A Survey of Data Mining Clustering Algorithms

Mihika Shah
Dwarkadas J. Sanghvi College of Engineering
Mumbai-400056
Maharashtra, India

Sindhu Nair
Dwarkadas J. Sanghvi College of Engineering
Mumbai-400056
Maharashtra, India

ABSTRACT
Clustering is a technique used in data mining that groups similar objects into one cluster, while dissimilar objects are grouped into different clusters. The clustering techniques can be categorized into partitioning methods, hierarchical methods, density-based methods and grid-based methods. The different partitioning methods studied here are k-means and k-medoids. The different hierarchical techniques studied here are BIRCH and CHAMELEON. The different grid-based techniques which are described are DBSCAN and DENCLUE. Lastly, the different techniques which are used in grid-based technique, like STING and CLIQUE are described. This paper aims to provide a brief overview and comparison of these different clustering algorithms and methods.

General Terms
Data Mining,Clustering

Keywords
Data mining, clustering, clustering algorithms, clustering methods.

1. INTRODUCTION
Data mining refers to extracting information from large amounts of data, and transforming that information into an understandable and meaningful structure for further use. Data mining is an essential step in the process of knowledge discovery from data (or KDD). It helps to extract patterns and make hypothesis from the raw data. Tasks in data mining include anomaly detection, association rule learning, classification, regression, summarization and clustering [1].

2. CLUSTERING
Clustering is an important technique in data mining and it is the process of partitioning data into a set of clusters such that each object in a cluster is similar to another object in the same cluster, and dissimilar to every object not in the same cluster. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures. Clustering analyses the data objects without consulting a known class label. This is because class labels are not known in the first place, and clustering is used to find those labels. Good clustering exhibits high intra-class similarity and low inter-class similarity, that is, the higher the similarity of objects in a given cluster, the better the clustering. The superiority of a clustering algorithm depends equally on the similarity measure used by the method and its implementation. The superiority also depends on the algorithm's ability to find out some or all of the hidden patterns [2]. The different ways in which clustering methods can be compared are partitioning criteria, separation of clusters, similarity measures and clustering space [3].

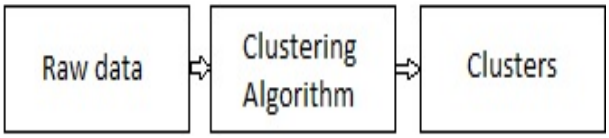

Figure 1: Stages of Clustering [2]

3. CLUSTERING TECHNIQUES
Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms and grid-based algorithms. These methods vary in (i) the procedures used for measuring the similarity (within and between clusters) (ii) the use of thresholds in constructing clusters (iii) the manner of clustering, that is, whether they allow objects to belong to strictly to one cluster or can belong to more clusters in different degrees and the structure of the algorithm [2].

The different clustering techniques are stated as follows:

- Partitioning clustering
  - K- Means
  - K-Medoids
    - PAM
    - CLARA
- Hierarchical clustering
  - Agglomerative
    - BIRCH
    - CHAMELEON
  - Divisive
- Density-based clustering
  - DBSCAN
  - DENCLUE
- Grid-based clustering
  - STING
  - CLIQUE

3.1 Partitioning Clustering
In the partitioning algorithm, the data is split into k partitions, where each partition represents a cluster and k<=n, where n is the number of data points. Partitioning methods are based on the idea that a cluster can be represented by a centre point.

Two conditions that must be satisfied in partitioning algorithms are: (i) At least one object should be present in one group or cluster and (ii) Each object must belong to only one cluster [2] [4].

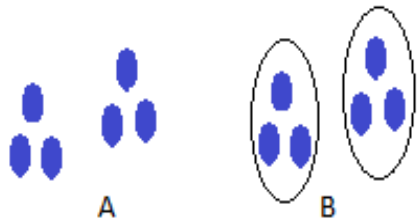

Figure 2: (A) Original Points (B) Partitioning Clustering [1]

3.1.1 K-Means

A cluster is represented by its centroid, which is usually the mean of points within a cluster. The objective function used for k-means is the sum of discrepancies between a point and its centroid expressed through appropriate distance [5]. The time complexity of k-means is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number of iterations [3]. The clusters formed by k-means have convex shapes.

