Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey of Clustering Techniques

Ramandeep Kaur*
Department of Computer Engineering,
University College of Engineering,
Punjabi University, Patiala, Punjab, India.

Dr. Gurjit Singh Bhathal
Department of Computer Engineering,
University College of Engineering,
Punjabi University, Patiala, Punjab, India

Abstract— Clustering can be considered the most important unsupervised learning technique; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. Clustering is "the process of organizing objects into groups whose members are similar in some way." A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In this paper, we are describing the clustering techniques and algorithms used for it.

Keywords— Clustering, Goals of clustering, clustering techniques, clustering algorithms.

I. INTRODUCTION

Clustering is the process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a dataset. A good clustering method will produce high quality clusters in which the intra-class (i.e., intra-clusters) similarity is high and the inter-class similarity is low. The quality of clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or the entire hidden pattern.

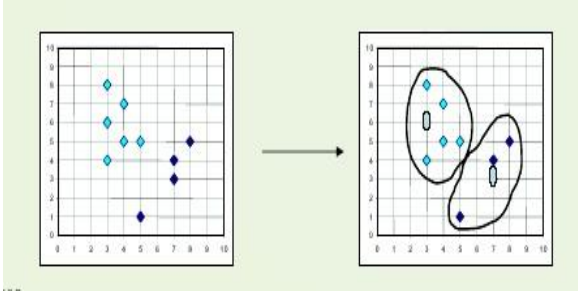


Fig.1 The result of Cluster analysis

The following are typical requirements of clustering in data mining:

1) *Scalability:* Many clustering algorithms work well on small data sets containing fewer than 200 data objects. However, a large database may contain millions of objects.
2) *Ability to deal with different types of attributed:* Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.
3) *Discovery of clusters with arbitrary shape:* Many clustering algorithms determined clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures 'end to find spherical clusters with similar size and density'. It is important to develop algorithms that can detect clusters of arbitrary shape.
4) *Minimal requirements for domain knowledge of determine input parameters:* Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters..
5) *Ability to deal with noisy data:* Most real-world databases contain outliners or missing, unknown, erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
6) *Insensitivity to the order of input records:* Some clustering algorithms are sensitive to the order of input data; for example, may generated dramatically different clusters. It is important to develop algorithms that are insensitive to the order of input.

7) *High dimensionality:* A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions.
8) *Constraint-based clustering:* Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic cash-dispensing machines (ATMs) in a city. To decide upon this, we may cluster household while considering constraints such as the city's rivers and highway networks and customer requirements per region.
9) *Interpretability and usability:* Users expect clustering results to be interpretable, comprehensible, and usable. That is, clustering may need to be tied up with specific semantic interpretations and applications. It is important to study how an applications goal may influence the selection of clustering methods.

II. 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 clustering. Consequently, it is a user which must supply this criterion, in such a way that the results of 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).

III. Clustering Techniques

A. Hierarchical Clustering

A hierarchical clustering creates a hierarchical decomposition of the given set of data objects. A hierarchical clustering can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition if formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects that are close to one another, until all the groups are merging into one. The divisive approach, also called as top down approach, starts with all of the objects in the same clusters. In this, a cluster is split up into smaller clusters, until eventually each object is in one another. In this clustering, Dendograms are great for visualization. It provides hierarchical relations between clusters. it shown to be able to capture concentric clusters. In this clustering, it is not easy to define levels for clusters.

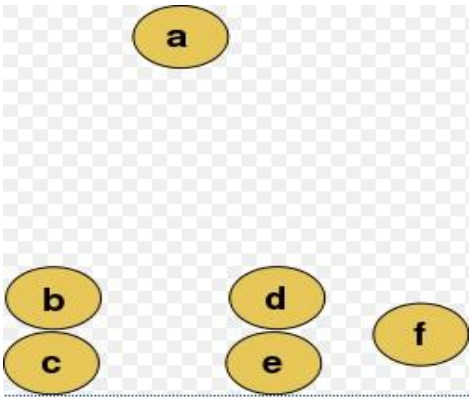
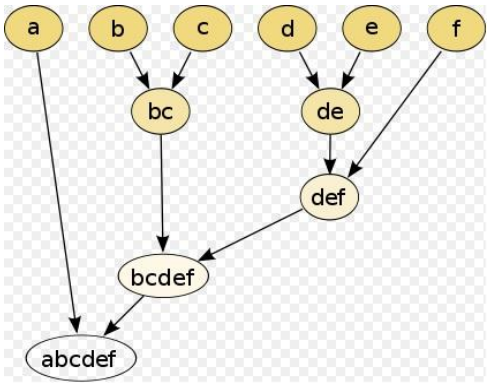


