Comprehensive review on Clustering Techniques and its application on High Dimensional Data

Afroj Alam[1], Mohd Muqeem[2] and Sultan Ahmad[3*]
[1,2] Department of Computer Application
Integral University, Lucknow(U.P), India
[3*]Department of Computer Science, College of Computer Engineering and Sciences,
Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
Corresponding Author: Sultan Ahmad

Summary

Clustering is a most powerful un-supervised machine learning techniques for division of instances into homogenous group, which is called cluster. This Clustering is mainly used for generating a good quality of cluster through which we can discover hidden patterns and knowledge from the large datasets. It has huge application in different field like in medicine field, healthcare, gene-expression, image processing, agriculture, fraud detection, profitability analysis etc. The goal of this paper is to explore both hierarchical as well as partitioning clustering and understanding their problem with various approaches for their solution. Among different clustering K-means is better than other clustering due to its linear time complexity. Further this paper also focused on data mining that dealing with high-dimensional datasets with their problems and their existing approaches for their relevancy

Key words:
Data mining, Clustering, K-means, PAM, CLARA, ETL, High-dimensional datasets, curse of dimensionality.

1. INTRODUCTION

An abundance of Intelligence is embedded in a different data warehouses in different corporation like financial data warehouse, telecom data warehouse, yield management data warehouses. In these data warehouses they have tremendous interest in the area of knowledge discovery and data mining. Clustering in data mining, is a more powerful, useful and fundamental technique for discovering hidden and interested patterns in the underlying data [1].

In Data mining clustering is also considered as a most important un-supervised machine learning techniques that deals with the large amount of data. Role of clustering is division of instances into homogeneous group of similar data, in which similar behavior between themselves i.e. intra-cluster will be in one group and dis-similar behavior is in other groups [4].

Sometimes grouping of instances is very necessary for different purposes in different fields, such as healthcare, agriculture, image processing, market research, pattern recognition, medical science, text-mining and our daily activity life. For example, in healthcare people having disease and with symptoms of that disease we can put the people in the right group. So that patients will be treated accordingly of that group and patients who not came on that group will be treated differently [2].

Clustering follows the idea of unsupervised learning methodology for finding the basic structure in a group of unlabeled data. Then makes a cluster of similar instances in one group and dis-similar instances in other cluster. This similarity among the instances can be determine by the intrinsic distance value between the instances [1]. The density and shape of the cluster can be constructed by the far and near distances among the instances. This distance is being calculated by squared Euclidean or Manhattan or Hamming distances. Angle vector is used as a distance for making cluster in high-dimensional data [5].

Clustering acts as a business intelligence for business decision maker to predict their customers interest based on their purchasing patterns, so that they can put the customers according to their characteristic in the right cluster. In molecular biological sciences taxonomies of animals and plants can be derived by clustering technique for their gene expression and also find the hidden inherent structure in that population. Clustering technique is also being implemented by geo-logical scientist for identifying the location similarity of lands, similarity of buildings in a specific area [3].

1.1 ETL (Extract transfer and load): It is a process through which Data warehouse is updated periodically as well as immediately. ETL has responsibility to transform the data in standard format that is also called clean data, then upload it in the data warehouse table. This transformation process is taken place before loading the data in data warehouse. In an extraction process, data will be retrieved from external as well as internal data sources [10].

1.2 Feature Selection: It is an important process in clustering, which can be used to determine the unique features in the data sets. That can be later used to define the cluster. If this feature selection process is in-correct that may lead to increase complexity as well as creation of

Manuscript received June 5, 2021
Manuscript revised June 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.6.31

irrelevant cluster. Feature selection procedure may be quantitative as well as qualitative. Qualitative features defined at higher level abstraction in nature whereas quantitative features selection means by ratio, nominal or ordinal scale [10].

1.3 Clustering Algorithm: There are so-many clustering algorithm that solves the real life problems. We have to select the algorithms carefully according to their domain knowledge. Algorithms are based on input the no of clusters, their termination condition, optimization criteria, synchronization of threads etc. Hence we have to analyze all these parameters before choosing the algorithm [10].

1.4 Cluster Validation: According to analysis, same algorithm of clustering on same datasets gives different results when we are choosing different initial points as a centroid for clustering. Also this is un-supervised learning so there is no exact way to judge the goodness of the cluster. Using cluster segregation and compactness we can evaluate the goodness of the clustering [10].

