STUDY OF A FEW SIGNIFICANT CLUSTERING METHODS

Abstract

Data analysis is an essential tool in contemporary scientific research, including that conducted in the biology sciences, computer sciences, and social sciences. In order to accurately analyze the enormous number of data produced by contemporary applications, clustering algorithms have become a potent alternative meta-learning technique. However, because of the complex nature of the information, each clustering algorithm has sole advantages and disadvantages. This paper presents some significant clustering methods, their technique, performance along with advantages and disadvantages.

Keywords: Partition based clustering, particle swarm intelligence, ant colony optimization, fuzzy-c-means, bee colony optimization.

Author

Srikanta Kumar Sahoo
Department of Computer Science and Engineering
SOA Deemed to be University
Bhubaneswar, India
srikantasahoo@soa.ac.in

I. INTRODUCTION

Late advancements in information technology and innovations in data processing have brought numerous applications in different business areas. Every organization uses a large amount of raw data for different applications. Therefore, the importance of information management has increased. Discovery of knowledge from the huge amount of data used by these organizations is a big challenge. Knowledge discovery is among the notable parts of data mining. The objective of data mining is the extraction of patterns and knowledge from large data sets, not the extraction of data itself [1]. Different data mining tasks include classification, summarization, association rule learning, regression, clustering, and anomaly detection.

One of the important data mining study fields is clustering. Clustering is the process of collecting a set of data objects such that objects in onecluster are progressively comparative to one another than to those in a different cluster [2, 3]. Clustering techniques can be used in many different fields including machine learning, image processing, data compression, pattern recognition, information retrieval, bioinformatics, etc [3]. There are nine categories that the conventional clustering algorithms fall under. Hierarchy and partitioning based are most popular ones along with distribution and density based methods. Again, fuzzy theory, graph theory and fractal theory based methods along with grid and model based methods are also having important characterizations. However there are two key classes: (i) hierarchy based clustering [1, 2, 4] and (ii) partition based clustering [1, 2, 4]. Hierarchical clustering aims at building a hierarchy of data objects/groups. Hierarchical clustering is again of two types: agglomerative or divisive [4]. The agglomerative approach begins with each item framing different groups, and progressively combines the objects/groups near each other until all the groups are converged into one. The divisive method begins with all the items in the same group and progressively splits the group into sub-groups until each item is in one group. Whereas the partition based clustering techniques mostly rely on a distance factor. For *n* objects and *k* clusters, the partitioning technique starts with creating initial partitioning and then in several iterations tries to improve the partitions by moving the objects in between them. Some of the problems with hierarchy based clustering include: once the cluster formation has done it cannot be modified, thus reallocation of objects is not possible; it cannot detect erroneous data, and because of its hierarchical nature, difficulty in handling larger datasets [4]. The partitional clustering overcomes all these problems, but with a limitation that it is not suitable for non-convex data [5].

The modern clustering techniques are basically based on kernels, swarm intelligence, quantum theory, spectral graph theory, affinity propagation, etc [5]. Swarm intelligence-based methods of modern clustering are the most widely employed.. Swarm intelligent algorithms are basically optimization algorithms that have been successfully applied for clustering technique. The fundamental goal of these clustering algorithms is to mimic the biological population's natural process of change. The primary benefit of these algorithms is that they avoid being easily drawn into local optimality while obtaining global optimality. This report is a study of certain significant clustering approaches that are used traditionally along with some modern algorithms along with their benefits and limitations. Here we have studied clustering approaches based on hierarchy, k-means, k-medoids, fuzzy-c-means and density under traditional techniques and approaches based on particle swarm optimization, ant colony optimization, bee colony optimization, whale optimization, kernel, and graph theory under modern techniques.

Following is how the rest of the paper is organized: Sections II and III offer various traditional and modern clustering strategies, respectively, and Section IV concludes up the work with a strategy for further research.

II. TRADITIONAL CLUSTERING TECHNIQUES

1. **Hierarchy Based Clustering:** The essential idea behind this kind of clustering procedure is to organize information by creating hierarchical connections between them [6].It is presumable that every point of data originally represents a different cluster. Then the two clusters that lie nearest to one another are merged to form a single cluster. As an alternative, turn it around.The BIRCH [7], CURE [8], and ROCK [9] are the algorithms that use this kind of clustering. By building the feature tree(CF tree) of clustering, BIRCH realizes the clustering result. If a new data point is received, the CF tree will grow dynamically. CURE, which is appropriate for extensive clustering, uses a sampling method based on randomization to group each model separately before integrating the outcomes. For handling enumeration-type data, ROCK is an enhancement of CURE that takes into account the influence of the data surrounding the cluster on the similarity of the data.Hierarchy-based clustering has the advantage of being able to handle data of any size and shape, but it also has a high time complexity.

