Data Clustering Techniques and Methods: A Comprehensive Literature Review

Lana M. Ilawi[1], Wael M. S. Yafooz[2]

[1] *Faculty of Computer and Information Technology, Al-Madinah International University, Malaysia, lana_ialwi@yahoo.ca*
[2]*Faculty of Computer and Information Technology, Al-Madinah International University, Malaysia, wael.mohamed@mediu.edu.my*

Received 30 July 2017; accepted 20 October 2017

Abstract

In the literature there are many researches dealing with the data clustering techniques and methods. In this paper, a detailed explanation of the current methods of data clustering will be presented. However, we will discuss a number of the common used data clustering techniques and methods

Keywords: (Data Clustering, Partitional Clustering, Hierarchical Clustering, Density-Based Clustering Algorithms, Categorical Clustering Methods, Grid-based Clustering, Correlation Clustering, Spectral Clustering; Gravitational Clustering, Herd Clustering)

1. Introduction

In literature there are many methods of data clustering and data classifications. In this paper, we will focus on the current and common methods. First of all, the different definitions of the data clustering and data classification will be introduced from different researchers' points of view.

The following are some definitions of data clustering:

1. Jain et al. [1] define the data clustering as "clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)".
2. Jain and Dubes [2] define it as "cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments".
3. Jain [3] defines the clustering as "given a representation of n objects, find K groups based on a measure of similarity such that objects within the same group are alike but the objects in different groups are not alike".

Within this paper, the following data clustering methods and techniques will be discussed in details:

1. Partitional Algorithms
1.1 k-Means
1.2 Partitioning Around Medoids (PAM)
1.3 Clustering Large Applications (CLARA)
1.4 Clustering Large Applications based on RANdomized Search (CLARANS)
2. Hierarchical Algorithms
2.1 Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)
2.2 Clustering Using Representatives (CURE)
3. Density-Based Clustering Algorithms
3.1 Density-based spatial clustering of applications with noise (DBSCAN)
3.2 DENsity CLUstEring (DENCLUE)
3.3 Ordering Points To Identify the Clustering Structure (OPTICS)
4. Categorical Clustering Methods
4.1 K-Prototypes
4.2 RObust Clustering using linKs (ROCK)
4.3 Sieving Through Iterated Relational Reinforcement (STIRR)
4.4 Clustering Categorical Data Using Summaries (CACTUS)
4.5 Expectation-Maximization (EM)
5. Grid-based Clustering
6. Correlation Clustering
7. Spectral Clustering
8. Gravitational Clustering
9. Herd Clustering

Within the next section the details of the above methods and techniques will be discussed. However, section 2 explains the partitional algorithms; section 3 presents hierarchical algorithms, section 4 presents the density-based clustering algorithms, section 5 describes the categorical clustering methods, section 6 discusses the Grid-based Clustering , section 7 explains the correlation clustering, section 8 presents the spectral clustering, section 9 illustrates the gravitational clustering, and section 10 discusses the herd clustering method. Finally, section 11 discusses the results of this literature review.

2. Partitional Algorithms

Clusters are framed by taking after either a base up approach or a top-down approach. For instance, single-linkage clustering [4] is a great base up approach in which information focuses are step by step agglomerated together to shape groups. In each progression, all combine insightful separations are figured to recognize the base. The gatherings required in the insignificant combine insightful separation are connected together [4]. Such a stage is rehashed until all information focuses are connected together. A progressive tree is developed to interface all information focuses toward the end. A tree profundity level can be cut the tree, framing clusters. To model information progressively, an uncommon various leveled grouping technique called Chameleon has been proposed [5]. It makes utilization of the between network and closeness idea to union and gap groups. In the event that the between network and closeness between two groups are higher than those inside the clusters, then the two groups are combined [5]. Next, in this section we will discuss some of these algorithms.

2.1 k-Means

Clustering is one of the unsubstantiated learning method in which a set of fundamentals is separated into uniform groups. The k-means method is one of the most common and intensively used clustering methods for different applications [6]. However, K-means clustering [7] is a partition-based cluster analysis method [6].

