International Journal of Academic Research and Development
ISSN: 2455-4197, Impact Factor: RJIF 5.22
www.academicsjournal.com
Volume 2; Issue 3; May 2017; Page No. 34-38

Modified clustering algorithms: Survey paper

Mayuri K Botre

Student, Computer Department, D. Y. Patil College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract
Clustering is the process of grouping a set of objects in such a way that object in one group which known as cluster are more similar to each other than the objects in other group. It is used to find structure in unlabeled data. In this paper, we survey on different clustering algorithms. It involves Hierarchical clustering, Grid based clustering, partitioning clustering, Density based clustering, and Model-based clustering. There is also given the combination of these algorithms for better performance.

Keywords: hierarchical clustering, grid-base clustering, partitioning clustering, density based clustering, model-based clustering

1. Introduction

Cluster is nothing but a group of object. In clustering, we create a group of similar object as compare to the object of other cluster. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, information retrieval, data compression, pattern recognition, bioinformatics, image analysis and computer graphics [1].

Our study basically focuses on the different types of clustering algorithms and which algorithm is better. We also have the combination of different clustering algorithms for better performance. There is also some modification is done on the clustering algorithm to improve the performance and efficiency. The clustering methods involves in this paper are:

Hierarchical clustering: Hierarchical clustering involves creating clusters that have predetermined ordering from top to bottom in order to build a hierarchy of clusters.

Grid-based clustering: Grid-based clustering is to partition the whole space into cells which is also known as grids and then merge the cells to build clusters.

Partitioning clustering: A partitioning clustering is nothing but a simply a division of the set of data objects into non-overlapping subsets such that each data object is in exactly one subset. Here subset is a cluster.

Density based clustering: In Density based clustering, given a set of points in some space, it group together those point which are closely packed together i.e. points with many nearby neighbors, marking other points as outlier that lie alone in low-density regions i.e. whose nearest neighbors are too far away.

Model-based clustering: Model-based clustering assumes that the data were generated by a model and tries to recover the original model from the data. The model that we recover from the data then defines clusters and an assignment of objects to clusters.

2. Types of Clustering Algorithms

2.1 Hierarchical clustering

Hierarchical clustering is a well-known clustering method that can be thought as a set of plain clustering methods organized in a tree structural form. These methods construct the clusters by recursively partitioning the data in either a top-down or bottom-up fashion, which is applicable to different domain area. Hierarchical methods are commonly used for clustering in Data Mining problems. Among both hierarchical algorithms, bottom-up approaches tends to be more accurate, but have higher computational cost than top-down approaches [2].

Hierarchical Clustering can be done in three different ways [5]:

- Single-linkage cluster: The distance from one cluster to the other should be shortest. It is also known as connectedness or minimum method.
- Complete-linkage cluster: The distance from one cluster to the other should be greatest. It is also known as diameter or maximum method.
- Average-linkage cluster: The distance from one cluster to the other should be the average distance.

Algorithm of Hierarchical Clustering Algorithm (Single-linkage cluster) [6]

- Assign a cluster to each item, such that N cluster for N items.
- Find and merge the pair of clusters which are closest to each other.
- Calculate the distances between the new and each of the old clusters. (using single-linkage cluster)
  a) Start with the disjoint clustering having level l(0) = 0 and sequence number n = 0.
  b) In the current clustering, now find the least dissimilar pair of clusters say pair (a), (b), according to d[(a), (b)] = min d[(u), (v)] where the minimum is over all pairs of clusters in the current clustering.
  c) Increment the sequence number: n = n + 1 and merge cluster (a) and (b) into a single cluster to form the next clustering n. Set the level of this clustering to l(n) = d[(a), (b)].
  d) Now the next step is to update the proximity matrix, M, by deleting the rows and columns corresponding to clusters (a) and (b) and adding a row and column correspond to the newly formed cluster. The proximity between the new cluster,

denoted (a, b) and old cluster (k) is defined in this way: d[(k), (a,b)] = min d[(k), (a)], d[(k), (b)]

- If all the objects are in one cluster then stop the process else, go to step 2.

Hierarchical method can be subdivided as following [2]

1) Agglomerative hierarchical clustering: A bottom-up approach in which each object initially represents a cluster of its own, and then similar clusters are iteratively getting merged until the desired cluster structure is obtained. This algorithm for N samples begins with N cluster and each cluster contains a single sample. After that two clusters with the nearest similarity will get merge until the number of cluster becomes one. The criteria used in this algorithm are min distance, max distance, average distance, and center distance.

