Data Clustering: Algorithms and Its Applications

Jelili Oyelade[1,2], Itunuoluwa Isewon[1,2], Olufunke Oladipupo[1], Onyeka Emebo[1], Zacchaeus Omogbadegun[1], Olufemi Aromolaran[1,2], Efosa Uwoghiren[1,2], Damilare Olaniyan[1,2] and Obembe Olawole[3]

1. Department of Computer and Information Sciences, Covenant University, Ota, Ogun State, Nigeria.
2. Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Ogun State, Nigeria.
3. Department of Biological Sciences, Covenant University, Ota, Ogun State, Nigeria.

{jelili.oyelade, itunu.isewon, funke.oladipupo, onye.emebo, zacchaeus.omogbadegun, olawole.obembe}@covenantuniversity.edu.ng
{olufemi.aromolaran, efosa.uwoghiren, damilare.olaniyan}@stu.cu.edu.ng

Abstract— **Data is useless if information or knowledge that can be used for further reasoning cannot be inferred from it. Cluster analysis, based on some criteria, shares data into important, practical or both categories (clusters) based on shared common characteristics. In research, clustering and classification have been used to analyze data, in the field of machine learning, bioinformatics, statistics, pattern recognition to mention a few. Different methods of clustering include Partitioning (K-means), Hierarchical (AGNES), Density-based (DBSCAN), Grid-based (STING), Soft clustering (FANNY), Model-based (SOM) and Ensemble clustering. Challenges and problems in clustering arise from large datasets, misinterpretation of results and efficiency/performance of clustering algorithms, which is necessary for choosing clustering algorithms. In this paper, application of data clustering was systematically discussed in view of the characteristics of the different clustering techniques that make them better suited or biased when applied to several types of data, such as uncertain data, multimedia data, graph data, biological data, stream data, text data, time series data, categorical data and big data. The suitability of the available clustering algorithms to different application areas was presented. Also investigated were some existing cluster validity methods used to evaluate the goodness of the clusters produced by the clustering algorithms.**

Keywords- Clustering, cluster analysis, clustering algorithms, model-based clustering, cluster validity

I. Introduction

Data clustering, also refers to as unsupervised classification, is defined as a technique of groups of objects creation, such that objects in one cluster are very identical and objects in different cluster(s) are relatively different. The major goal of clustering is to discover set of patterns, points, or objects from the natural grouping(s) [1].

The rapid advancement in clustering applications such as digital imaging, Internet search, video surveillance, etc. and the advances in storage technology have brought about many high-dimensional and high-volume data sets. According to [1], digital universe exhausted almost 281 exabytes in 2007, and this has geometrically estimated to be 10 times the size by 2011 (That is, 1 exabyte is 1018 bytes or 1,000,000 terabytes).

In addition to the fast growing in the volume of data, the availability of data type such as video, image, text, etc. has also multiplied and many of these data streams are not in structured form (unstructured), which make it difficult in analysing them.

Therefore, there is urgent need the advanced procedure that will automatically understand, process, and summarize this enormous volume and variety of data as required to manage this increased data [1].

There are three purpose Data clustering has been used for:

- Natural classification: This is very important for the identification among organisms, the degree of their similarity or relationship, for instance, the phylogenetic relationship.
- Structurally underlying: This must be underlying structurally so as to understand the perception of the data, hypotheses generation, anomalies detection, and also the identification of the salient features.
- Compression: This is also important as a technique for organizing and summarizing the data through cluster prototypes [1].

In data mining and machine learning, the problem of data clustering technique has been extensively studied in these areas because of its usefulness in the various applications to segmentation, summarization, target marketing, bioinformatics etc [2]. The clustering problems such as cluster validity, robustness of clustering techniques, nature of data and others have been addressed extensively in several applications such as segmentations, text mining, image processing etc. The understanding of the types of data used is very significant as they are the major determinant of the data analysis method to be used.

The major two classification of Data analysis techniques are stated below:

(i) Exploratory Data analysis (EDA) describes the art of an investigator to get maximum understanding into the fundamental structure of the data and it is often the first step in data analysis

(ii) Inferential Data Analysis (IDA) describes the procedure by which an investigator extracts information from data for further reasoning or makes prediction from data.

978-1-7281-2847-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCSA.2019.000-1

A. Types of Data

Information explosion, both structured and unstructured form generate different set of data and also large quantities of data. In unstructured data, information does not follow a specific format or organized in a pre-defined way and also does not have a pre-defined data model. For example, it is usually text-heavy but may also contain video, numbers, audio, images, etc. On the other hand, the structured data typically dwells in relational databases, there are semantic relationships within each object, fields store length-delineated data phone numbers, each text strings of variable length are contained in records which make searching easier unlike unstructured data.

Categorical Data: This represent types of data collected in groups and the number of events in each group is counted numerically. This data consists of finite number of attributes, examples of categorical variables data are race, sex, age group, and educational level. There are some challenges in applying Bayesian/frequentist methods on categorical data [3]

Text Data: This is a human-readable order of characters that is encoded into computer-readable formats such as EBCIDIC, ASCII etc.

Multimedia Data: This type of data consists of various media types such as text, audio, video, and animation. They are time-dependent, and their processing is subject to time constraints. They are discrete representation of fact or figure; however, they appear contiguous to physical observation when presented regularly and periodically at sufficiently high frequencies.

Stream Data: This is an arrangement of digitally encoded packets of data used to communicate or receive information that is in the course of being communicated. Stream data can be equally regarded as a subset of multimedia data, such as video, audio, and animation.

Uncertain Data: This is the type of data that involves noise which allows it deviate from the planned, precise, or original values. Uncertainty or data veracity is one of the important features of data in this era of big data.