Fig.2 Example for Agglomerative Clustering

The hierarchical clustering dendrogram would be as such:



B. Density- based Clustering

Most partitioning methods cluster object based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Density-based clustering does not require one to specify the number of clusters in the data a priori, as opposed to k-means. It has a notion of noise. It requires just two parameters and is mostly in sensitive to the ordering of the points in the database. The quality of density-based clustering depends on the distance measure used in the function.

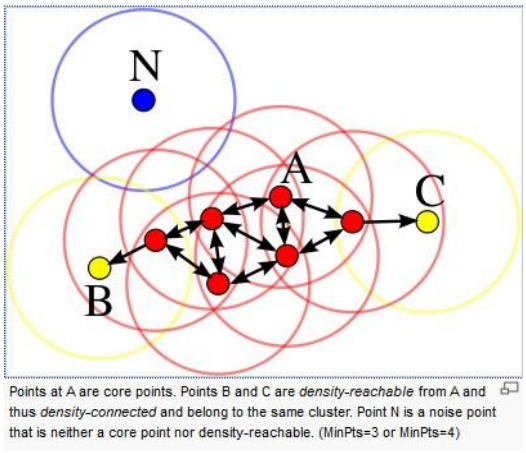


Fig.3 Example for Density- based Clustering

C. Grid-based Clustering

Grid-based clustering quantizes the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. "STING" is a typical example of grid-based method.

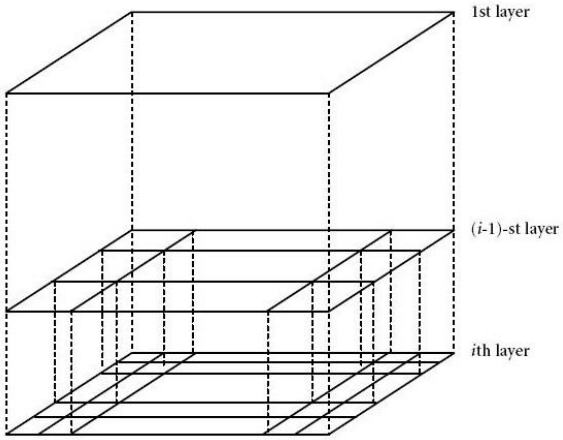


Fig.4 Example for Grid-based Clustering

D. Model-based Clustering

Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of data points.

E. Partitioning Method

In this method, given a database of 'n' objects, a partitioning methods construct k partitions of the data, where each partition represents a cluster and k<= n. That is, it classifies the data into 'k' groups, which together satisfy the requirements: each group must contain at least one object and each object must belong to exactly one group.

Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. It uses an "iterative relocation technique" that attempts to improve the partitioning by moving object from one group to another.

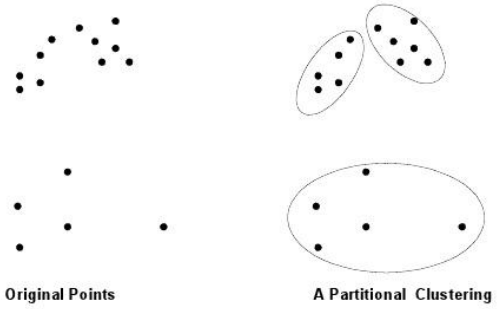


Fig.5 Example for Partitioning Method

a) K-mean algorithm

1. It accepts the number of clusters to group data into, and the dataset to cluster as input values.
2. It then creates the first K initial clusters (K= number of clusters needed) from the dataset by choosing K rows of data randomly from the dataset.
3. The K-Means algorithm calculates the Arithmetic Mean of each cluster formed in the dataset. The Arithmetic Mean of a cluster is the mean of all the individual records in the cluster. In each of the first K initial clusters, their is only one record. The Arithmetic Mean of a cluster with one record is the set of values that make up that record.
4. Next, K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure.
5. K-Means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset. The arithmetic mean of a cluster is the arithmetic mean of all the records in that cluster.
6. K-Means re-assigns each record in the dataset to only one of the new clusters formed. A record or data point is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity
7. The preceding steps are repeated until stable clusters are formed and the K-Means clustering procedure is completed. Stable clusters are formed when new iterations or repetitions of the K-Means clustering algorithm does not create new clusters as the cluster center or Arithmetic Mean of each cluster formed is the same as the old cluster center. There are different techniques for determining when a stable cluster is formed or when the k-means clustering algorithm procedure is completed.