2. **K-Means Clustering:** K-Means clustering is used in the fields of data science and machine learning to address clustering problems which is an unsupervised learning algorithm and falls under partitioning technique. The unlabeled dataset is partitioned into K different clusters in this instance and every cluster is given a centroid. The primary goal of this strategy is to decrease the total distances between the data points and the cluster to which they are connected [10].

   The k-means clustering technique primarily accomplishes the following tasks.
   - Start by initializing K points, as initial cluster centroids, at random.
   - Create cluster by assigning the data elements to the nearby centroid with the help of a distance measure.
   - Update the cluster centroid by averaging the data points of the individual clusters.
   - Repeat steps 3 and 4 as necessary to determine the best centroids and ensure that the data objects are being allotted to the right clusters without changing their placement.

   When, n is the number of data elements, k is the number of clusters and t is the number of iterations the time complexity of the algorithms is $O(nkt)$. Hence, it is a faster clustering technique. But the main drawbacks of k-means clustering include its susceptibility to initial set of cluster centers, only being applicable to symmetric data, and tendency to fall in the local optimum regions.

3. **K-Medoids Clustering*:*** The K-Medoids [11] clustering approach creates K clusters from a set of data points utilizing a distance measure.In this case, K-medoids function like K-means clustering and medoid indicates the centroid of a cluster. The key benefit of this approach is its resistance to anomaly for spherical data. More significantly, its

convergence is quicker and with fewer steps. Despite the possibility that it may fail for irregular data, it can be a helpful tool for research across disciplines.

The basic steps for K-medoids clustering is as follows:
- Pick Kmedoids at random.
- To create the initial cluster, assign data objects to the nearby medoids using a distance metric.
- Add up the distances between each cluster and its corresponding medoid.
- Replace the medoids with a different data member within the cluster randomly.
- To obtain an updated clustering result, assign each data point to the nearest medoids.
- Add up the distances between each cluster and its corresponding medoid.
- Update the medoids if the distance summationfornewly created medoids is smaller than that with old medoids.
- Repeat steps 4 to 7 until sum of the distances of old and new medoids are not same.

Empty cluster construction, K-means problem solving, and sensitivity to noise are some of its advantages. Additionally, it chooses the cluster member with the greatest degree of centering. Its drawbacks include the need for accuracy and complexity.

4. **Fuzzy-c-means Clustering*:*** Hard or crisp clustering refers to the clustering methods we've covered up to this point in which just one cluster is assigned to each object. This restriction is eased when using soft clustering or fuzzy clustering, and an object can have some degree of membership in all of the clusters. This is especially helpful when the boundaries between the clusters are unclear and poorly defined. Additionally, the memberships might let us identify more complex connections between a certain object and the exposed clusters. The fuzzy c-means clustering (FCM) approach is the most popular soft clustering technique [12].FCM seeks to minimize the cost function while trying to locate a partition (fuzzy clusters) for a set of data elements.

Each data point's membership score is determined by the FCM clustering algorithm usinga distance measure (Euclidean distance) with its cluster center. The membership ratings are higher for the nearby data points. The membership score and cluster centers are updated after every iteration as follows:

$$a_{ij} = 1/\sum_{k=1}^{c} (dis_{ij}/dis_{ik})^{(2/m-1)}, \tag{1}$$

$$b_j = \frac{\sum_{i=1}^{n} a_{ij}{}^m x_i}{\sum_{i=1}^{n} a_{ij}{}^m}, \tag{2}$$

In this equation, a_{ij} is the membership score,i is data point index,$j$is cluster center index, b_j is the j^{th} cluster center, dis is the Euclidean distance,c is clusters count,m is the fuzziness parameter and n is data points count.

The steps that the FCM clustering algorithm takes to complete its work are as follows:
- Create n random centers to initiate n clusters.
- Utilize (1) to determine the membership score for every data point.
- Up until membership levels exceed a threshold value, repeat steps 4 through 6 as necessary.
- Find the cluster centers using (2).

- Calculate the centroid's Euclidean distance from each data point.
- Utilize (1) to update the membership score for every data point.
- Print the centroids of the clusters.