The basic algorithm for k-means clustering is as follows [3]:

a) Arbitrarily choose k objects from D as the initial centers, where k is the number of clusters and D is the data set containing n objects.

b) Repeat

i. (Re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.

ii. Update the cluster means, that is, calculate the mean value of the objects for each cluster.

iii. Until no change.

Disadvantages of k-means

1. Usually terminates at the local optimum, and not the global optimum.
2. Can only be used when the mean is defined and hence requires specifying k, the number of clusters, in advance.

3.1.2 K-Medoids

In this algorithm we use the actual object to represent the cluster, using one representative object per cluster. Clusters are generated by points which are close to respective methods. The partitioning is done based on minimizing the sum of the dissimilarities between each object and its cluster representative [3].

3.1.2.1 PAM

Like all partitioning methods, PAM works in an iterative, greedy way. The initial representative objects are chosen randomly, and it is considered whether replacing the representative objects by non-representative objects would improve the quality of clustering. This replacing of representative objects with other objects continues until the quality cannot be improved further. PAM searches for the best k-medoids among a given data set. The time complexity of PAM is $O(k(n-k)^2)$ [3]. For large values of n and k, this computation becomes even more costly than the k-means method.

Algorithm [3]:

a) Arbitrarily choose k objects in D as the initial representative objects or seeds.

b) Repeat

i) Assign each remaining object to the cluster with the nearest representative object

ii) Randomly select a non-representative object, o_{random}

iii) Compute the total cost, S, of swapping representative object o_j with o_{random}

iv) If $S<0$ then swap o_j with o_{random} to form the new set of k representative objects.

c) Until no change

3.1.2.2 CLARA

CLARA uses 5 samples, each with 40+2k points, each of which are then subjected to PAM, which computes the best medoids from the sample [5]. A large sample usually works well when all the objects have equal probability of getting selected. The complexity of computing medoids from a random sample is $O(ks^2+k(n-k))$, where s is the size of the sample [3]. CLARA cannot find a good clustering if any of the best sampled medoids is far from the best k-medoids.

3.2 Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy (or a tree of clusters), called as dendogram. This method is based on the connectivity approach based clustering algorithms. It uses the distance matrix criteria for clustering the data and constructs clusters step by step [2]. Hierarchical clustering can be categorized into agglomerative (bottom-up) and divisive (top-down). Examples of hierarchical clustering are BIRCH, CHAMELEON, and CURE.

3.2.1 Agglomerative clustering

It is also known as AGNES. An agglomerative cluster starts with singleton clusters and recursively merges two or more most appropriate clusters. The algorithm forms clusters in a bottom-up manner, as follows [5]:

a) Initially, put each article in its own cluster.
b) Among all current clusters, pick the two clusters with the smallest distance.
c) Replace these two clusters with a new cluster, formed by the smallest distance.
d) Repeat the above two steps until there is only one remaining cluster in the pool.

3.2.1.1 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):

BIRCH is an agglomerative hierarchical based clustering algorithm. It is used for clustering large amounts of data. It is based on the notion of a clustering feature (CF) and a CF tree. A CF tree is a height-balanced tree. Leaf nodes consist of a sequence of clustering features, where each clustering feature represents points that have already been scanned. It is mainly used when a small number of I/O operations are needed. BIRCH uses a multi clustering technique, wherein a basic and good clustering is produced as a result of the first scan, and additional scans can be used to further improve the quality of clustering [3]. The time complexity of BIRCH is O(n) where n is number of clusters [3].

3.2.1.2 CHAMELEON

CHAMELEON is an agglomerative hierarchical clustering technique where, unlike other algorithms which use static model of clustering, CHAMELEON uses dynamic modelling to merge the clusters. It does not need users to provide information, but adjusts the clusters to be merged automatically according to their intrinsic properties. It has a

complexity of $O(Nm+Nlog(N)+m^2log(m))$, where N is the number of data points and m is the number of partitions [6]. CHAMELEON uses the interconnectivity and compactness of the clusters to find out the similarity between them. It can be used to find clusters of varying densities [4]. However, the processing time for high-dimensional data may go up to $O(n^2)$.