The goal of the K-means clustering algorithm is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The K-means clustering algorithm consists of the following two separate phases:

1 Define k centroids, one for each cluster.

2 Take each point belonging to the given data set and associate it to the nearest centroid.

However, Euclidean distance is generally considered to determine the distance between data points and the centroids. When most of the points are available in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once k new centroids was found, a new binding is to be made between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for clustering [8]. Generally speaking, the k-means algorithm is one of the most widely studied clustering algorithm and is generally operative in producing good results. The main drawback of the k-means algorithm is that it gives different clusters for various sets of values of the initial centroids. The final clusters' quality is heavily based on the selection of the initial centroids. The k-means algorithm is computationally costly and time consuming to the product of the number of data items, number of clusters and the number of iterations [8].

2.2 Partitioning Around Medoids (PAM)

The Partitioning Around Medoids (PAM) is one of the most well-known versions of K-medoids algorithms [9]. PAM is iterative optimization that merges relocation of points between perception clusters with re-nominating the points as potential medoids [9]). The PAM algorithm can be summarized is as follows [9]:

1. Choose K of the n data points randomly, make them as the medoids.
2. Associating each data point to the nearest medoid.
3. Loop for each medoid m
4. Loop for each non-medoid data point

(i) o Swap m and o

(ii) Compute the total cost of the configuration

5. Make a configuration by selecting the lowest cost.
6. Repeat steps 2 to 4 until there is no change in the medoid.

2.3 Clustering Large Applications (CLARA)

Clustering Large Applications (CLARA) has been designed by Kaufman and Bousseeuw [10] to handle the huge data sets, it is completely relies on sampling [10]. As a replacement of finding representative objects for the entire data set, CLARA draws a sample of the data set, applying PAM on the sample, and finding the medoids of that sample. However, the point here is, if the sample is drawn in a suitably random way, the medoids of the sample would estimate the medoids of the whole data set. Thus, to get better approximations, CLARA draws multiple samples and produces the best clustering as an output. Here, for accurateness, the quality of a clustering can be measured based on the average difference of all objects in the whole data set, and not only of those objects in the samples. Kaufman and Bousseeuw [10] in their experiments indicate that five samples of size 40 + 2L give satisfactory results. In addition, CLARA is not designed for small data sets [11].

2.4 Clustering Large Applications based on RANdomized Search (CLARANS)

Clustering Large Applications based on RANdomized Search (CLARANS) consists of both CLARA and PAM in terms of their efficiency and effectiveness [12]. It can be used to offer active spatial data mining [12]. Furthermore, it is more efficient than PAM and CLARA for both small and large data sets [12]. In addition, it is the best and common algorithm used when considering outlier detection [13].

CLARANS begins with the randomly selection of medoids. Then, it dynamically draws the neighbor. Also, it checks "maxneighbour" for swapping. If find negative pair, then it selects another medoid set. Otherwise, it selects existing selection of medoids as local optimum and it randomly begins with the new selection of medoids. It

stops the process when returns the best [13]. CLARANS algorithm can be summarized as the following [13]:

1. Input parameters numlocal and maxneighbour.
2. Select k objects from the database object D randomly.
3. Mark these K objects as selected Si and all other as non-selected Sh.
4. Calculate the cost T for selected Si
5. If T is negative update medoid set. Otherwise selected medoid chosen as local optimum.
6. Restart the selection of another set of medoid and find another local optimum.
7. CLARANS stops when returns the best.

3. Hierarchical Algorithms

Information is isolated into non-covering subsets with the end goal that every information occasion is doled out to precisely one subset [14]. The initial step is to pick the methods for groups as the centroids, while the second step is to allot information focuses to their closest centroids. Its computational speed and straightforwardness request to individuals [14, 15]. Its fundamental downside is the weakness to its irregular seeding method. As such, if the underlying seeding positions are not picked effectively, the clustering result quality will be influenced unfavorably [14]. In light of that, Arthur and Vassilvitskii [16] proposed a technique called k-means++ to enhance k-implies in 2007. From Arthur and Vassilvitskii work in [16], we can watch that the means 2-4 of k-means++ are precisely the same as those of k-means. The primary distinction lies in the progression 1 which is the seeding method. Another seeding method is proposed to supplant the self-assertive seeding strategy of k-mean. Given an arrangement of seeds picked, the seeding method supports the information focuses which are a long way from the seeds as of now picked. In this manner the seeds are picked probabilistically as scattered as could be allowed. As k-means++ is the augmented adaptation of k-means technique, we led numerical examinations to assess and think about their execution under 1000 recreate runs [16].