Time Series Data: This is a series of data points that is indexed in a particular time order or time periods or intervals. To determine the rate of unemployment monthly for example, would involve a time-series. The unemployment rate is well-defined and regularly assessed at determined periods or equally spaced interval.

Big Data: This is the term used to describe any volumes amount of both structured, semi-structured and unstructured data that has the ability to be mined for useful information. It describes data sets that are not only complex but also voluminous to the extent that the traditional data processing application software which are not capable to deal with them. Information privacy, updating, querying, transfer, visualization, sharing, search, data analysis, data storage and data capturing are major challenges of big data.

II. Methods of Clustering and Their Algorithms

There are several categories of clustering; these categories have different types of algorithms. These categories and types are discussed below:

A. Hierarchical Clustering (Agnes, Diana, Cure, Chameleon)

This clustering category groups data based on the proximity of their data points. The Hierarchical Clustering (HC) permits sub-clusters within a cluster which resulted into a nested cluster organized in tree-like structure. This method has two main approaches, they include; Agglomerative and Divisive approach.

The Hierarchical agglomerative clustering is accomplished by initially, placing each point in a cluster of its own, then find and combine the two points nearest to it, a point in this case refers to an individual object or a cluster of objects.

Since the a priori number of clusters within the data does not require by the HC method, it, however, uses dissimilarity between data objects which is basically the computation of distances to group them. The dissimilarity is computed using distance metrics such as Euclidean distance, Manhattan distance, Unweighted Pair Group Method with Arithmetic Mean (UPGMA) etc [4]. Agglomerative Nesting (AGNES) [5] is an example of HC algorithm that have a rigid clustering process in which once two data points are merged they cannot be unmerged even when it is discovered that the 2 points are dissimilar. This rigidity enables it to have small computation times.

The Divisive clustering is achieved by assuming the whole population is one cluster then starts dividing into smaller groups. It is a top-down approach. Divisive analysis (DIANA) [5] is an example of Divisive clustering and suffers similar drawback as AGNES in terms of its rigidity in clustering process.

AGNES and DIANA are biased against non-spherical clusters due to the centroid-based approach employed by both methods. Clustering Using Representatives (CURE) [6] technique overcomes the biases of AGNES and DIANA to non-spherical clusters by initializing scatter points with a constant number that holds the level and the cluster's shape; the scatter points selected shrink close to the centroid and then becomes the representatives of the cluster.

CURE is less sensitive or stronger to outliers; it can find clusters with non-spherical shapes and size variances; it engages partitioning to handle large databases very efficiently.

```
Procedure merge (a, b)
begin
1. $s := a \cup b$
2. $s$.mean := $\frac{|a|\,a.\text{mean}+|b|\,b.\text{mean}}{|a|+|b|}$
3. tempSet := $\emptyset$
4. for $j := 1$ to $k$ do
```

```
    maxiDist := 0
    foreach point t in cluster s do {
        if j = 1
            miniDist := distance(t, s.mean)
        else
            miniDist := min{distance(t,r) : r E tempSet}
        if (miniDist >= maxiDist){
            maxiDist := miniDist
            maxiPoint := t   }
    }
    tempSet := tempSet ∪ {maxiPoint}
}
foreach point t in tempSet do
    s.rep :=s.rep U {t + α*(s.mean-t) }
return s
end

procedure cluster(M, k)
begin
K := build_kd_tree(M)
H := build_heap(M)
while size(H) > i do {
    a := extract_min(H)
    b := a.closest
    delete(H, b)
    z := merge(a, b)
    delete_rep(K, a); delete_rep(K, a); insert_rep(K, z)
    z.closest := c /* c is an arbitrary cluster in H */
    for each c ϵ H do {
            if dist(z, c) < dist(z. z.closest)
                    z.closest := c
            if c.closest is either a or b {
                    if dist(c, c.closest) < dist(c, z)
                        c.closest := closest_cluster(K, c, dist(c, z))
                    else
                        c.closest := z
                        relocate(H, c)    }
            else if dist(c, c.closest) > dist(c, z) {
            c.closest := z
            relocate(H, c)                    }
    }
    Insert(H, z)
}
end
```

CHAMELEON [7] technique make use of graph-based partitioning algorithm to initially group the data objects into a large amount of relatively small sub-clusters so that objects in each cluster are highly related and consequently less affected by outliers. This technique uses agglomerative clustering algorithm to group objects into clusters by continually merging them using connectivity and closeness measures. If data sets is very large for example, hierarchical methods may not be effective unless other methods are combined, because hierarchical approaches are $O(n^2)$ and $O(n^3)$ for space and time complexities respectively [8,9,10]; where n represents number of data points in the collection of data (dataset). Data types that use this clustering methods are the categorical data [11], time series data [12]

CHAMELEON Algorithm:

Chameleon uses Divide and Conquer approach, where it will first partition the items of data into sub-clusters and then continually merges these sub-clusters to get the final clusters.

Input: Adjacency matrix of data points

Output: a file of clustered data points

1: Construct a k-nearest neighbour graph of the input data items

2: Divide the data points in the graph using multi-level graph dividing technique hMETIS

3: redo step 2

4: Merge the clusters that best conserve the group self-likeness as regards to Relative interconnectivity and Relative proximity

5: Perform step 4 until it is impossible to merge more clusters

B. Partitional Clustering (K-Means, PAM)

Partitional clustering algorithms is a non-hierarchical clustering that usually deals with statics sets. The goal of partitional clustering is to discover the groupings present in the data through optimization techniques of the objective function which improve the quality of the partitions iteratively. In these methods, the desired number of clusters, k would be supplied by the user which are then improved iteratively. However, hierarchical clustering algorithms, is define as the clustering method by creating a binary tree-based data structure which is called dendrogram. Hierarchical methods, on the other hand, is a nested cluster which organizes inform of a tree and does not require a particular value of k unlike non-hierarchical clustering. The hierarchical clustering algorithm develop the clustering in a tree-like form from individual data points in single cluster [2]. Partitional clustering remains one of the most popular and applied techniques because of its simplicity, efficiency, very easy to implement and its empirical success. It does not impose a hierarchical structure and computes all feasible clusters simultaneously [1].