Due to the need to calculate each data point's membership in each cluster, the technique is comparatively slower and depends on how the weight matrix is initialized.

5. **Density Based Clustering:** The fundamental tenet of these methods of clustering is that all data obtained in an area of dense space for information is considered to belong to the same cluster [13]. Common ones include DBSCAN [14], OPTICS [15], and Mean-shift [16]. Directly derived from the core concept of clustering procedures of this category is the DBSCAN method. The OPTICS technique is an enhancement over DBSCAN process and it fixes the flaw that DBSCAN had in that it was sensitive to the neighborhood radius and the required adequate quantity of points. In mean-shift process, the current data point's mean offset is first determined, followed by the calculation of the next data point's meanusing the current offset value and recent data points, and finally, the process repeated until certain termination measures are met. Benefits include high-efficiency clustering that works with arbitrary-shaped data; and the drawbacks include resulting in poor-quality clustering outcomes when the data space's density is uneven, memory requirement increases when the sizeof data increases, and a clustering outcome that is very parameter-sensitive.

III. MODERN CLUSTERING TECHNIQUES

1. **Particle Swarm Optimization (PSO) Based Clustering:** The concept of the particle swarm optimization [17] is founded on the collective bird-like seeking behavior of swarms. A meta-heuristic population search method is used here. By altering their position and velocity, the particles fly towards the location of food. Every position of the particle represents a possible solution (pbest), and the best position of the particle represents the global solution. Up until the ideal solution (gbest) is discovered, the particles travel. Every pass of the algorithm computes the velocity and updates the locations according to an objective function defined. The following equations can be used to update the velocity and position:

$$V_i^{k+1} = \omega V_i^k + c1 * r1 * \left(P_i^k - X_i^k\right) + c2 * r2 * \left(G_i^k - X_i^k\right) \tag{3}$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \tag{4}$$

Where,V_i,P_i, X_i, G_i, respectively, stand for velocity, pbest, current location, and gbest.The steps in PSO based clustering are as follows:

- Make the cluster centers (Particle) initialized at random.
- For a maximum number of iterations, repeat steps 3 through 7.
- Repeat steps 4 through 7 for every particle.
- Repeat steps 5-7 for each data vector.
- Determine the distance between each data vector and the centroids, then, allocate them to the closest cluster.

- Update pbest and gbest and compute the fitness value.
- Utilizing position formula (4) and velocity (3), update cluster centers.

Better precision, compact clusters, and repeatability are benefits of PSO-based clustering. The limitation is that for huge datasets it requires higher execution time. Again, PSO falls in the local optimum and is dependent on initial cluster centers.

2. **Ant Colony Optimization Based Clustering:** The ant'smovement to different locations is the basis for ant colony optimization. The direction and distance of the pheromone have an impact on this movement [18]. (5) is used to determine whether an ant will migrate to a destination (d) from a source (s) at a time instancet. After some time instancet' the intensity of the pheromone is computed using (6).

$$P_{sd}^{k}(t) = \frac{[\tau_{sd}(t)]^{\alpha}[\eta_{sd}]^{\beta}}{\sum_{k\in\text{allowed}(k)}[\tau_{sk}(t)]^{\alpha}[\eta_{sk}]^{\beta}} \quad \text{if } d \in \text{allowed}(k) \tag{5}$$

$$\tau_{sd}\left(t+t'\right) = (1-\rho)\tau_{sd}(t) + \Delta\tau_{sd} \tag{6}$$

Here, τ_{sd} represents the pheromone's intensity along the path, η_{sd} = 1/ distance_{sd}, and α and β indicate the pheromone's impact. The evaporationrate is represented by ρ, and the overall amount of pheromones released by all ants is referred by$\Delta\tau_{sd}$.Although ant colony optimization (ACO) performs well in the area of discrete problem solving, it unavoidably has some drawbacks. It has good stability, however when working with a lot of data, it has some issues with convergence speed and solution accuracy. ACO has drawbacks including a sluggish rate of convergence, weak similarity, large computational complexity, and a tendency to settle in local optima.