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) has been proposed by Tian Zhang et al. [17] as an agglomerative hierarchical clustering method, and they verified that it was especially appropriate for large databases. BIRCH can usually produce a good cluster with a single scan of the data, and enhance the quality additionally with a few further scans of the data. In addition, BIRCH was the first clustering algorithm proposed in the database area that can solve the produced noise effectively. Zhang et al. [17] also evaluated efficiency the time and space of the BIRCH's, the order data input sensitivity, and cluster quality through several experiments [18].

Clustering Using Representatives (CURE)

Clustering Using Representatives (CURE) uses multiple representative points for each cluster that are produced by choosing well-scattered points from the cluster and then decrease them toward the center of the cluster by a stated fraction. In this way, CURE is enabled to be well-adjusted to the geometry of clusters which having non-spherical shapes and wide variances in size. To deal with large databases, CURE identifies a combination of random sampling and partitioning which lets it to hold large data sets efficiently. However, random sampling, coupled with outlier handling techniques, also makes it promising for CURE to filter outliers enclosed in the data set effectively. In addition, the labeling algorithm in CURE uses various random representative points for each cluster to allocate data points on disk. This allows it to properly label points even when the shapes of clusters are non-spherical and the sizes of clusters differ. For a random sample size of s, the time complexity of CURE is 0(s2) for low-dimensional data and the space complexity is linear in s [19].

4. Hierarchical Algorithms

Aside from the outstanding grouping strategies, there are distinctive clustering standards. In thickness based grouping, information is clustered in light of some availability and thickness capacities. For instance, DBscan [20] utilizes thickness based ideas to characterize clusters. Two availability capacities thickness reachable and thickness associated have been proposed to characterize every information point as either a center point or an outskirt point. DBscan [20] visits focuses subjectively until the sum total of what focuses have been gone to. On the off chance that the fact of the matter is a center point, it tries to extend and frame a group around itself. In view of the test comes about, the creators have shown its vigor toward finding self-assertively formed clusters [20].

4.1 Density-based spatial clustering of applications with noise (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) depends on a density-based notion of clusters which is designed to discover clusters of arbitrary shape [21]. DBSCAN needs only one input parameter and supports the user in defining a suitable value for it. The results of their experiments confirmed that [21]:

1. DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS.
2. DBSCAN outperforms CLARANS by factor of more than 100 in terms of efficiency.

However, DBSCAN needs only one input parameter and supports the user in defining a suitable value for it. It realizes clusters of arbitrary shape. Finally, DBSCAN is also efficient for large spatial databases [21].

4.2 DENsity CLUstEring (DENCLUE)

The main concepts of the Density based clustering (DENCLUE) [22] are the influence and density functions.

However, the influence of each data point can be showed as mathematical function and resulting function is called Influence function. The Influence function defines the influence of data point within its neighborhood. While, the density function is the sum of influence of all data points. Within DENCLUE there are two defined types of clusters, which are, center defined and multi center defined clusters .In center defined cluster a density attractor. The influence function of a data objects y Є F is a function. Which is defined in terms of a basic influence function F, F (x) =- F (x, y). The density function is defined as the sum of the influence functions of all data points. Furthermore, DENCLUE generalizes other clustering methods such as density based clustering algorithm, partition based clustering algorithm, and hierarchical clustering algorithms [22].

4.3 Ordering Points To Identify the Clustering Structure (OPTICS)

Ordering Points To Identify the Clustering Structure (OPTICS) can be identified as a generalization of DBSCAN to various ranges, effectively substituting the ε parameter with a maximum search radius [23]. OPTICS is an algorithm for discovering density-based clusters in spatial data. Its main idea is similar to DBSCAN, but it discourses one of DBSCAN's major weaknesses, that is, the problem of perceiving meaningful clusters in data of changing density. For this, the points of the database are linearly ordered such that points which are spatially closest become neighbors in the ordering. Furthermore, a special distance is kept for each point that denotes the density that required to be accepted for a cluster in order to have both points belongs to the same cluster [24]. OPTICS generalizes DB clustering by creating an ordering of the points that allows the extraction of clusters with arbitrary values for ε [23]. There are many textual document clustering as proposed in [48, 49, 50, 51]

5. Categorical Clustering Methods

5.1 K-Prototypes

The K-Prototype algorithm is one of the most important algorithms for clustering the data objects when they are described by both numeric and categorical features. Huang [25] proposed k-prototypes algorithm. This algorithm is completely based on the k-means paradigm, but it differs in removing the numeric data limitation whilst conserving its efficiency. However, it combines both the K-Means and K-Modes processes to cluster data with mixed numeric and categorical values. The random selection of starting centroids in these algorithms may lead to different clustering results and falling into local optima [26].

