process of dividing a collection of items into smaller groups (called clusters) based on how similar they are to each other. There isn't just one algorithm for clustering; rather, it's a job that may be accomplished in a number of different ways, depending on the individual algorithm used. The goal of clustering is to collect comparable picture pixels into a single cluster based on some feature, with the resultant cluster exhibiting strong intra-cluster similarities and low inter-cluster similarities. In order to classify data points into groups or clusters, the clustering method may be used [1, 4]. Several distinct types of clustering algorithms have been developed.The following are few examples of data clustering methods.

Agglomerative vs. divisive

Clustering's agglomerative method is sometimes called the bottom-up method. Each data point is treated as a cluster in this approach, and clusters are merged into larger ones at each iteration until a stopping criterion is fulfilled. Using this strategy, the hierarchy is constructed from the ground up by merging clusters in a progressive manner. Finding out which parts of a cluster may be combined is the first step. In most cases, the two items with the shortest distance to one another are used.The top-down, or divisive, technique to clustering begins with the whole dataset as a single cluster, and then divides the data into sub-clusters at each iteration.

A measure of dissimilarity between the image data points is assessed to determine whether clusters should be merged in an agglomerative method or where a cluster should be divided in a divisive approach. Most hierarchical clustering techniques do this by using a suitable metric, or distance measure, between the cluster nodes. In hierarchical clustering, the linking criteria is just as crucial as the distance function. Given that each cluster in this approach has many components, calculating distances requires the participation of more than one variable. Single-linkage and complete-linkage criteria are the most used ones [1].

Monothetic vs. polythetic

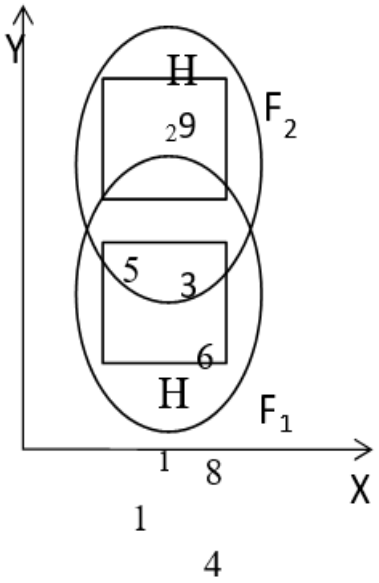
This strategy suggests using characteristics in either a sequential or simultaneous fashion throughout the clustering procedure. A straightforward Monothetic method takes into account characteristics in order to partition the input collection of patterns. A Monothetic divisive clustering approach employs just a single variable for a division in a specific stage. Figure 2.1 depicts this concept. Feature X2, represented by the line H in the supplied data set, is first used to split the dataset in half. Then Each of these groups, or clusters, is then split into the two separate clusters represented by the vertical lines V1 and V2 using feature X1.This method's fundamental flaw is that it produces 2d clusters with dimensionality disproportional to the original data. As a result, when d is big, a great many clusters are created, and the data set is then split up into many, tiny pieces.

In order to properly classify a dataset, polythetic algorithms take into account all of its attributes concurrently. Most of the clustering algorithms are polythetic, which employs all attributes concurrently into the calculation of distances between patterns and grouping is based on those distances [4].

the membership value of the cluster is high. Fuzzy clustering often provides more realistic results than rigid grouping. Fuzzy clustering gives membership value to the object to more than one class, for example, if the item is located on the boundary between many classes and does not have to entirely belong to one class. This clustering approach produces groups that are not mutually exclusive [4], [7].

Clusters generated by hard clustering and fuzzy clustering are shown in figure 2.2. The two squares represent the mutually exclusive hard clusters H1= 3,2,8,1,4 and H2 = 9,7,5,6, whereas the two ellipses F1 and F2 represent clusters formed by fuzzy clustering. The results show that there are commonalities between the two groups, rather than completely separate ones. Fuzzy clustering makes it possible for clusters to overlap if a given pattern has a membership value in the range [0, 1] for more than one cluster.

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022



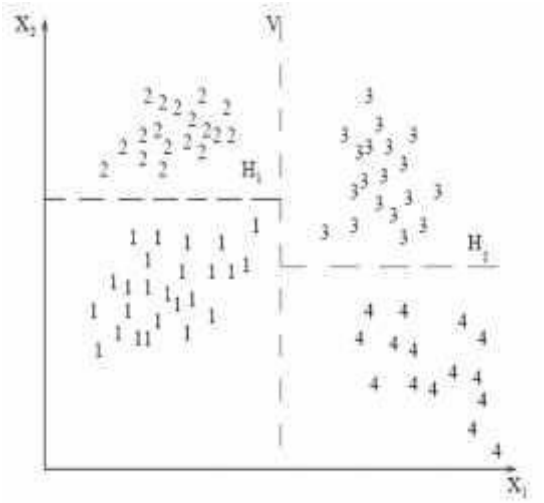


Fig. 2.1 Monothetic Partitional Clustering ofcluster, a single pattern is not belongs to more than one cluster. A hard partition can beobtained from a fuzzy partition clustering by thresholding the membership value. However fuzzy clustering algorithms allow the objects or patterns to belong more than one cluster simultaneously. It assigns different membership value to each pattern for several clusters. The fuzzy clustering can be converted to hardclustering by assigning each pattern to theFig. 2.2 Hard and Fuzzy Clustering of DataPoints