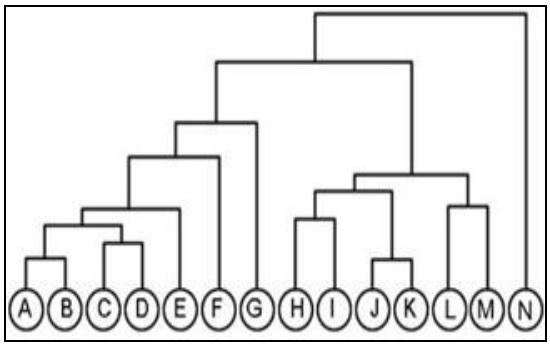

Fig 1: The dendrogram for an agglomerative hierarchical clustering (bottom-up) [2].

2) Divisive hierarchical clustering: A top-down approach that all the objects initially belongs to a single root cluster and iteratively partition existing clusters into sub-clusters.

2.2 Grid-Based Clustering

Grid-based clustering algorithm uses a multi resolution grid data structure. It partitions the space which contain the objects into fixed number of cells and that will form a grid structure on which all the clustering operations are performed. Fast processing time is the main advantage of this approach. A grid-based clustering algorithm consists of the following five steps [7]:

- Creating the grid structure, i.e. partitioning the data space into a finite number of cells.
- Sorting of the cells according to their densities.
- Identifying cluster centers.
- Traversal of neighbor cells.

2.3 Partitioning Clustering

Divide data objects into non-overlapping subsets which is a cluster such that each data object is in exactly one subset. This partitioning method consists of a set of N clusters and each object belongs to only one cluster. There are various types of partitioning methods are [7]:

2.3.1 K-means algorithm

K-means algorithm partitions the data into K clusters (C1; C2;……CK), represented by their centers or means. The center of each cluster is calculates as the mean of all the instances belonging to that cluster. The algorithm starts with an initial set of cluster centers, which are chosen randomly or sometime according to some heuristic procedure. In each iteration. each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated. The center of each cluster is calculated as the mean of all the instances belonging to that cluster:

$$\mu_k = \frac{1}{Nk}\sum_{q=1}^{Nk} xq$$

Where N_k is the number of instances belonging to cluster k and ^1k is the mean of the cluster k.

2.3.2 K-modes algorithm

The K-means algorithm is not effective for categorical data so to solve this problem use K-modes. The different between K-means and K-modes is that K-modes replace the means of clusters with the modes of the clusters. It uses the latest dissimilarity procedures to deal with categorical objects and use a frequency-based method to revise modes of clusters.

3. Modified Clustering Algorithm

3.1 New Hierarchical Clustering Algorithm

The new hierarchical clustering algorithm is a bottom-up agglomerative hierarchical clustering approach. Suppose that a set of points S = $\{p_1,\ldots,p_n\}$ in R^t is given and we want to cluster them. The first step is to find the nearest neighbor for each data to make pairs, then check for those pairs that have a point in common to make primary clusters. The next step is to find the biggest distance (D) in each cluster. So, calculate the mean value for each primary clusters and then measure the distance between mean and all data points of clusters. The final step is to measure the distance between data points of different clusters. If there are two data points from different clusters which their distance is <= (D of cluster 1) or (D of cluster 2), then merge these two clusters [2].

Algorithm of New Hierarchical Clustering Algorithm [2]

1) Find the nearest neighbor for each data point to make pairs, by calculating Euclidean distance between all data points (Fig.4.).

$$D(p, q) = \sqrt{(q1 - p1)^2 + (q2 - p2)^2}$$

2) Merge those pairs that have a point in common to make primary clusters.
3) Calculate the mean (μ) for each primary clusters.
4) Calculate the distance between mean (μ) and all data points of one cluster.

$$D(p, \mu) = \sqrt{(\mu1 - p1)^2 + (\mu2 - p2)^2}$$

5) Find the Maximum value of (d) in each cluster and name it D.
6) Calculate the distance between the data points of different clusters, if there are two points (point 1 from cluster 1 and point 2 from cluster 2) from different clusters which their distance is <= (D of cluster 1) or (D of cluster 2), then merge these two clusters (Fig.5).

3.2 Grid-Based Clustering Algorithm

The authors, Thomas Wagner *et al.,* has presented grid-based clustering algorithm which is based on the ideas of the inner loop of CLIQUE, i.e. the depth-first search, well suitable for radar applications, and optimized towards T low computational burden. There are two modifications done in this algorithm [8] i.e.