3.2.2 Divisive clustering

It is also known as DIANA. A divisive cluster starts with one cluster of all data points and recursively splits the most appropriate cluster. This process continues until a stopping criterion (usually the number of requested clusters k) is reached [3]. It uses a top-down strategy. The algorithm for divisive clustering is as follows [5]:

a) Put all objects in one cluster.

b) Repeat until all clusters are singletons

   i) Choose a cluster to split

   ii) Replace the chosen cluster with the sub-cluster.

3.3 Density Based Clustering

Density based algorithms find the cluster according to the regions which grow with high density. They are one-scan algorithms. There are two major approaches for density-based methods. The first approach called the density-based connectivity clustering pins density to a training data point. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The second approach pins density to a point in the attribute space and is called Density Functions. It includes the algorithm DENCLUE.

3.3.1 DBSCAN (Density-Based Spatial Clustering of Application with Noise)

This algorithm is based on the user defined parameters, and on the same database with different parameters, it can create multiple clusters. The number of clusters is not required initially, because it produces the clusters only on the density basis. The data points in DBSCAN fall into three categories: (i) Core points i.e. points that are at the interior of a cluster, (ii) Boundary points i.e. non-core points inside a boundary and (iii) Outliers i.e. points that are neither core nor boundary points [6]. DBSCAN cannot handle clusters of different densities.The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius must contain at least minimum number of points.

3.3.2 DENCLUE (DENsity-based CLUstEring)

Denclue is a clustering method that depends upon density distribution function. DENCLUE uses a gradient hill-climbing technique for finding a local maxima of density functions [6]. These local maxima are called density attractors, and only those local maxima whose kernel density approximation is greater than the noise threshold are considered in the cluster. [3].

3.4 Grid Based Clustering

Grid-based clustering method maps all the objects in a cluster into a number of square cells, known as grids. Grid based clustering has a fast processing time that typically depends on the size of the grid instead of the data.

Unlike other clustering methods which are data-driven, a grid-based clustering uses a space-clustering approach which partitions the space into cells independent of the distribution of input objects.

Grid Density based algorithms require the users to specify a grid size or the density threshold.The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE.

3.4.1 STING (STatistical INformation Grid approach)

This approach breaks the available space of objects into cells of rectangular shapes in a hierarchy. It follows the top down approach and the hierarchy of the cells can contain multiple levels corresponding to multiple resolutions [4]. The statistical information like the mean, maximum and minimum values of the attributes is precomputed and stored as statistical parameters, and is used for query processing and other data analysis tasks. The statistical parameters for higher level cells can be computed from the parameters for lower level cells. The complexity is O (K) where k denotes the total count of cells in last tier [4].

Advantages of STING

1. Query independent.
2. The grid structure facilitates parallel processing and incremental updating.

Disadvantages of STING

1. The quality of the cluster can be degraded as the final cluster does not have any diagonal boundary.

2. THE QUALITY OF THE CLUSTER DEPENDS ON ITS GRANULARITY; IT CAN'T BE TOO COARSE NOR TOO FINE.

3.4.2 CLIQUE

Clique is a grid-based method that finds density-based clusters in subspaces. CLIQUE performs clustering in two steps. In the first step, CLIQUE partitions each dimension into non-overlapping rectangular units, thereby partitioning the entire space of data objects into cells. At the same time it identifies the dense cells in all the subspaces. When the fraction of total data points contained in the unit exceeds the input model parameter then a unit is dense. In the second step, CLIQUE uses these dense cells to form clusters, which can be arbitrary.

