ISSN: 2278 – 1323

International Journal of Advanced Research in Computer Engineering &Technology (IJARCET)
Volume 2 Issue 11, November 2013

A REVIEW ON CLUSTERING TECHNIQUES AND THEIR COMPARISON

W.Sarada, Dr.P.V.Kumar

Abstract— Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. A cluster is intended to group objects that are related, based on observations of their attribute's values. Clustering is often confused with classification, but there is some difference between the two. In classifications the objects are assigned to pre-defined classes, where as in clustering the classes are formed. The term "class" is in fact frequently used as a synonym to the term "cluster". Clustering is used in data analysis, pattern recognition and data mining for finding unknown groups in data. This paper is intended to study and compare different clustering algorithms. The algorithms under investigation are hierarchical, partitioned; density based clustering according to the factors: methodology, structure, model, application or suitability, usefulness. Clustering is a main task of explorative, data mining and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bio informatics.

Index Terms— **Cluster Analysis, hierarchical clustering, partitioned clustering, density based clustering, Sub space clustering**

I. INTRODUCTION

Cluster analysis groups objects based on their similarity and has wide applications .[1],[2]Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be “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. We can show this with a simple graphical example in fig:1 below:

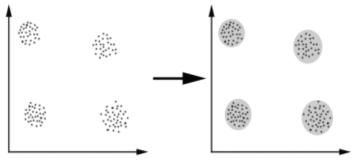

Fig:1 A simple graphical example

to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering*:* two or more objects belong to the same Cluster, if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

A. An Example of Clustering

In order to elaborate the concept a little bit, let us take the example of the library system. In a library books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster. For example, books on the database are kept in one shelf and books on operating systems are kept in another cupboard, and so on. To further reduce the complexity, the books that cover same kind of topics are placed in same shelf. And then the shelf and the cupboards are labeled with the relative name. Now when a user wants a book of specific kind on specific topic, he or she would only have to go to that particular shelf and check for the book rather than checking in the entire library.

B. Concepts of Clustering

Clusters are represented in different ways [3] such as Division with boundaries, Spheres, Probabilistic, dendrograms. The quality of a clustering result depends on both the similarity measure used by the method and its application. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns The types of data in clustering analysis are Interval-scaled variables, Binary variables, Nominal, ordinal, and ratio variables, Variables of mixed types.

In this paper, different clustering techniques and comparisons among them are discussed. This paper is organized as follows: Section 2 describes the goals, possible applications, classification, to provide a self-contained review of the concepts underlying clustering techniques. Section 3 discusses cluster models and some common distance measures. Then, Section 4 introduces some of the clustering techniques their advantages and limitations and compares them with regard to the various factors such as methodology, structure, model, application, usefulness and conclusion in Section 5.

II. THE GOALS OF CLUSTERING

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. 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*).*

A. Possible Applications

Clustering algorithms can be applied in many fields [4], for instance:

Marketing*:* finding groups of customers with similar behavior 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 epicenters to identify dangerous zones;
WWW*:* document classification; clustering weblog data to discover groups of similar access patterns.

B. Classification

Clustering algorithms may be classified [5] as Exclusive Clustering, Overlapping clustering, and Hierarchical clustering, Probabilistic Clustering. In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the Fig-2 below, where the separation of points is achieved by a straight line on a bi-dimensional plane. On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

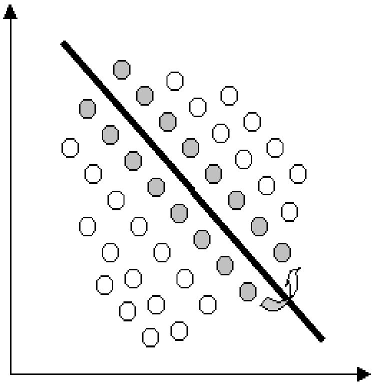

Fig: 2- separation of points is achieved by a straight line on a bi-dimensional plane

Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering uses a completely probabilistic approach.