5.2 RObust Clustering using linKs (ROCK)

The ROCK algorithm is a strong clustering algorithm in which it connects the distance based on the notion of links and the number of links between two tuples which is the number of common neighbors they have in the dataset through the cluster [27]. These are argued with the non-metric similarity measures which are based on the relevant situations. ROCK clustering has been developed to lessening the query response time by searching the documents in the resulted clusters instead of searching the whole database [28]. When using ROCK, Guha et al. [29] produces better quality result comparing to the traditional methods. In addition, The ROCK algorithm is operative on vector epitomizes a tuple in the data where the entries are recognizing as categorical values [30].

5.3 Sieving Through Iterated Relational Reinforcement (STIRR)

One of the most powerful methods is the Sieving Through Iterated Relational Reinforcement (STIRR) method [31]. This method uses an iterative approach when the data objects are being similar and a large overlap appear in the database items [31]. In addition, this method has the following key features [31]:

1. No priori quantization.
2. Define the similarity between database items even to items that never occur together in a tuple.
3. Viewing each tuple in the database as a set of values.

5.4 Clustering Categorical Data Using Summaries (CACTUS)

Clustering Categorical Data Using Summaries (CACTUS) algorithm is considered a fast summarization-based algorithm since its purpose is construct a summary information from the dataset which is necessary for discovering well-defined clusters [32]. However, it has the following two important characteristics [32]:

1- Two scan of dataset are required, and it is very fast and scalable.

2- Find clusters in subsets of all attributes and can thus perform a subspace clustering of the data.

5.5 Expectation-Maximization (EM)

The Expectation-Maximization algorithm is used in maximum likelihood estimation where the problem includes two sets of random variables of which one X, is observable and other Z, is hidden [33]. The EM algorithm is summarized in the following to steps [33]:

1. Estimates the expectancy of the missing value by unlabeled class information. This step works in performing classification of each unlabeled document. This step called E-Step.
2. Maximizes the likelihood of the model parameter using the earlier computed expectation of the missing values as if were the true ones.

So far, EM algorithm is applied with administered approach. However, the disadvantages of this algorithm are [33]:

1. The whole data must be labeled.
2. There are no new classes which dynamical generation would be there.
3. It is time consuming and leads to decrease in classification speed.

6. Grid-based Clustering

In framework based grouping, the information space is partitioned into numerous bits (grids) at various granularity levels to be clustered independently [34]. For instance, Inner circle [34] can consequently discover subspaces with high thickness clusters. No information dispersion supposition has been made. The exact outcomes exhibited that it could scale well with the quantity of measurements. Consequently it is particularly proficient in grouping high-dimensional information [34].

7. Correlation Clustering

Correlation clustering [35] was propelled from an archive clustering issue in which one has a couple savvy closeness work f gained from past information. The objective is to parcel the present arrangement of records in a way that connects with f however much as could reasonably be expected. At the end of the day, there are a total diagram of N vertices, where each edge is marked either + or −. We will probably deliver a parcel of vertices (a clustering) that concurs with the edge names. The creators have demonstrated that this issue is a NP-finish issue. Consequently they proposed two estimation calculations to accomplish the dividing [35]. The main strategy called Cautious is to limit the contradictions (number of − edges inside clusters in addition to the quantity of + edges between clusters), while the second technique called PTAS is to augment the assertions (number of + edges inside clusters in addition to the quantity of − edges between clusters) [35]. Fundamentally, the thoughts of the over two techniques are the same (to total the vertices which concur with their edge names). The main strategy is examined in detail in this work [35].

To start with, we subjectively pick a vertex v. At that point we get all the positive neighbors (the neighbor vertices with + edge) of the vertex and place them into a set A. Having grabbed all the positive neighbors of the vertex, we perform pruning. That is the 'Vertex Removal Step'. In this progression, we proceed onward to check 3δ-badfor all the positive neighbors of the vertex, where $\delta = 1/44$. On the off chance that there are, we expel it from the set A. After the evacuation step, the following stride is 'Vertex Addition Step' in which we attempt to include back some vertices which are 7δ-good with the picked vertex v to the set A. The vertices in the set A are then picked as one group. The above strides are rehashed until no vertices are left or the set A ends up plainly unfilled [35].