1. First is to consider the periodicity and data-alignment of the outcome of the 3D-fast Fourier transform (FFT). Typically at the first index of the outcome of the FFT is the DC component. Then the frequency components are sorted ascending with the index. Thus the negative frequencies are located in the upper half of the FFT's output array. Static target contains the DC value in the Doppler dimension, i.e. v = 0 m/s. The extension step on clusters that contain the DC will result in two distinct clusters: one will contain zero and the positive frequencies and the other one will contain the negative frequencies.
2. Second is to treats a special property of the reflections from a pedestrian. Depending on the resolution of the radar system, as well as orientation and gait of the pedestrian, a single person might appear as several distinct clusters in the range/Doppler/DoA map.

Pseudo-code describing the algorithm is given below. The significance as well as the addressed map are efficiently stored as an $N \times N_C \times N_A$ Boolean matrix. The cluster number of each cell is then stored as entry in an $N \times N_C \times N_A$ integer matrix.

Pseudo-Code Explaining Our Grid-Based Clustering Algorithm [8]

```
Main loop
1.  foreach cell as c
2.    if c is not checked
3.      if c is significant
4.        clusterIndex++
5.        set clusterNumber of c to clusterIndex
6.        expand (c)
7.      else
8.        mark c as checked
9.      end
10.   end
11. end

Expand (a)
1.  foreach cell within k steps to a as c
2.    if c is not checked
3.      mark c as checked
4.      if c is significant
5.        set clusterNumber of c to clusterIndex
6.        expand (c)
7.      end
8.    end
9.  end
```

The result of our algorithm applied to a minimal example is shown in Fig.2. The dark cells are significant cells whereas the white cells are insignificant. The numbers in the cells corresponds to the assigned cluster numbers.

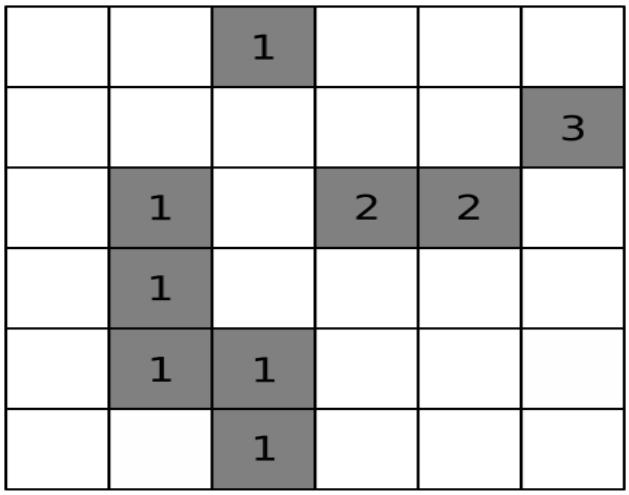

Fig 2: Clustering example: Cells below threshold are drawn white, while cells above threshold are drawn gray. The result map contains the cluster numbers as printed in center of the cells. In this example 3 clusters have been found where cluster on has been wrapped around [8].

3.3 Modified K-medoid algorithm

The K-medoid algorithm is a very popular clustering algorithm. It is having application in image segmentation, data mining, bioinformatics and various other fields. The author Raghvi chouhan and Abhishek Chauhan, proposed an algorithm that will work well with large datasets. Modified K-medoid algorithm reduces the presence of cluster – error criterion and up to some degree it also avoids getting into a locally optimal solution [9].

Algorithm: Modified K-medoid (S, k), S = {x1, x2, …, xn} [9]

Input: The number of clusters k1(k1>k) and a dataset containing n objects (Xij+).

Output: Set of k clusters represented by Cij which minimizes the Cluster – error criterion.

Algorithm (Exixting)

1. Distance between each data point and all other data-points in the set D will be computed.
2. Find the closest pair of data points from the set D and form a data-point set Ax (1<=x<=k) which contains these two data-points, Delete these two data points from the set D.
3. Find the data point in D that is closest to the data point set Ax, Add it to Ax and delete it from D.
4. Repeat step 4 until the number of data points in Ax reaches (n/k).
5. If x<k, then x = x+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ax and delete them from D, Go to step 4.

Sub Algorithms

Algorithm First

- For each data-point set Am (1<=x<=k) find the arithmetic mean of the vectors of data points Cx(1<=x<=k) in Ax.