The notion of a cluster varies between algorithms and is one of the many decisions to take when choosing the appropriate algorithm for a particular problem. At first the terminology of a cluster seems obvious: a group of data objects.

However, the clusters found by different algorithms vary significantly in their properties, and understanding these cluster models is key to understanding the differences between the various algorithms.

III. CLUSTER MODELS

A. Typical cluster models include:

Centroid models: for example the k-means algorithm represents each cluster by a single mean vector.
Connectivity models: for example hierarchical clustering builds models based on distance connectivity.
Distribution models: clusters are modeled using statistic distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.
Density models: for example DBSCAN defines clusters as connected dense regions in the data space.
Subspace models: in Bi-clustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.
Group models: some algorithms (unfortunately) do not provide a refined model for their results and just provide the grouping information.

C. *Distance measure*

An important step in most clustering techniques is to select a distance measure [6], which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (x = 1, y = 0) and the origin (x = 0, y = 0) is always 1 according to the usual norms, but the distance between the point (x = 1, y = 1) and the origin can be 2, or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

TABLE: 1 Some Common distance functions

A cluster is an ordered list of objects, which have some common characteristics. The objects belong to an interval [a, b], in our case [0, 1]. [7]The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed. A similarity measure SIMILAR (Di, Dj) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement by using clustering an enormous amount of time in finding the exact match from the database is reduced.

IV. CLUSTERING TECHNIQUES

A. *Hierarchical clustering*

[9] Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Hierarchical clustering creates 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, in which one starts at the leaves and successively merges clusters together; or divisive, in which one starts at the root and recursively splits the clusters. Another variation of the agglomerative clustering approach is conceptual clustering.

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

Find the two closest objects and merge them into a cluster
Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
If more than one cluster remains, return to step 2
Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

B. *Partition clustering*

Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. This algorithm typically determines all clusters at once. Most applications adopt one of two popular heuristic methods like k-means algorithm, k-medoids algorithm. Partition algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering. K-means and derivatives

S. no	Distance Measure	Distance Type	More common measure
1	Euclidean	crow flies or 2-norm	Euclidean
2	Manhattan	Taxicab or 1-norm	
3	Maximum norm	Infinity norm	
4	Hamming	minimum number of substitutions required to change one member into another	

A. *k-means clustering*

[10]-[12] The k-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters). It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always. For such datasets the k-medoids variants is appropriate. An alternative, using a different criterion for which points are best assigned to which centre k-medians is clustering.

The reason behind choosing the k-means algorithm to study is its popularity for the following reasons: Its time complexity is O (nkl), where n is the number of patterns, k is the number of clusters, and l is the number of iterations taken by the algorithm to converge. Its space complexity is O (k+n). It requires additional space to store the data matrix. It is order-independent; for a given initial seed set of cluster centers, it generates the same partition of the data irrespective of the order in which the patterns are presented to the algorithm.

Basic K-means Algorithm for finding K clusters

1. Select *K* points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.

B. K-medoids algorithm:

The basic strategy of k-medoids algorithm is each cluster is represented by one of the objects located near the center of the cluster. PAM (Partitioning around Medoids) was one of the first k-medoids algorithm is introduced.

The advantage is K-medoids method is more robust than k-mean in presence of noise and outliers because a medoids is less influenced by outliers or other extreme values than a mean. The disadvantage is, it is relatively more costly; complexity is O(i k (n-k)2),where i is the total number of iterations, k is the total number of clusters, and n is the total number of objects. Need to specify k, the total number of clusters in advance. Result and total run time depends upon initial partition.

C. Density-based clustering algorithms

In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are two typical algorithms of this kind. [6]DBscan Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers.

The key idea of the DBSCAN algorithm is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold. DBSCAN does not deal very well with clusters of different densities.