8. Spectral Clustering

A portion of the current grouping methodologies may discover neighborhood minima and require an iterative calculation to discover great clusters utilizing distinctive introductory group beginning stages. Interestingly, spectral clustering [36-38] is a moderately encouraging methodology for clustering in light of the main eigenvectors of the matrix got from a separation framework. The primary thought is to make utilization of the range of the similarity matrix of the information to perform dimensionality decrease for k-implies grouping in less measurements. The fundamental work [36] is talked about in this work.

Toward the starting, we frame an affinity matrix A, which is an NxN matrix and N is the aggregate number of data points. Every section Aij relates to the similitude measure between the data points si and sj . The scaling parameter $\sigma 2$ controls how quickly Aij tumbles off with the separation amongst si and sj. After we have framed the affinity matrix A, we develop the Laplacian framework L from the standardized affinity matrix of A. At that point we discover the k driving eigenvectors (i.e. with k driving eigenvalues) of L and shape the framework X by stacking the eigenvectors in section. After we have stacked the eigenvectors to frame the framework X, we standardize each line. At that point we regard each line in X as an information vector and utilize k-implies clustering calculation to group them. The clustering results are anticipated back onto the original data (i.e. it allocates the original point si toward cluster j if and just if push i of the matrix X is doled out to cluster j) [36].

9. Gravitational Clustering

Unmistakable from the works we have specified gravitational grouping is considered as a somewhat extraordinary method. It was first proposed by Wright [39]. In his method, every information example is considered as a molecule inside the component space. A physical model is connected to mimic the developments of the particles. As portrayed in [40], another gravitational clustering technique utilizing Newton laws of movement has been proposed. A rearranged rendition of gravitational clustering was proposed by Long et al. [41]. Wang et al. [42] proposed a nearby contracting method to move information toward the medians of their k closest neighbors. Blekas and Lagaris [43] proposed a similar method called Newtonian Clustering in which Newton's conditions of movement are connected to therapist and separate information, trailed by Gaussian blend show building. Sub-atomic progression like system was likewise connected for Clustering by Junlin et al. [44].

10. Hard Clustering

To handle the clustering issue, a novel clustering method, Hard Clustering (HC), has been proposed by Wong et al. [45]. It curiosities lie in two perspectives:

1. HC is enlivened from the nature, group conduct, which is a regularly observed wonder in this present reality including human versatility designs [46]. Along these lines it is extremely instinctive and simple to be comprehended for its great execution.

2. HC likewise exhibits that cluster analysis should be possible in a non-conventional manner by making information alive [46].

HC varies from the conventional ones. Rather than making a decent attempt to break down information alone, it additionally spends exertion on moving information. Two phases are proposed in HC. Motivated by the herd behavior [47], a fascination model is utilized to guide data movements in the main stage. Every data instance is spoken to by a molecule. The organize position of a molecule is given by the estimations of the corresponding data instance it speaks to. The particles draw in each other if their separations are littler than a limit. Every particle has its own speed (initially zero). In every emphasis, the speed of a molecule is influenced by the area particles. On the off chance that most particles are found in a specific course, the speed of the molecule is quickened toward that bearing [47].

After every one of the emphases in the principal organize, all information examples ought to be very much isolated and consolidated. They are significantly less demanding to be grouped than some time recently. Along these lines a natural approach is proposed to cluster information in the second stage. A rundown of cluster centroids is kept up. Toward the starting, the centroid rundown is void. For each point, we check whether its separation to any centroid is littler than the limit. In the event that a centroid is detected, at that point the fact of the matter is allocated an indistinguishable cluster from the centroid. On the off chance that its separations to all centroids are higher than or equivalent to the edge, the fact is added to the rundown and begins another cluster around it. After all information examples are filtered, a clustering result is acquired. At the main look, HC is like Gravitation Clustering (GC) [39]: information occasions are moved by a model. Regardless, their points of interest are entirely unexpected. For example, the model in GC is a physical model after Newton Laws of movement, while that in HC is a simulated model which is intended for computational effectiveness. The molecule speeding up abatements as the between molecule remove increments in GC while they are free in HC. Math is included in GC though just computationally proficient operations are permitted in HC [47].