K-Means algorithm finds a partition of n observations into k clusters such that each observation belongs to the cluster with the closest mean. K-means algorithm requires an initial k value which specifies the number of partitions to be obtained and assign objects to groups so as to minimize the squared error. It has a relative low computational cost and is suited for spherical or ball-shaped clusters. There are a number of extensions to K-means to enhance its performance such as Intelligent Kernel K-means (IKKM) etc [13].

***K-Means Algorithm**:*

Let A = {a_1,a_2,a_3,……..,a_n} represents the set of data points and C = {c_1,c_2,…….,c_k} denote centers c_i .

1) Specify 'k' cluster centers randomly.
2) Compute the gap existing between cluster centers and each data point
3) Place the data item into the cluster with least distance to the data point among all cluster centers.
4) Recompute the updated center for all clusters using:

$$V_i = (1/C_i)\sum_{j=1}^{C_i} x_i$$

where, 'k_i' stands for the quantity of data items in i_{th} cluster.

5) Recompute the gap between the updated cluster centers and each data point
6) Check to see if any data point has changed cluster then stop, else redo from step (3).

Partitioning Around Medoid (PAM) The partitional clustering algorithm associated to k-means algorithm and the medoid shift algorithm is refer to as Partitioning Around Medoid (PAM). PAM algorithm clusters objects on a given *m* interval-scaled variables, and also can be applied when input data matrix are a dissimilarity [5]. Both PAM and k-mean algorithms try to minimize the distance between points labelled in a particular cluster and a point that is selected as the center of that particular cluster. But PAM is stronger than k-means because the sum of dissimilarities it minimized as opposed to the sum of squared Euclidean distances in the case of k-mean. It is vulnerable to the issue of initial input, and also failure to compute large datasets, highly connected clusters and high-dimensional datasets makes it less required for clustering a group of data such as gene expression data. Data type that use this clustering methods are the categorical data [11], discrete data [14], text data [14], multimedia data [16], uncertain data [17].

***PAM Algorithm**:*

The algorithm has two phases, build and swap phase. The build phase sequentially selects centrally located k elements while the Swap phase computes the total cost for each pair of selected and non-selected element.

```
Input:
          $S = \{s_1, s_2, ..., s_n\}$ // Set of data points
          M // Adjacency matrix displaying gap between data
          points
          C // Number of preferred clusters
Output
          C // Set of clusters.
Algorithm
          Arbitrarily select $c$ medoids from $S$;
          repeat
                    for each $s_n$ not a medoid do
                              for each medoid $s_m$ do
                                        Calculate $TC_{mn}$;
                    find $m$, $n$, where $TC_{mn}$ is the smallest.
                    if $TC_{mn} < 0$ then
                              replace medoid $s_m$ with $s_n$;
          until $TC_{mn} \geq 0$;
          for each $s_m \in S$ do
                    assign $s_m$ to $C_j$ where $dis(s_m, s_j)$ is the
                    smallest over all medoids;
```

where TC is total cost for each pair of selected and non-selected element that is computed by cost *(x, c)*

$$Cost(x,c) = \sum_{i=1}^{d} x - c$$

where x is any data object, c is the medoid, and d is the dimension of the element.

C. Grid based Clustering (STING, OPTIGRID)

Grid based clustering is renowned for extracting clusters in a huge multidimensional space quantize into a group of cells that form a grid structure on which all of the activities for clustering is carried out. In this approach, clusters are regarded as more condensed than their surroundings.

Statistical information grid-based (STING) algorithm [18] uses statistical information for the approximation of the expected of the query results. STING algorithm has a very low computational cost; also has the capacity of handling large spatial dataset. Graphical representation of the cluster is obtained from the hierarchical structure of the grid cells and the associated statistical information. One major drawback of this technique is that the user is required to provide the density parameter which determines the quality of clustering.

STING Algorithm:

```
1.  Determine a level to begin with.
2.  Develop grid tree-like structure as indicated by the database and the parameters of every cell is generated;
3.  Specify a level in any case;
4.  For every cell in the specify level, the confidence interval of the likelihood is calculated;
5.  if the level not equal leave level then
6.  Go to the succeeding level in the tree-like structure and go to step 3 for the pertinent cells of the upper-level layer;
7.  else if the query specification is met then
8.  Locate the areas of applicable cells and restore those areas that meet the requirements of the query;
    else
9.  Reprocess data contained in the related cells and return the results of the requirements of the query that is met;
    end if
```

Optimal grid (OPTIGRID) [19] is a grid based clustering technique where the partitioning is done such that the dataset is partitioned in a region of low density and the cutting plane should be distinguish clusters as much as possible. The grid cell method safeguards unbiasness for size and shape of cluster as well as handles high dimensional data by ensuring the data points are scattered over various grid cells. Data type that use this clustering methods are the uncertain data [20].