D. SNN Algorithm

[8], [10] The SNN algorithm, as DBSCAN, is a density-based clustering algorithm. The main difference between this algorithm and DBSCAN is that it defines the similarity between points by looking at the number of nearest neighbors that two points share.

Using this similarity measure in the SNN algorithm, the density is defined as the sum of the similarities of the nearest neighbors of a point. Points with high density become core points, while points with low density represent noise points. All remainder points that are strongly similar to a specific core points will represent a new clusters.

E. Subspace clustering

Subspace clustering methods look for clusters that can only be seen in a particular projection (subspace, manifold) of the data. In these methods not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously. [13]Two important parameters fundamentally affect the performance of subspace clustering algorithms :(1) the distance between subspaces and (2) the number of samples we have on each subspace. Subspace clustering refers to the task of finding a multi-subspace representation that best fits a collection of points taken from a high-dimensional space.

Subspaces can either be axis-parallel or affine. The term is often used synonymous with general clustering in high-dimensional data. The image below shows a mere two-dimensional space where a number of clusters can be identified. In the one-dimensional subspaces, the clusters C_a(in subspace $\{x\}$) and C_b, C_c, C_d(in subspace $\{y\}$) can be found. C_c Cannot be considered a cluster in a two-dimensional (sub-) space, since it is too sparsely distributed in the $x$axis. In two dimensions, the two clusters $C_{ab}$and C_{ad}can be identified.

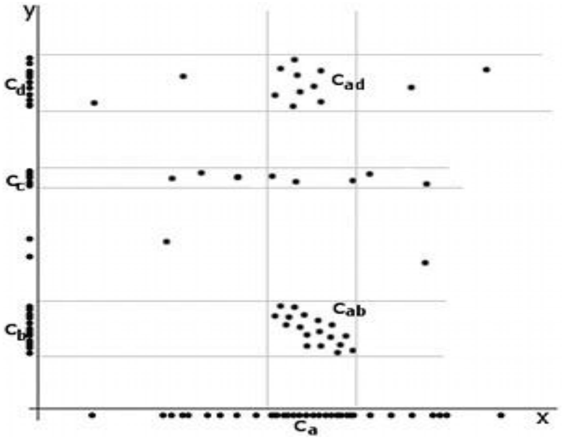

Fig:3 - Example 2D space with subspace clusters

The problem of subspace clustering is given by the fact that there are 2^ddifferent subspaces of a space with $d$dimensions. If the subspaces are not axis-parallel, an infinite number of subspaces is possible. Hence, subspace clustering algorithms utilize some kind of heuristic to remain computationally feasible, at the risk of producing inferior results.

It has been observed that the clustering algorithms should satisfy the following:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- Interpretability and usability.

From the above algorithms it has been found that there are also a number of problems with clustering. They are:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of "distance" (for distance-based clustering);
- if an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces;
- The result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

A comparison between various cluster techniques is discussed in the table: 2 given below:

TABLE 2.Comparison of various cluster techniques based on Methodology, structure, model, application, use