2.CLUSTERING METHODOLOGY

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

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

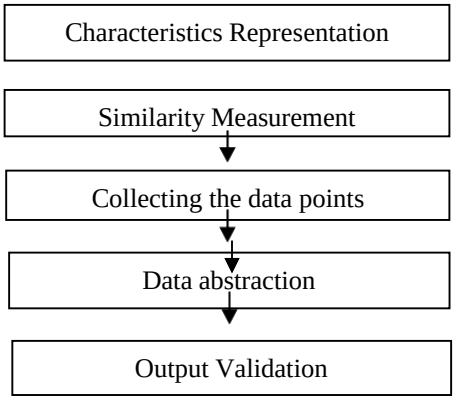


Fig. 3.1 Flow Of Clustering Clustering is a task which can be performed by various algorithms that differs from each other in their methods of computingor finding the clusters. Thus the clustering results may depend on various parameters such as mean calculation formula,

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

distance measures, threshold selection criteria and cluster modelsused etc in clustering algorithms. Typical clustering process used the following methods tobe 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 forcalculating 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 asingle 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 typesof 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 *n*th 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 algorithmor a thresholding algorithm can also used which

Automatic thresholding is a process in which the threshold value is calculated on its own.

If object pixels are brighter than the background, they should also be brighter than the average, which is why the mean or median value is often used as the threshold in more straightforward approaches. The noise-free picture with a consistent backdrop may utilize the mean or median as a threshold, but this is not always the

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

case.

Making a histogram of the intensity of the image's pixels and using the valley points as the threshold is another option. The histogram method presupposes not only that there are average values for background and object pixels but also that there is some variance in the actual pixel values. Selecting an exact threshold, however, may be computationally costly, because picture histograms may not have clearly defined valley spots. The following are some examples of thresholding:

,

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 ordynamic 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 oncorrelations between variables by whichdifferent patterns can be identified and analyzed.It measures similarity of an unknown sample setto a known one. It differs from Euclideandistance in that it takes into account thecorrelations 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 valueswith 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 popularmeasure is the Minkowski metric, It is definedas follows

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

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.

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

- **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 whethertwo 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. Ifthe data points of a given data set are all in samephysical units then simple Euclidean distancemeasure are employed to group the similar data together.

3. 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 formcluster 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 ofa data points. However the cluster found by different algorithms vary significantly in their properties. For understanding the differences between the various algorithms, understandingof cluster model acts as key factor. Some typicalcluster models are

- **Connectivity models:** In this model clusters are form based on the distance connectivity.
- **Centroid models:** In this model clustersare form by assigning data points to a mean point which is treated as centre.

In this model clusters are represent by asingle mean vector.

- **Distribution models:** In this modelclusters 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 somemost commonly used clusteringalgorithms 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 providea hierarchy of clusters that merges with eachother or divide at certain distances. Inconnectivity based clustering along with usual distance measures user also have to take intoaccount 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.

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

- **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 ofthese algorithms.

- **Distribution-based clustering**

The distribution based clustering algorithms cluster the given data by comparing the data set with any standard distribution models. Theclusters 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 fixednumber of Gaussian distributions. First it randomly initialised the distribution model to thedata set then its parameters are iterativelyoptimized to fit appropriately to data set. This iterative method converges to local optimum andproduces 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 thedensest 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 thelinkage 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 thecomplexity 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.

4. CONCLUSION :

In this research, we investigate a wide variety of approaches to developing clustering algorithms.The steps involved in clustering are dissected, and the metrics used at each stage are discussed in detail. Every clustering algorithm must include at least one of these methods. This study demonstrates how several factors, including the similarity measures used, the mean calculation formula, the threshold, and the cluster models selected, may significantly affect the outcomes of a clustering procedure..

ISSN NO: 9726-001X
Volume 10 Issue 4 April 2022

REFERENCES

1. The first is "A survey on partition clustering method," by S. Anitha, Akilandeswari.J, and Sathiyabhama.B, published in the International Journal of Enterprise Computing and Business System International Systems in 2011, volume 1, pages 1-13.
2. "An Experiment with Distance Measures for Clustering," by Vimal A., S. R. Valluri, and K. Karlapalem; presented at the International Conference on Management of Data (COMAD) 2008; Mumbai, India; December 17–19; pages 17–19.
3. Data Clustering Techniques, Andritsos Periklis, 11 March 2002, Department of Computer Science, University of Toronto.
4. Data clustering: a review, ACM Computing Surveys, volume 31, pages 264–323, 1999. Jain. A., MURTY. M. N., and FLYNN. P. J.
5. School of Computer Science andSoftware Engineering, Monash University, Australia; Ray, Turi, Determination of Number of Clusters in K-Means Clustering and Application in Color Image Segmentation (2017).
6. Sixth Chapter, Digital Image Processing, by Gonzalez and Woods, Third Edition, Pearson Education, 2009, Pages 1- 976.
7. Microarray image segmentation utilizing clustering methods. Department of Computer Engineering, Fatih University, 34500. 7. Volkan Uslan and hsan mür Bucak.