Optigrid Algorithm

```
1.  Input: data set Z, $x$, min_slice_score
2.  Compute the set of contracting projections $C = \{C0, C1, . . . , Ck\}$ and determine every projection of the data set $Z$: $Ci(Z)$, $i = 1, 2, . . . , n$;
3.  Initialize  list of cutting planes $BEST\ \_SLICE \Leftarrow 0$, $SLICE \Leftarrow 0$;
4.  for $j = 0$ to $m$
    (a) $SLICE$ ⇐best local slice ($Ci(P)$);
    (b) $SLICE\ \_SCORE$ ⇐Score best local slice ($Ci(P)$);
    (c) Insert all the cutting planes with a score ≥ min_slice_score into BEST _SLICE ;
5.  if $BEST\ \_SLICE = 0$ then return $P$ as a cluster;
                else
6.  Examine the $x$ cutting planes with the highest score from BEST _SLICE and delete the rest
7.  Construct a multidimensional grid $M$ define by the $x$ cutting planes;
8.  Insert all data points in $Z$ into $G$ (i.e. $x \in G$).
9.  Establish the highly populated grid cells in $G$; add them to the set of clusters $S$;
10. Refine ($S$);
11. for each cluster $S_i \in S$ do
    (a) Perform the same process with data set $S_i$ ;
                end for // inner for loop
                end if
                end for  // outer for loop
```

D. Density based Clustering (DBSCAN, DENCUE)

Density-based clustering algorithm is an algorithm that play an important role in discovery non-linear shapes structure based on the density and which consider clusters as dense regions of objects in the data space and clusters are divided by region of low density. A density value is associated with each object evaluated as the number of its neighbour objects within a given radius. The quality of these techniques is not affected by outliers and shape of cluster.

Similar to the Grid based approach, Density based clustering is also efficient in handling high dimensional data.

DENsity-based CLUstEring (DENCLUE) [21] uses map to calculate the density function and outliers are regarded as cubes with low cardinality and subsequently eliminated from the clustering process and subsequently uses the local density function for the determination of the connectedness of the data points. Map representation of the data enhances compactness of the clusters and also computationally efficient to handle large dataset.

DENCLUE Algorithm:

A Denclue clustering algorithm is defined by the local maxima of the estimated density function. The procedure of a hill-climbing is started for each instance of the data point, and this assigns the instance to a local maximum. The hill climbing is guided by the gradient of $\nabla\hat{C}(u)$ for a Gaussian kernel, which takes the form

$$\nabla\hat{C}(u) = \frac{1}{l^{d+2}}\sum_{j=1}^{N} k\left(\frac{u-u_j}{l}\right)\left(u_j - u\right) \qquad \text{(i)}$$

The procedure for hill climbing begins at a data point and iteratively until the density remain unchanged or does not grow further. The updated formula of the iteration to proceed is given in equation (ii) below:

$$u^{l+1} = u^{(1)} + \delta\frac{\nabla\hat{C}(u^{l})}{\left\|\nabla\hat{C}(u^{l})\right\|_2} \qquad \text{(ii)}$$

Density-based Spatial Clustering of Applications with Noise (DBSCAN) [22] that groups a set of closely parked points in some space pattern as outliers points that lie alone in low-density areas or whose nearest neighbours are too distant away. Major advantages of this method are that a-priori specification of number of clusters is not require, it able to manage outliers while clustering the data set, also, the algorithm is very robust in separating clusters of high density against cluster of low density, and arbitrary size and clusters can be find easily with the method. However, it performance in handling high dimensional dataset is very poor or weak [21]. It is applicable to large datasets. Data type that use this clustering methods are the multimedia data [21], stream data[22], time series data [23].

DBSCAN Algorithm

```
DBSCAN_Method (D, $\varepsilon$, least_points):
    K =0
    for each unvisited element T in D
        indicate T as unvisited
        circle_points = regQuery(T, $\varepsilon$)
        if sizeof(circle_points) < least_points
            omit T
        else
            K = next cluster
            expandClust(T, circle_points, K,
            $\varepsilon$, least_points)

expandClust(T, circle_points, K, $\varepsilon$, least_points):
    include T to cluster K
    for each element T' in circle_points
        if T' is not visited
            indicate T' as visited
            circle_points' = regQuery(T', $\varepsilon$)
            if sizeof(circle_points) >=
            least_points
                circle_points =
                circle_points included
                with circle_points'
            if T' is not yet member of any
            cluster
                add T' to cluster K

regQuery(T, $\varepsilon$):
  return all points within the n-dimensional sphere centered
at T with radius $\varepsilon$ (including T)
```

E. Soft Clustering (Fuzzy C-Means, FLAME)

Soft clustering is a form of clustering approach where data points belong to more than one cluster with a certain degree of membership typically between 0 and 1. The weighted sum for each data point (object) must equal to one.

In practice, when each object is assigns to the cluster of its highest membership weight, then a fuzzy clustering is referred to as hard clustering.

Fuzzy C-means is a soft clustering technique where data points belong to more than one clusters and is distinguished by its fuzzy membership function. In this method, the membership matrix of the input dataset is preserved and this is updated in each iteration. This method has these two major advantages; (1) It has potential of clustering overlapping data points and (2) its rate of convergence is very high. However, Its major drawback is the issue of cluster validity as a result of the a priori requirement of value *c* required for quality of clustering results; and outliers can be allocated indistinguishable membership in each cluster and this may resulted in less desirable for gene expression data [24].

Fuzzy C-means Algorithm

1. Initialize $V = [v_{ij}]$ matrix, $V^{(0)}$
2. At t-step: determine the centers vectors $S^{(t)} = [s_j]$ with $V^{(t)}$

$$s_j = \frac{\sum_{i=1}^{N} v_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} v_{ij}^{m}}$$

3. Update $V^{(k)}$, $V^{(k+1)}$

$$v_{ij} = \frac{1}{\sum_{k=1}^{s} \left(\frac{\|x_i - s_j\|}{\|x_i - s_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\| V^{(k+1)} - V^{(k)} \| < \varepsilon$ then EXIT; else return to step 2

Fuzzy clustering by Local Approximation of MEmbership (FLAME) is a soft clustering technique that describes clusters in the dense parts of a dataset and implement cluster based on the neighbourhood interconnections among the objects [25]. FLAME has the potential for capturing nonlinear interconnections and non-spherical clusters, and also strong in the identification of cluster outliers. Example of data types that use this clustering method are the categorical data [26], time series data [27].