S.no	Algorithm	Types	methodology	Structure	Model	Application
1	Hierarchial	a) Agglomerative(Bottom-up)	Divide and Conquer	Tree structure called as dendrogram	Connectivity based on distance connectivity	Larger clusters
		b) Divisive(Top-down)				Smaller clusters
	Comparison	Agglomerative begins with each element as a separate cluster and merge them into successively larger clusters. Divisive: begin with the whole set and proceed to divide it into successively smaller clusters.				
	Use	Any non-negative-valued function may be used as a measure of similarity between pairs of observations				
2	Partitioning	a)K-means b)K-medoids	The center is the average of all the points/objects in the cluster	spherical shaped clusters in small to medium sized data sets	Centroid	Large datasets
	Comparison	K-medoids method is more robust than k-mean in presence of noise and outliers because a medoids is less influenced by outliers or other extreme values than a mean.				
	Use	Simplicity, speed and can also be used as divisive algorithms in the hierarchical clustering.				
Comparison between hierarchical and partition clustering		Hierarchical Algorithms find successive clusters using previously established clusters; Partitioned algorithms determine all clusters at once.				
3	Density-based	a) DBSCAN b)SNN algorithm	Density of points	Arbitrary-shaped clusters	Density	Density based notion of clusters
	Comparison	The main difference between SNN algorithm and DBSCAN is that it defines the similarity between points by looking at the number of nearest neighbors' that two points share.				
	Use	These methods are devised to discover arbitrary-shaped clusters				
4	Sub space clustering	a)Correlation clustering	Clusters seen in a particular projection (subspace, manifold) of the data.	General Problem	Subspace	Bio-informatics
		b)Bi-Clustering(also known as Co-clustering or two-mode-clustering)		special case of axis-parallel subspaces		
	Comparison	Correlation works with arbitrary feature combinations where as in Bi-Clustering it doesn't however work with.				
	Use	These methods can ignore irrelevant attributes.				
Comparison between hierarchical ,Partition, DBSCAN		DBScan is better than K-means, hierarchical clustering				

V. CONCLUSION

Clustering is a descriptive technique. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery that involves try and failure. It will often be necessary to modify preprocessing and parameters until the result achieves the desired properties. We have presented a comparison of different clustering approaches with regard to various factors and our conclusion is that the approaches regardless of performance, each approach have its own benefits. In this sense our aim is to establish a baseline for further research.

REFERENCES

[1] A. K. Jain and R. C. Dubes, "Data clustering: A review.," *ACM Computing Surveys, vol. 31, 1999*.

[2] RCT Lee Cluster Analysis and Its Applications In J.T.Tou, editor, *Advances in Information Systems Science.Plenum* Press. New York

[3] Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*, 2001 (Academic Press, San Diego, California, USA).

[4] Pham, D.T. and Afify, A.A. Clustering techniques and their applications in engineering. Submitted to *Proceedings of the Institution of Mechanical Engineers, IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725 ISSN: 2250-3021*

[5] *Model-based Methods of Classification*: Using the mclust Software in Chemo metrics Chris Fraley University of Washington AdrianRaftery University of Washington.

[6] P. Zhang, X. Wang, and P. X. Song, "Clustering categorical data based on distance vectors," *The Journal of the American Statistical Association, vol.101,no. 473, pp. 355–367, 2006*.

[7] Athman Bouguettaya "On Line Clustering", *IEEE Transaction on Knowledge and Data Engineering Volume 8, No. 2, April 1996*.

[8] T.Soni Madhulatha , " An overview on clustering methods" *IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725*.

[9] Grabmeier, J. and Rudolph, A. *Techniques of cluster algorithms in data mining.*Data mining and K nowledge discovery,2002,6,303-360

[10] Comparision between clustering algorithms- Osama Abu Abbas. *International journal of Arab Information and Technology,vol.5 No.3,2008*

[11] Bottou, L. and Bengio, Y. *Convergence properties of the k-means algorithm.*

[12] Jain, A.K. and Dubes, R.C.Algorithms for Clustering Data, 1988(Prentice Hall, EngleWood Cliffs, New Jersey,U.S.A)

[13] Mahdi Soltanolkotabi, Ehsan Elhamifar and Emmanuel J. Candes"Robust Subspace Clustering" January 2013

W.Sarada is a research scholar at Rayalaseema University, Kurnool, Andhra Pradesh, India. She is working as an Assistant Professor in the Department of Computer Science at RBVRR Women's College .Her areas of interest include Data Mining, Software Engineering, Digital Image Processing, Computer Net Works.

Dr.P.V.Kumar is a professor at University College of Computer Science and Engineering, Osmania University, Hyderabad. He has vast experience in teaching,guiding and in admin.He is a research supervisor/guide to M.Tech, M.Phil and PhD students.