3. **Bee Colony Optimization Based Clustering:** The traditional Bee Colony Optimization (BCO) [19] clustering technique operates in alternating stages.The first one is the forward pass and the other is backward pass. The forward pass's main purpose is to survey the search area and gather useful information. By assessing an objective function, the fitness of these practicable solutions is calculated. In the initial step of the backward pass, each bee chooses whether to stick with its own answer or to adopt one from another. The bee transforms into a recruiter if it chooses to keep going. Other bees in the hive searching for a recruiter are recruited by it, and they all move to where it has been located. If not, the bee chooses a recruiter and follows it in finding a solution.Once the backward pass is finished, all bees have a possible solution. The optimal possible solution among the bees is chosen as the local best (LB). If the strength of the most recent LB is greater than the strength of the most recent global best (GB) solution created up until the preceding iteration, the GB upgrades to the current LB.The algorithm must meet a stopping criteria in order to stop operation.

   The following steps are performed in the algorithm:
   - Initialize all the bees and their clusters.
   - Repeat steps 3 to 6 until termination.
   - Perform forward pass to collect feasible solutions.
   - Perform backward pass and do the following:

- Compute the stickiness probability of bee.
- If the probability of stickiness is less, chose another bee to follow.

- After all phases are finished; assign the remainder of the data objects.
- Update global best.
- Return global best.

The likelihood that a bee will adhere to a solution is determined by the following equation [11].

$$P_b(k+1,t) = e^{-O_b(k,t)/(k\times t)}, \tag{7}$$

$$O_b(k,t) = \frac{SICD_b(k,t) - SICD_{min}(k,t)}{SICD_{max}(k,t) - SICD_{min}(k,t)}, \tag{8}$$

Where, the bee is referred by b, the stage by k, and $t$is iterations. $O_b(k,t)$is the normalized value of $SICD$, $SICD_{max}$, and $SICD_{min}$ are the largest and smallest $SICD$ value. The BCO clustering is renowned for its adaptability, dependability, and capacity to investigate regional solutions. However, it has numerous drawbacks, including a stumpy convergence rate, uneven exploitation and exploration, and delayed sequential processing.

4. **Whale Optimization Algorithm Based Clustering:** The Humpback Whale Optimization Algorithm is founded on how humpback whales hunt. Around 12 metres below the surface, whales dive and blow bubbles all around the object. Afterward, ascend in the water (bubble-net attacking) to assault the target [20]. The bubble-net attack works like this: The whales constrict and surround the prey by referring to its current best location as the target subject's position, while the other whales adjust their places in response.

The steps involved in this clustering algorithm are as follows:

- Chose the centroids (Particle) at random.
- Repeat the steps 3 to 8 until termination.
- Repeat steps 4 to 6 for each particle.
- Repeat steps 5 and 6 for each data vector.
- Determine the distance between each data vector and the centroids, then, allocate them to the closest cluster.
- Determine the fitness value and make updates to the top search agent X*.
- Repeat step 8 for every agent X.
- Using either a spiral move or a shrink, update the location of agent X.

Different areas have favored the whale optimization algorithm (WOA), an advanced optimization method with a straightforward layout. WOA does have certain drawbacks, including a sluggish convergence rate, poor precision, and a tendency to quickly converge to local optimal values.

5. **Kernel Based Clustering*:*** The fundamental tenet of this class of clustering techniques is that nonlinear mapping is used to transfer information from the space of inputs into a high dimension space of features for the cluster study. Kernel K- means [21], kernel SOM [22], and kernel FCM [23] are common clustering techniques. Kernel K-means, kernel SOM, and kernel FCM are algorithms that use the kernel method and transform the

original information into an extremely large feature space. Benefits include: easier clustering in extremely large feature space, suitability for information of any shape, ability to investigate noise and distinguish overlying clusters, and lack of need for prior knowledge of data topology. The clustering outcome is very dependent on the kernel type and its factors, the computational time is considerably high, and the method is not appropriate for handling big amounts of data.

6. **Graph Theory Based Clustering:** The fundamental concept behind these clustering methods is to turn the clustering task into a graph partitioning task by regarding every element as a vertex and the level of similarity over items as a weighted edge. The aim is to discover a strategy for graph partitioning that maximizes the overall weight of connections between the edges inside a group while minimizing the weight of connections between distinct groups. Recursive spectral and multi-way spectral are the two categories into which the standard algorithms for this type of clustering can be broadly subdivided, and the classic algorithms for these two classes are, respectively, SM [50] and NJW [51].The fundamental principle of SM is typically applied to image segmentation, is to reduce the Normalized Cut using an exploratory technique using the eigenvector. In the feature space created by the eigenvectors corresponding to the k greatest eigenvalues of the Laplacian matrix, NJW also performs the clustering analysis. These algorithms are appropriate for the arbitrary-shaped, high-dimensional data set, which converged to the global optimal. The time complexity is relatively large, the clustering outcome is sensitive to the scaling parameter.