- Select nearest object of each Cx(1<=x<=k) as initial centroid.
- Compute the distance of each data-point di (1<=i<=n) to all the centroids cj (1<=j<=k) as d(di, cj)
- For each data-point di, find the closest centroid cj and assign di to cluster j.
- Set ClusterId[i]=j; //j:Id of the closest cluster
- Set Nearest_Dist[i]=d(di, cj)
- For each cluster j (1<=j<=k), recalculate the centroids
- Repeat

Algorithm (Improved)

1. for each data-point di
   - Compute its distance from the centroid of the present nearest cluster
   - If this distance is less than or equal to the present nearest distance, the data-point stays in the
   - cluster

   Else
   - For every centroid cj (1<=j<=k) Compute the distance (di, cj); Endfor
   - Assign the data-point di to the cluster with the nearest centroid Cj
   - Set ClusterId[i]=j
   - Set Nearest_Dist[i] = d (di, cj): Endfor
2. For each cluster j (1<=j<=k), recalculate the centroids: until the convergence Criteria is met

Table 1: Comparison on the basis of Execution Time Performed for different values of K (K= Number of Cluster) [9].

No. of Cluster	Time taken to execute (In millisecond) K-Medoids Algorithm	Time taken to execute (In millisecond) Modified K-Medoid Algorithm
3	19723	16432
4	32372	24876
5	61412	43221
6	67562	45102

Table 2: Comparison on the basis of Execution Time Performed for different values of data points N [9].

No. of Cluster	Time taken to execute (In millisecond) K-Medoids Algorithm	Time taken to execute (In millisecond) Modified K-Medoid Algorithm
300	81897	49210
400	99856	73105
500	108352	98530
600	128951	105241

4. Conclusion

In this paper, we present different clustering algorithm used to find structure in unlabeled data. This paper includes hierarchical clustering, grid-based clustering, partitioning clustering, density based clustering and model based clustering. We also presents sub categories of this algorithm. Modified clustering algorithm in addition to comparison with the un-modified clustering algorithm.

5. References

1. Bijuraj LV. Clustering and its Applications, Proceeding of National Conference on New Horizons in IT – NCNHIT, 2013.
2. Zahra Nazari, Dongshik Kang, Reza Asharif M, Yulwan Sung, Seiji Ogawa. A New Hierarchical Clustering Algorithm, ICIIBMS, Track2: Artificial Intelligence, Robotics, and Human-Computer Interaction, Okinawa, Japan, 2015.
3. WANG Hao, LIU TangXing, BU Qing, YANG Bo. An Algorithm based on Hierarchical Clustering for Multi-target Tracking of Multi-sensor Data Fusion, Proceeding of the 35th Chinese conference Chengdu, China. 2016, 27-29.
4. Shiwani Rana, Roopali Garg. Application of Hierarchical Clustering Algorithm to Evaluate Students Performance of an Institute, 2016 Second International Conference on Computational Intelligence and Communication Technology.
5. Punitha SCP, Ranjith Jeba Thangaiah, M. Punithavalli, Performance Analysis of Clustering using Partitioning and Hierarchical Clustering Techniques, International Journal of Database Theory and Application 7. 2014; 6:233-240.
6. A Tutorial on Clustering Algorithms. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html
7. Sukhvir Kaur. Survey of Different Data Clustering Algorithms, International Journal of Computer Science and Mobile Computing. 2016; 5(5):584-588.
8. Thomas Wagner, Reinhard Feger, Andreas Stelzer. A Fast Grid-Based Clustering Algorithm for Range/Doppler/DoA Measurements, Proceedings of the 13th European Radar Conference.
9. Raghavi Chouhan, Abhishek Chauhan. An Ameliorated Partitioning Clustering Algorithm, 2014 Sixth International Conference on Computational Intelligence and Communication Networks.
10. Cichosz P. Data Mining Algorithms Explained Using R, John Wiley & Sons, Ltd. 2015, 349-362.
11. Maimon O, Rokach L. The Data Mining and Knowledge Discovery HandBook, Springer Science + Business Media, Inc. 2005, 321-340.
12. Alpydin E. Introduction to Machine Learning, the MIT Press 2010, 143-158.
13. Masciari E, Mazzeo GM, Zaniolo C. A New, Fast and Accurate Algorithm for Hierarchical Clustering on

Euclidean Distances, Springer-Verlag Berlin Heidelberg, LNAI 7819, 2013, 2:111-114.
14. Pradeep Rai Shubha Singh. A Survey of Clustering Techniques, International Journal of Computer Applications (0975-8887). 2010, 7(12).
15. Ashwini Gulhane, Prashant L. Paikrao DS. Chaudhari. A Review of Image Data Clustering Techniques, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307. 2012, 2(1).
16. Pavel Berkhin. Survey of Clustering Data Mining Techniques, Accrue Software, Inc, 1-56.
17. 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. 2015, 206-111.