11. Discussion

In literature there are many researches dealing with the data clustering techniques and methods. In this paper, a detailed explanation of the current methods of data clustering was presented. However, we have discussed twenty-four data clustering techniques and methods.

Many of the discussed data clustering techniques and methods are coming in their usage and application, for example:

1 K-Means

2 K-Means++ (only changes how to initialize centroids)

3 CURE

4 DBSCAN

5 STING

6 K-Prototypes

7 OPTICS

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, pp. 264-323, September 1999.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*: Prentice-Hall, 1988.

[3] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," presented at the The 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, USA, 2008.

[4] R. Xu and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, pp. 645-678, May 2005.

[5] G. Karypis, E.-H. S. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, pp. 68-75, August 1999.

[6] S. Shinde and B. Tidke, "Improved K-means Algorithm for Searching Research Papers," *International Journal of Computer Science & Communication Networks*, vol. 4, pp. 197-202, 2014.

[7] A. Alrabea, A. V. Senthilkumar, H. Al-Shalabi, and A. Bader, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with PCA " *Journal of Advances in Computer Networks*, vol. 1, June 2013.

[8] K. A. A. Nazeer and M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," in *Proceedings of the World Congress on Engineering (WCE'09)*, London, U.K., 2009.

[9] F. B. A. Abid, "A Novel Approach for PAM Clustering Method," *International Journal of Computer Applications*, vol. 86, pp. 1-5, January 2014.

[10] L. Kaufman and P. J. Rousueeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*: John Wiley & Sons, 1990.

[11] R. Ng and J. Han., "Effective and Eflective Clustering Methods for Spatial Data Mining," BC, Canada.

[12] S. Vijayarani and S. Nithya, "An Efficient Clustering Algorithm for Outlier Detection," *International Journal of Computer Applications*, vol. 32, pp. 22-27, October 2011.

[13] R. T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," in *The 20th VLDB Conference Santiago*, Chile, 1994, pp. 144-155.

[14] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: A data-distribution perspective," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, pp. 318-331, 2009.

[15] G. Stockman and L. G. Shapiro, *Computer Vision*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[16] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," presented at the the 18th annual ACM-SIAM symposium on Discrete algorithms,, Philadelphia, PA, USA, 2007.

[17] T. Zhang, R. Ramakrishnan, and MironLinvy, "BIRCH: An eEficient Data Clustering Method for Large Databases," in *International Conference on Management of Data (ACM-SIGMOD'96)*
Montreal, Quebec, 1996

[18] Y. Rani and H. Rohil, "A Study of Hierarchical Clustering Algorithm," *International Journal of Information and Computation Technology*, vol. 3, pp. 1225-1232, 2013.

[19] S. Guha, R. Rastogi, and K. Shims, "Cure: An Efficient Clustering Algorithm for Large Databases," *Information Systems*, vol. 26, pp. 35-58, 2001.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," presented at the 2nd International Conference on Knowledge Discovery and Data Mining, 1996.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *The Second International Conference on Knowledge*

Discovery and Data Mining (KDD'96), Portland, Oregon, USA, 1996, pp. 226-231.

[22] A. Hinneburg and D. Keim, "An efficient approach to clustering Large multimedia databases with noise," in *The 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, 1998, pp. 58-65.

[23] R. Prabahari and V. Thiagarasu, "A Comparative Analysis of Density Based Clustering Techniques for Outlier Mining," *International Journal of Engineering Sciences & Research Technology*, vol. 3, pp. 132-136, November 2014.

[24] A. Ram, S. Jalal, A. S. Jalal, and M. kumar, "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases," *International Journal of Computer Application*, vol. 3, June 2010.

[25] I. Ahmad, "K-Mean and K-Prototype Algorithms Performance Analysis," *American Research Institute for Policy Development*, vol. 2, pp. 95-109, 2014.

[26] K. A. Prabha and N. K. K. Visalakshi, "Particle Swarm Optimization based K-Prototype Clustering Algorithm," *IOSR Journal of Computer Engineering*, vol. 17, pp. 56-62, April 2015.

[27] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," in *The 15th International Conference on Data Engineering (ICDE'99)*, Sydney, Australia, 1999, pp. 512-521.