Table 1. Comparison of the features of the various clustering algorithms

Algorithm	Scalability and Efficiency	Noise	Shape of cluster	Input data
K-Means	Scalable in processing large datasets.	Sensitive to noise and outliers.	Works well only with clusters of convex shapes	Works only on numerical data.
PAM [3]	Works well for small datasets but not for large datasets.	Not very sensitive to noise and outliers.		Works on data of all attributes.
CLARA [3]	Can deal with larger datasets in comparison to PAM. Efficiency depends on sample size.	Not very sensitive to noise and outliers.		Works on data of all attributes.
BIRCH	One of the best algorithms for large databases in terms of running time, space required, quality, number of I/O operations applied. Shows linear scalability with respect to a number of objects.		Performs clustering well only with spherical data.	Works on data of all attributes.
CHAMELEON			Good at finding clusters of arbitrary shape.	Works on data of all attributes.
DBSCAN	Does not work well for high dimensional data.	Handles noise effectively.	Good at finding clusters of arbitrary shape.	
DENCLUE	Does not work well for high dimensional data.	Invariant against noise.	Can find clusters of arbitrary shape.	
STING				Used mainly with numerical values.
CLIQUE	Scales linearly with the size of the input and shows good scalability when the number of dimensions are increased.			It is not sensitive to input order.

4. CONCLUSION AND FUTURE SCOPE

This paper aims to provide an overview of the algorithms used in different clustering techniques along with their respective advantages and disadvantages. The different clustering methods that have been studied are partitioning clustering, hierarchical clustering, density based clustering and grid based clustering. Under partitioning method, a brief description of k-means and k-medoids algorithms have been studied. In hierarchical clustering, the BIRCH and CHAMELEON algorithms have been described. The DBSCAN and DENCLUE algorithms under the density based methods have been studied. Finally, under grid-based clustering method the STING and CLIQUE algorithms have been described. The challenge with clustering analysis is mainly that different clustering techniques give substantially different results on the same data. Moreover, there is no algorithm present which gives all the desired outputs. Because of this, there is extensive research being carried out in 'ensembles' of clustering algorithms, i.e. multiple clustering techniques done on a single dataset. Along with this, research on improving the existing algorithms is also being carried out.

5. ACKNOWLEDGMENTS

I would like to thank the Department of Computer Engineering and Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.

6. REFERENCES

[1] K. Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies (0975-9646), Vol. 5(2), 2014

[2] Amandeep Kaur Mann and Navneet Kaur, "Review Paper on Clustering Techniques", Global Journal of Computer Science and Technology, Software and Data Engineering (0975-4350), Volume 13 Issue 5 Version 1.0 Year 2013

[3] Han, J. and Kamber, M. Data Mining- Concepts and Techniques, 3rd Edition, 2012, Morgan Kauffman Publishers.

[4] Nisha and Puneet Jai Kaur, "A Survey of Clustering Techniques and Algorithms", IEEE (978-9-3805-4415-1), 2015

[5] Pradeep Rai and Shubha Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications (0975-8887) Vol 7-No. 12, pp. 1-5, October 2010

[6] Pavel Berkhin, "Survey of Clustering Data Mining Techniques", Accrue Software, Inc.

[7] Anil K. Jain, "Data Clustering: 50 years beyond K-means", Pattern Recognition Letters – ELSEVIER, 2009

[8] Suman and Mrs. Pooja Mittal, "Comparison and Analysis of Various Clustering Methods in Data

Mining on Education data set using the weka tool", International Journal of Emerging Trends & Technology in Computer Science (2278-6856), Volume 3, Issue 2, March-April 2014

[9] Namrata S. Gupta, Bijendra S. Agrawal, Rajkumar M. Chauhan, "Survey on Clustering Techniques of Data Mining", American International Journal of Research in Science, Technology, Engineering & Mathematics (2328-3491), 9(3), December 2014-February 2015, pp. 206-211

[10] Anoop Kumar Jain, Prof. Satyam Maheswari, "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research (2278-733X), Vol 1 Issue 1 Aug 2012

[11] Kotsiantis, S. B.; Pintelas, P. E., "Recent Advances in Clustering: A Brief Survey", Department of Mathematics, University of Patras.

[12] Murtagh, Fionn; Contreras, Pedro, "Methods of Hierarchical Clustering", CSIR, Vol 1, pp, 1-21, May 3, 2011.