b) Fuzzy c-means

1. It is an extension of k-means.
2. Fuzzy c-means allows data points to be assigned into more than one cluster.
3. In c-mean, each data point has a degree of membership (or probability) of belonging to each cluster.

Fuzzy c-means algorithm

Let x_i be a vector of values for data point g_i.

1. Initialize membership $U^{(0)} = [\, u_{ij} \,]$ for data point g_i of cluster cl_j by random
2. At the *k*-th step, compute the fuzzy centroid $C^{(k)} = [\, c_j \,]$ for $j = 1, .., nc$, where *nc* is the number of clusters, using

$$c_j = \frac{\sum_{i=1}^{n} (u_{ij})^m x_i}{\sum_{i=1}^{n} (u_{ij})^m}$$

where *m* is the fuzzy parameter and *n* is the number of data points.

3. Update the fuzzy membership $U^{(k)} = [\, u_{ij} \,]$, using

$$u_{ij} = \frac{\left(\frac{1}{\| x_i - c_j \|} \right)^{\frac{1}{(m-1)}}}{\sum_{j=1}^{n_c} \left(\frac{1}{\| x_i - c_j \|} \right)^{\frac{1}{(m-1)}}}$$

4. If $\|U^{(k)} - U^{(k-1)}\| < \varepsilon$, then STOP, else return to step 2.
5. Determine membership cutoff

   For each data point g_i, assign g_i to cluster cl_j if u_{ij} of $U^{(k)} > \alpha$

IV. Why We Use Fuzzy C-Means Algorithm

In fuzzy K-means algorithm, clusters are identical whereas, In fuzzy C-means, clusters are sometimes similar. Fuzzy C-means provides the better results than K-means because membership exits in their clusters. Fuzzy C-means algorithm also provides flexibility, So changes can easily done.

V. Conclusion And Future Work

This paper represents the survey of different techniques of clustering. In future, we work on the fuzzy C-means algorithm to get the improved results of image.

References

[1] X. Chang, W. Li, and J. Farrell, "*A C-means clustering based fuzzy modeling method,*" The Ninth IEEE International Conference on Fuzzy Systems, Vol. 2, 2000, pp. 937-940.

[2] Amiya Halder, Soumajit Pramanik, Arindam Kar, "*Dynamic Image Segmentation using Fuzzy C-Means based Genetic Algorithm*," International Journal of Computer Applications, Volume 28-No.6, August 2011.

[3] Chaloemchai Lowongtrakool and Nualsawat Hiransakolwong, "*Design of Image Segmentation by Automatic Unsupervised Clustering using Computation Intelligence*," International Conference on Machine Learning and Computer Science (IMLCS'2012) August 11-12, 2012 Phuket.

[4] B.Sowmya, B.Sheela Rani "*Color Image Segmentation usingFuzzy Clustering techniques and Competitive Neural Network*," Applied Soft Computing, Vol. 11(3), 3170-3178, April 2011.

[5] M.Ameer Ali, Laurence S Dooley and Gour C Karmakar "*Object Based Image Segmentation Using Fuzzy Clustering*," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), 14-19 May 2006, Toulouse.

[6] Ehsan Nadernejad and Amin Barari, "*A Novel Pixon-Based Image Segmentation Process Using Fuzzy Filtering and Fuzzy C-mean Algorithm*,**"** International Journal of Fuzzy Systems, Vol. 13, No. 4, December 2011.

[7] C. H. Li, W. C. Huang, B. C. Kua, and C. C. Hung, "*A novel fuzzy weighted c-means method for image classification*," International Journal of Fuzzy Systems, vol.10, no. 3, pp. 168-173, 2008.

[8] Sahaphong and N.Hiransakolwong, "*Unsupervised Image Segmentation Using Automated Fuzzy c-Means*", Seventh International Conference on Computer and Information Technology, 690-694, 2007.