IV. CONCLUSION

The aim of this study is to present a summary of the algorithms used in various clustering procedures, alongside each algorithm's benefits and drawbacks. The various clustering techniques that have been studied include density-based, hierarchy based, and partition based under traditional techniques. Additionally, we have provided PSO-based, ACO-based, BCO-based, Whale optimization-based, and kernel-based strategies under modern clustering algorithms. Different clustering algorithms produce noticeably different findings on the same data, which is the major difficulty with clustering analysis. Furthermore, no algorithm is currently available that provides all needed results. Due to this, a significant amount of research is being done.

For application domains like medicine and healthcare, the future work will concentrate on building clustering algorithms based on BCO, Fuzzy-c-means, k-means, and k-medoids. The field of anomaly detection may also be further tested.

REFERENCES

[1] Jain AK. Data clustering: 50 years beyond K-means. Pattern recognition letters. 2010 Jun 1;31(8):651-66.
[2] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on neural networks. 2005 May 9;16(3):645-78.
[3] Oyelade J, Isewon I, Oladipupo O, Emebo O, Omogbadegun Z, Aromolaran O, Uwoghiren E, Olaniyan D, Olawole O. Data clustering: Algorithms and its applications. In2019 19th International Conference on Computational Science and Its Applications (ICCSA) 2019 Jul 1 (pp. 71-81). IEEE.
[4] Han J, Pei J, Tong H. Data mining: concepts and techniques. Morgan kaufmann; 2022 Jul 2.
[5] Xu D, Tian Y. A comprehensive survey of clustering algorithms. Annals of Data Science. 2015 Jun;2:165-93.

[6] Johnson S (1967) Hierarchical clustering schemes. Psychometrika 32:241–254
[7] Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very largedatabases. ACM SIGMOD Rec 25:103–104
[8] Guha S, Rastogi R, Shim K (1998) CURE: an efficient clustering algorithm for large databases. ACMSIGMOD Rec 27:73–84
[9] Guha S, Rastogi R, Shim K (1999) ROCK: a robust clustering algorithm for categorical attributes.In: Proceedings of the 15th international conference on data engineering, pp 512-521
[10] Li Y, Wu H. A clustering method based on K-means algorithm. Physics Procedia. 2012 Jan 1;25:1104-9.
[11] ParkHae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." *Expert systems with applications* 36.2, 3336-3341, 2009.
[12] Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Computers & geosciences. 1984 Jan 1;10(2-3):191-203.
[13] Kriegel H, Kröger P, Sander J, Zimek A (2011) Densitybased clustering. Wiley Interdiscip Rev1:231–240
[14] Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second ACM SIGKDD international conferenceon knowledge discovery and data mining, pp 226–231.
[15] Ankerst M, Breunig M, Kriegel H, Sander J (1999) OPTICS: ordering points to identify the clustering structure. In: Proceedings on 1999 ACM SIGMOD international conference on management of data,vol 28, pp 49–60.
[16] Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEETrans Pattern Anal Mach Intell 24:603–619.
[17] W. Liu, Z. Wang, X. Liu, N. Zeng, and D. Bell, “A novel particle swarm optimization approach for patient clustering from emergency departments,” IEEE Transactions on Evolutionary Computation, vol. 23, no. 4, pp. 632–644, 2018.
[18] İnkaya T, Kayalıgil S, Özdemirel NE. Ant colony optimization based clustering methodology. Applied Soft Computing. 2015 Mar 1;28:301-11.
[19] R. Forsati, A. Keikha, and M. Shamsfard, “An improved bee colony optimization algorithm with an application to document clustering,” Neurocomputing, vol. 159, pp. 9–26, 2015.
[20] Nasiri, J., Khiyabani, F.: A whale optimization algorithm (WOA) approach for clustering.Cogent Math. Stat. ISSN: 2574–2558 (2018).
[21] Schölkopf B, Smola A, Müller K (1998) Nonlinear component analysis as a kernel eigenvalue problem.Neural Comput 10:1299–1319.
[22] MacDonald D, Fyfe C (2000) The kernel self-organising map. Proc Fourth IntConfKnowl-BasedIntellEngSyst Allied Technol 1:317–320
[23] Wu Z, XieW,Yu J (2003) Fuzzy c-means clustering algorithm based on kernel method. In: Proceedingsof the fifth ICCIMA, pp 49–54.