FLAME Algorithm

Fuzzy clustering by Local Approximation of Membership (FLAME) algorithm is into three phases as stated below:

Phase One: Information extraction from the dataset:

for (each object)

a. Find the neighbourhood graph of each object to its K-Nearest Neighbors (KNN) and calculate their proximity;

b. Use the proximity value calculated in (a) above to estimate the object density;

c. Use the object density calculated in (b) above to define the object types as follows:

   i. Cluster Supporting Object (CSO): i.e. object with density > all its neighbours;

ii. Cluster Outliers: i.e. object with density $<$ all its neighbors, and its predefined threshold;

iii. Else (i.e. the rest).

Phase Two: Local approximation of fuzzy memberships:

a. Initialization of initial fuzzy membership:

i. Assigned each (CSO/Outlier) with fixed and complete membership to their individual groups.

ii. Equal memberships are assigned to the rest of all clusters and the outlier group;

b. Then:

i. Update fuzzy memberships of all type 3 objects by a linear combination of the fuzzy memberships of its nearest neighbors.

Phase Three: Clusters are constructed from fuzzy memberships:

a. One-to-one object-cluster assignment;

i. That is, the assignment of each object to the cluster of the highest membership;

b. Assignment of one-to-multiple object-clusters;

i. That is, the assignment of each object to the cluster of the higher membership than a threshold.

F. Model based Clustering (SOM)

Self-organizing map (SOM) [28] is a type or developed based on artificial neural network (ANN) approach that is trained to produces an intuitive map with the use of unsupervised learning (i.e. the discretization of the input space of the training data) of a high-dimensional dataset in two-dimensional (2D) or three-dimensional (3D) space and similar clusters are placed near each other as a result of a single-layered neural network and this applicable to large datasets. Data type that use this clustering methods are the multimedia data [29], time series data [30].

SOM Algorithm

1. Initialization: choose the random value of the weight of node vectors in a map
2. Select input vector *A(t)* randomly from an input space
3. Matching every node in the map by:

a. Use the Euclidean distance formula to calculate the distance (closeness) between input vector and weight of the node vector of the map as follows:

$$d_j(x) = \sum_{i=1}^{D} (x_i - w_{ji})^2$$

b. Monitor the node (the best matching unit node- BMU) that creates the least gap.

4. Apply the weight update equation to the weight vectors of the nodes in the vicinity of the BMU by attracting them nearer to the input vector as follows:

a.
$$W_v(t+1) = W_v(t) + \theta(u, v, t).\alpha(t).(A(t) - W_v(t))$$

5. Increment value of *t* and go back to step 2 while *t* $< \lambda$

Where Wv(t) = weight vector

X(t) = monotonically decreasing learning coefficient

D(t) = the input vector

θ(v, t) = neighbourhood function

λ is the iteration limit

G. Ensemble Clustering

Ensemble clustering [31] involves the application of a combination of several clustering methods on a given dataset into a probably better and more robust consensus clustering. The consensus function is then used to aggregate the results from the various clustering techniques to generate a single clustering result. This method avoids the drawback of priori input of number of clusters by using cluster validation indices for the selection of the optimum cluster numbers for each dataset. Graph-based partitioning is consequently applied to obtain the final result of the clustering which enables inconsistent edges (outliers) to be removed. Data type that use this clustering methods are the Stream data [32].

In conclusion, Table 1 compares the suitability of the available clustering algorithms to different data types.

III. Areas of Application of Data Clustering

Clustering techniques has successfully been applied in various areas or fields of life and example of some of these successful application areas are; Data Mining, Web cluster engines, Academics, Bioinformatics, Machine Learning, Image processing, Weather report analysis etc. According to [2], some common application platforms where clustering problem arises are as follows:

Table 1. Comparative Analysis of clustering algorithms

Data Type	Partitional	Hierarchical	Grid based	Soft Clustering	Density based	Model based	Ensemble
Categorical	✓	✓		✓		✓	
Text	✓					✓	
Multimedia	✓				✓	✓	
Stream					✓		✓
Uncertain	✓		✓				
Time Series	✓	✓	✓			✓	
Discrete data	✓					✓	

• **Method of Collaborative Filtering:** In this method, clustering gives a summary of like-minded users. The assessments provided by the different users are used for the performance of the collaborative filtering and this can be used for providing recommendations in a diverse of applications.

• **Social Network Analysis (SNA):** SNA is the process of qualitative and quantitative social structures through the use of graph theory and networks. In this method, the structure of a social network maps the structure in terms of nodes and the edges or link that connect them. Clustering also play some very important application roles to social network summarization.

• **Customer Segmentation:** This application is the method of grouping customers into different classes with respect to their common attributes. This approach is very similar to collaborative filtering because groups of similar customers are also created by this method in the data. But main difference here is the use of random attributes for clustering purposes about the objects instead of using the rating information.

• **Clustering serve as intermediate step for the other data mining problems:** Apart from data summarization, clustering also serves as an important intermediate step for many problems in data mining research domain; such as classification analysis or outlier analysis. Data summary is very useful or helpful for different categories of application-specific insights.

• **Dynamic Trend Detection:** The dynamic clustering stream algorithms are used for detecting trends in a wide diversity of social networking applications areas and patterns in dynamic stream data. In this algorithm, the data is dynamically grouped in form of a stream to determine the significant patterns of variations. Some streaming data examples are multidimensional data, text streams, trajectory data, and streaming time-series data etc.