[28] A. Tyagi and S. Sharma, "Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 4, pp. 809-815, 2012.

[29] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Stanford University, Stanford, USA2007.

[30] T. V. Jagatheesan S.M, "Design of a FUZZY logic based Categorical Text Clustering Algorithm for Information Retrieval," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, pp. 981-984, January 2014.

[31] S. Sajikumar and E. Ramadevi, "Clustering Algorithms on Data Mining in Categorical Database," *International Journal of Computer Systems*, vol. 3, pp. 244-247, March 2016.

[32] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS–Clustering Categorical Data Using Summaries," in *The 5th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'99)*, San Diego, California, USA, 1999, pp. 73-83.

[33] B. Nigam, P. Ahirwal, S. Salve, and S. Vamney, "Document Classification Using Expectation Maximization with Semi Supervised Learning," *International Journal on Soft Computing (IJSC)*, vol. 2, pp. 37-44, 2011.

[34] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," *SIGMOD Rec.*, vol. 27, pp. 94-105, June 1998.

[35] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning Journal,*, vol. Special Issue on Theoretical Advances in Data Clustering, pp. 86-113, 2004.

[36] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processingm Systems.*, ed: MIT Press, 2001, pp. 849-856.

[37] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal Mach. Intell.*, vol. 22, pp. 888-905, August 2000.

[38] M. Maila and J. Shi, "A random walks view of spectral segmentation," presented at the AI and STATISTICS (AISTATS), 2001.

[39] W. Wright, "Gravitational clustering," *Pattern Recognition*, vol. 9, pp. 151-166, 1977.

[40] J. Gomez, D. Dasgupta, and O. Nasraoui, "A new gravitational clustering algorithm," presented at the the SIAM Int. Conf. on Data Mining (SDM), 2003.

[41] T. Long and L.-W. Jin, "A new simplified gravitational clustering method for multi-prototype learning based on minimum classification error training," in *Advances in Machine Vision, Image Processing, and Pattern Analysis, ser. Lecture Notes in Computer Science*. vol. 4153, N. Zheng, X. Jiang, and X. Lan, Eds., ed Berlin / Heidelberg: Springer, 2006, pp. 168-175.

[42] X. Wang, W. Qiu, and R. H. Zamar, "Clues: A non-parametric clustering method based on local shrinking," *Computational Statistics & Data Analysis*, vol. 52, pp. 286-298, September 2007.

[43] K. Blekas and I. E. Lagaris, "Newtonian clustering: An approach based on molecular dynamics and global optimization," *Pattern Recogn.*, vol. 40, pp. 1734-1744, June 2007.

[44] L. Junlin and F. Hongguang, "Molecular dynamics-like data clustering approach," *Pattern Recognition*, vol. 44, pp. 1721-1737, 2011.

[45] K.-C. Wong, C. Peng, Y. Li, and T.-M. Chan, "Herd clustering: A synergistic data clustering approach using collective intelligence," *Applied Soft Computing*, vol. 23, pp. 61-75, 2014.

[46] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Lio, "Collective human`mobility pattern from taxi trips in urban area," *PloS one*, vol. 7, pp. 434-487, 2012.

[47] A. V. Banerjee, "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics*, vol. 107, pp. 797–817, August 1992.

[48] Yafooz, W. M., Abidin, S. Z., Omar, N., & Halim, R. A. (2014). Shared-Table for Textual Data Clustering in Distributed Relational Databases. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013) (pp. 49-57). Springer, Singapore.

[49] Yafooz, W. M., Abidin, S. Z., Omar, N., & Halim, R. A. (2014). Model for automatic textual data clustering in relational databases schema. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013) (pp. 31-40). Springer, Singapore.

[50] Wael M.S. Yafooz, Abidin, S. Z., & Omar, N. (2011, November). Towards automatic column-based data object clustering for multilingual databases. In Control System, Computing and Engineering (ICCSCE), 2011 IEEE International Conference on (pp. 415-420). IEEE.

[51] Yafooz, W. M., Abidin, S. Z., Omar, N., & Halim, R. A. (2013, August). Dynamic semantic textual document clustering using frequent terms and named entity. In System Engineering and Technology (ICSET), 2013 IEEE 3rd International Conference on (pp. 336-340). IEEE.