• **Multimedia Data Analysis:** Several data of different kinds of modalities are processed at the same time, such as video, images, audio, etc, and these categories of documents fall under multimedia data. The application of multimedia required efficient algorithms for the manipulation of media data, because of the highly stochastic nature of multimedia data, it is very difficult to theorize away the challenges of noise and the media selection.

• **Data Summarization:** This is a key data mining concept that involves approaches for finding compact data representations which make processing and interpretation easier in various of applications. Many clustering techniques are closely correlated to dimensionality reduction techniques, and such techniques can be regarded a form of data summarization.

• **Biological Data Analysis:** Biological data analysis is a scientific way of joining analytical tools with the biological contents for deeper and broader understanding of the relationships known to be connected to experimental observations. The biological data is organized either as networks or as sequences. As a result of the accomplishment of the work of human genome and the growing capability to gathering diverse types of gene expression data, in the last few years, the evolution of biological data analysis has increased exponentially.

A. Applications of Clustering in Data Mining

Raymond and Jiawei [32] built up the CLARANS calculation which is a blend of the Clustering LARge Applications (CLARA) and Partitioning Around Medoids (PAM) to group spatial information. They could approve with genuine information the viability of their calculation against existing calculations for mining spatial information. Hosseini *et al.* [33] utilized clustering to build up a Customer Relationship Management (CRM) strategy to decide client's reliability. They utilized an extended RFM (Recency, Frequency, Monetary) to identify the high-response customers in marketing promotion by including one extra parameter, joining Weighted RFM-based (WRFM-based method) strategy to K-means technique with K-optimum as indicated by Davies-Bouldin Index applied in data mining, and after that arranging client item dedication in under B2B idea. Huang [34] proposed two expansions of k-means technique for grouping huge categorical datasets. The k-modes technique which utilizes a straightforward coordinating disparity measure to manage categorical data replaces the methods for groups with modes, and utilizations a frequency-based strategy to refresh modes in the bunching procedure to limit the grouping cost work. With these expansions, the k-modes calculation empowers the bunching of unmitigated information in a manner like k-means. The k-model's calculation which additionally coordinates the k-means and k-modes calculations to take into consideration bunching objects portrayed by blended numeric and categorical characteristics. Soya bean disease and credit approval datasets were used to demonstrate the performance of the two methods and this showed that the two algorithms are very efficient when clustering large number of datasets, which is very critical and important to data mining applications. Chau *et al.* [35] implemented a variant of k-mean algorithm called UK-means clustering algorithm, this is an enhancement of the k-means method to manipulate data uncertainty. In their work, pattern of moving-object uncertainty was used for the implementation of UK-means algorithm. Their results showed that under uncertainty condition, one can also produce more accurate results with clustering algorithm.

B. Applications of Data Clustering in Search Engines

McCallum, Nigam, & Ungar [36] implemented canopy clustering which is accurate, very easy, fast, and cheap clustering technique. In their work, they group items into overlapping subsets to compare intractable items. In multidimensional feature space, objects are denoted as a point in canopy clustering algorithm. The technique makes use of two distance thresholds (TH1>TH2) and distance metric for clustering processing. It has the property that objects in a true cluster belong to the similar (same) canopy to ensure that no precision is lost by limiting the comparisons of objects in similar canopy. This method is often used as an initial step for some clustering algorithms such as k-means algorithm and well applicable in so many areas such as problem of reference matching from the bibliographic citations domain. Computational running time was decreased by more than an order of magnitude and marginally improved in terms of accuracy. A Strategy was proposed by [37] to show click-through records as a bipartite graph and to apply an iterative agglomerative clustering algorithm to the vertices of the graph to organize all web pages into groups independent of particular user. The Lycos search engine was used and results shows that the algorithm provided better search result for the users. An algorithm was also implemented by Liu *et al.* [38] for parallel distributed hybrid trees in high dimensional spaces for efficient batch searches or online for nearest neighbours of points. They employed two data sets of images for the research, firstly, the clusters with hand labeled for setting several algorithm parameters, and secondly, the larger set which is the target dataset for clustering of images a reality. An estimated scalable version of nearest neighbour search algorithm was developed and used for finding near duplicates among over a billion images.

C. Application of Data Clustering in Academics

El-Halees [39] used clustering as one of the four methods to mine student's data and analyses their e-learning behaviour. The purpose of the clustering was to group active students with non-active students for better student's performance. The Expectation-Maximization Algorithm was used to group students according to their performance. The Mean of each cluster for each attribute was computed and using these results the students were divided into 5 groups. The K-Means algorithm was implemented using the Euclidean distance as a measure of similarity distance by [40] to examine the performance of student's academic progress in higher institutions. The overall performance of this method based their evaluation with a deterministic model where the evaluation of group assessment in each cluster is done by totalling the mean of the individual scores in each cluster. This clustering algorithm served as a good benchmark for students' academic performance monitoring in higher institution which also enhanced the decision making by the academic planners by monitoring the progression of candidates' performance by improving on their future academic performance in the subsequence academic session. [41] also used K-Means algorithm as one of the important processes for academic trends prediction and patterns in educational databases, the algorithm was applied to group the student's profile.

D. Applications in Bioinformatics

CD-HIT is a greedy incremental method of clustering that sorts input sequences that begins with the longest sequence as the representative of first cluster to the shortest sequence, and then, the remaining sequences is processes sequentially from the longest to the shortest. Based on its similarities to the existing representative's sequences, the sequence is automatically classified as the first cluster representative sequence and this was initially built for reference databases creation to cluster protein sequences with redundancy reduction, this was then extended for supporting clustering nucleotide sequences [42]. It has also been used in various applications varying from non-redundant dataset creation [42, 43], protein family classifications [44, 45], artefact identification, metagenomics annotation [46], RNA analysis and in multi-core machines. [47] developed and implemented a Java suite application which is easy to use, platform independent and versatile for a large-scale analysis of gene expression data which referred to as GENESIS. This GENESIS application tool integrates various tools such as normalization, filters and visualization for microarray data analysis, it also incorporated distance measures and also clustering algorithms such as k-means, hierarchical clustering, support vector machines, self-organizing maps and principal component analysis. The results generated by this clustering tool across all the techniques applied are visible which enable the analysis of the results of different algorithms and parameters. The minimum spanning tree, an algorithm based on graph theory was used by [48] to cluster multidimensional gene expression data to study their functional relationship through clustering. This algorithm was implemented on S. cerevisiae data set on human fibroblasts to Serum and Arabidopsis expression data on chitin elicitation and therefore, the efficiency and effectiveness of this algorithm showed as the test results are highly encouraging [49] also implemented and applied association mining rule algorithm to identified the favourable secondary phenotype candidates. This algorithm identified 1967 secondary phenotype hypotheses that cover 244 genes and 136 phenotypes. For the evaluation analysis, one manual and two automated evaluation approaches were used, they were able to demonstrate with their method that the predicted genes constitute a biological relevance and good candidates to be experimentally tested and confirmed of the secondary phenotype candidates [48].

E. *Application in Image Processing*

A fuzzy logic algorithm was developed and implemented by [50] for fuzzy segmentation of Magnetic Resonance Imaging (MRI). The intensity of the MRI homogeneity can be related to weaknesses in the radio – frequency coils or issues related to acquisition sequences. This algorithm is a modification of the objective function of the fuzzy-c means (FCM) algorithm for the compensation of the intensity homogeneity which enable pixel labelling to have effect by its instant neighbours. The experimental result of the synthetic images data and also the MR data demonstrated the effectiveness of the algorithm. Lloyd's k-means clustering method was implemented by [51], this is a filtering algorithm that require kd-tree as its main data structure and also minimizes the mean square distance from each point to the nearest distance. The algorithm executes faster and better as the partition between data clusters grows and also on empirical studies on synthetically generated data and real data sets in data compression application, image segmentation, and colour quantization.

IV. Cluster Validation

Clustering or cluster analysis is an unsupervised learning process of finding structure in data without the assistance of a response variable. They play some important roles in the areas of science and technology, because of the sensitivity of most clustering algorithms to their initial parameters, they have the challenges of the estimation of appropriate number of groups or clusters and because of these challenges, evaluation is required for the assessment of the clustering results in most of the applications and therefore, validity of the clustering is required to perform to avoid finding patterns in a random data and also to assists in comparing two clustering algorithms. Cluster validation is the measure or procedure for the evaluation of the goodness of clustering algorithm results. It can also be defined as a means of checking qualities and reliabilities of clusters resulted from clustering process. Cluster validation can be group into three main categories namely: External, Internal and Relative Criterion Analysis [50]. External criteria measure the validity of a cluster based on external information not contained in the database. Internal criteria measure the validity of a cluster using information contained in the database itself while Relative criterion is used to evaluate the results from two or more different clusters.

V. Conclusion

There are several factors involved when considering appropriate clustering techniques. These factors include types of data, biasness of technique to a priori parameters, shape of cluster, presence of outliers, volume of data, dimension of data, Missing Values etc. Data analysis through clustering has touched all aspects of lives using various algorithms ranging from the distance-based approaches to graphs to machine learning with objectives of finding the most effective and efficient method for optimal clustering. There have also been tremendous improvements in most clustering algorithms in terms of run time, efficiency and processing device leading to a lot of novel variant algorithms to accommodate the large and complex data set being processed today. We have been able to review several data types and clustering algorithms, while also highlighted the comparative analysis of the various clustering algorithms with respect to the data types.

Acknowledgment

The authors would like to thank Covenant University for providing good platform in carrying out this work.

References

[1] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010.

[2] C. C. Aggarwal and C. K. Reddy (2014). Data Clustering: Algorithms and Applications, Taylor & Francis Group, LLC

[3] A. Agresti, “Two Bayesian/frequentist challenges for categorical data analyses,” METRON, vol. 72, no. 2, pp. 125–132, Aug. 2014.

[4] R. R. Sokal and C. D. Michener, “A statistical method for evaluating systematic relationships,” Univ. Kansas Sci. Bull., vol. 28, pp. 1409–1438, 1958.

[5] L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis, vol. 344. New York: John Wiley & Sons, 1990.

[6] S. Guha, R. Rastogi, and K. Shim, “CURE: an efficient clustering algorithm for large databases,” in ACM SIGMOD Record, 1998, vol. 27, no. 2, pp. 73–84.

[7] G. Karypis, E.-H. Han, and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” Computer (Long. Beach. Calif)., vol. 32, no. 8, pp. 68–75, 1999.

[8] J. A. Hartigan, “Printer graphics for clustering,” J. Stat. Comput. Simul., vol. 4, no. 3, pp. 187–213, Jan. 1975.

[9] F. Murtagh, “A Survey of Recent Advances in Hierarchical Clustering Algorithms,” Comput. J., vol. 26, no. 4, pp. 354–359, Nov. 1983.

[10] M. Zaït and H. Messatfa, “A comparative study of clustering methods,” Futur. Gener. Comput. Syst., vol. 13, no. 2–3, pp. 149–159, Nov. 1997.

[11] A. Ahmad and L. Dey, “A k-mean clustering algorithm for mixed numeric and categorical data,” Data Knowl. Eng., vol. 63, no. 2, pp. 503–527, Nov. 2007.

[12] Y. Kakizawa, R. H. Shumway, and M. Taniguchi, “Discrimination and Clustering for Multivariate Time Series,” J. Am. Stat. Assoc., vol. 93, no. 441, pp. 328–340, Mar. 1998.

[13] J. Oyelade et al., “Clustering Algorithms: Their Application to Gene Expression Data.,” Bioinform. Biol. Insights, vol. 10, pp. 237–253, 2016.

[14] S. Guha, R. Rastogi, and K. Shim, “Rock: a robust clustering algorithm for categorical attributes,” Inf. Syst., vol. 25, no. 5, pp. 345–366, Jul. 2000.

[15] I. S. Dhillon and D. S. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Mach. Learn., vol. 42, no. 1/2, pp. 143–175, 2001.

[16] V. Niennattrakul and C. A. Ratanamahatana, "On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping," in 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07), 2007, pp. 733–738.

[17] G. Cormode and A. McGregor, "Approximation algorithms for clustering uncertain data," in Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '08, 2008, p. 191.

[18] W. Wang, J. Yang, and R. Muntz, "STING: A statistical information grid approach to spatial data mining," in VLDB, 1997, vol. 97, pp. 186–195.

[19] A. Hinneburg and D. A. Keim, "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering," in Proceedings of the 25th International Conference on Very Large Data Bases, 1999, pp. 506–517.

[20] C. C. Aggarwal and P. S. Yu, "A Framework for Clustering Uncertain Data Streams," in 2008 IEEE 24th International Conference on Data Engineering, 2008, pp. 150–159.

[21] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in KDD, 1998, vol. 98, pp. 58–65.

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN)," in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996, vol. 96, pp. 226–231.

[23] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03, 2003, p. 216.

[24] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," Comput. Geosci., vol. 10, no. 2–3, pp. 191–203, 1984.

[25] L. Fu and E. Medico, "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data," BMC Bioinformatics, vol. 8, no. 1, p. 1, 2007.

[26] D. W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," Pattern Recognit. Lett., vol. 25, no. 11, pp. 1263–1271, Aug. 2004.

[27] X. Golay, S. Kollias, G. Stoll, D. Meier, A. Valavanis, and P. Boesiger, "A new correlation-based fuzzy logic clustering algorithm for FMRI," Magn. Reson. Med., vol. 40, no. 2, pp. 249–260, Aug. 1998.

[28] T. Kohonen, "The self-organizing map," Proc. IEEE, vol. 78, no. 9, pp. 1464–1480, 1990.

[29] C. Faloutsos, K.-I. Lin, C. Faloutsos, and K.-I. Lin, "FastMap," in Proceedings of the 1995 ACM SIGMOD international conference on Management of data - SIGMOD '95, 1995, vol. 24, no. 2, pp. 163–174.

[30] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," Bioinformatics, vol. 21, no. Suppl 1, pp. i159–i168, Jun. 2005.

[31] X. Hu and I. Yoo, "Cluster ensemble and its applications in gene expression analysis," in Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29, 2004, pp. 297–302.

[32] T. Raymond and H. Jiawei, "Efficient and Effective Clustering Methods for Spatial Data Mining," in Proceeding of VLDB '94, 1994.

[33] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty," Expert Syst. Appl., vol. 37, no. 7, pp. 5259–5264, Jul. 2010.

[34] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Min. Knowl. Discov., vol. 2, no. 3, pp. 283–304, 1998.

[35] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain Data Mining: An Example in Clustering Location Data," Springer, Berlin, Heidelberg, 2006, pp. 199–204.

[36] McCallum, Nigam, K. and Ungar, L.H. (2000) Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, Boston, 20-23 August 2000, 169-178.

[37] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00, 2000, pp. 407–416.

[38] T. Liu, C. Rosenberg, and H. Rowley, "Clustering Billions of Images with Large Scale Nearest Neighbor Search," in 2007 IEEE Workshop on Applications of Computer Vision (WAC V '07), 2007, pp. 28–28.

[39] A. El-Halees, "Mining Students Data To Analyze Learning Behavior : a Case Study Educational Systems," Work, no. February, 2008.

[40] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," Feb. 2010.

[41] S. Parack, Z. Zahid, and F. Merchant, "Application of data mining in educational databases for predicting academic trends and patterns," in 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), 2012, pp. 1–4.

[42] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, "WBCD breast cancer database classification applying artificial metaplasticity neural network," Expert Syst. Appl., vol. 38, no. 8, pp. 9573–9579, 2011.

[43] P. Jaganathan and R. Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification," Comput. Biol. Med., vol. 43, no. 12, pp. 2222–2229, 2013.

[44] A. Mitchell et al., "The InterPro protein families database: the classification resource after 15 years," Nucleic Acids Res., vol. 43, no. D1, pp. D213–D221, 2014.

[45] I. Pedruzzi et al., "HAMAP in 2013, new developments in the protein family classification and annotation system," Nucleic Acids Res., vol. 41, no. D1, pp. D584–D589, 2012.

[46] S. Wu, Z. Zhu, L. Fu, B. Niu, and W. Li, "WebMGA: a customizable web server for fast metagenomic sequence analysis," BMC Genomics, vol. 12, no. 1, p. 444, 2011.

[47] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: cluster analysis of microarray data," Bioinformatics, vol. 18, no. 1, pp. 207–208, Jan. 2002.

[48] Y. Xu, V. Olman, and D. Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," Bioinformatics, vol. 18, no. 4, pp. 536–545, Apr. 2002.

[49] A. Oellrich, I. Jacobsen, J. Papatheodorou, M. G. P. Sanger, and D. Smedley, "Using association rule mining to determine promising secondary phenotyping hypotheses," Bioinformatics, vol. 30, no. 12, pp. i52–i59, 2014.

[50] M. N. Ahmed, S. M. Yamany, N. Mohamed, A. A. Farag, and T. Moriarty, "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data," IEEE Trans. Med. Imaging, vol. 21, no. 3, pp. 193–199, Mar. 2002.

[51] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881–892, Jul. 2002.