1.5 Results Interpretation: The final step of clustering process is result interpretation. The main goal of the clustering process is to provides the controlling on data with significant insights on general data. This is done for identifying the pattern of the data after that we can efficiently make analysis on that data. That's why clustering is hot and buzz research topic for business intelligence and knowledge interpretation. All these steps of clustering are depicted in the above Figure 1 [10].

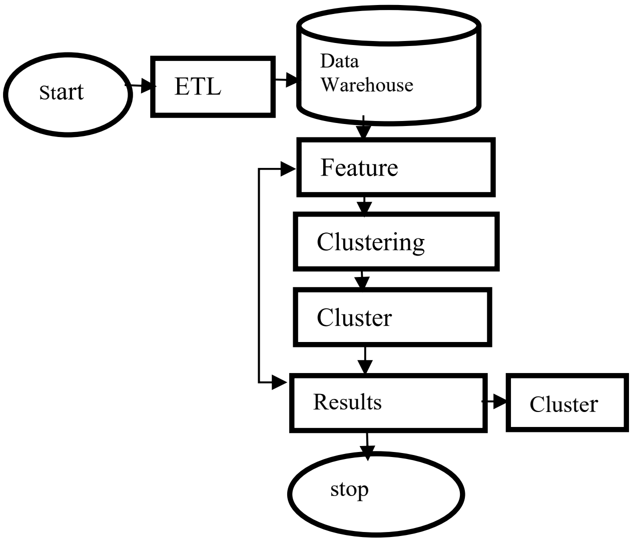

Fig. 1. Clustering process [10]

2. RELATED WORKS

According to recent researchers overlapping k-means algorithm is very attractive algorithm for detecting the overlapping clusters. But also there is one dis-advantage which is selecting the cluster centroids is randomly which gives different results on same datasets on different iteration. This limit has been removed by hybrid k-harmonic means (KHM) and overlapping k-means (OKM) algorithms. The researcher in this article have focused only in healthcare datasets. Applications of KHM-OKM can be implemented also in other domains where we have problem of overlapping cluster. Despite their encouraging results, there are some limitations also in this algorithm that need to be address in future work [22].

With the help of clustering and support vector machine researcher can perform the forecasting of load of electricity for the next 48 hours. Every day average loading of electricity is calculated with the help of periodical data in the data warehouse and the patterns are clustered between the daily average basis uses patterns and daily average basis loading training pattern. The forecasted load using clustering training pattern were compared with actual load without clustering. According to researcher forecasted load was near to actual load. One most important conclusion by researcher in this paper is to verify the importance of clustering for choosing training pattern with support vector machine for getting the better results of short term load forecasting [23].

In this paper Authors give an overview of nature inspired metaheuristic algorithms for partitional clustering. According to the author the old gradient based partitional clustering were easier and was computationally taking less time, but the result was inaccurate due to local minima. On the other hand, using nature inspired metaheuristic algorithms the entire space will be considered as a population for searching and will get the guarantee of optimal partition. Again the multi-objectives algorithm has facility of selecting the desired solution from a group of optimal solution which is not in single-objective algorithms. Automatic clustering having promising solution is generally far better because they don't require apriori information about the number of clusters in the datasets [24].

In this paper the researcher used CBA-ANN-SVM (a recent amalgam clustering model in which both ANN and SVM have been used widely) for the forecasting of load of electricity demand for short-term in Bandarabbas power consumption. The researcher has minimized the rate of error of prediction of load for short-term electrical energy demand in bandarabbas using SVM-ANN hybrid clustering. Using CBA-ANN-SVM model the mean absolute percentage error (MAPE=1.474) when number of cluster is 4 and

MAPE=1.297 when number of cluster is 3. On the other hand, with SVM MAPE=2.015 and with ANN model MAPE=1.7990. That's why CBA-ANN-SVM has been selected as a final model. At the last as a future work researcher saying that either Fuzzy method or genetic algorithm or combined of both will be used for the forecasting of each cluster. Which will give more accuracy in the results [25].

3.TYPES OF CLUSTERING

We have various approaches of clustering techniques because of its different inclusion principle. According to some researcher clustering approach is divided into two categories: one is hierarchical and other is partitioning technique. Some researcher saying that it has been divided into three categories: grid-based, model-based and density based method. Given below shows the clustering approach classification [2][6].

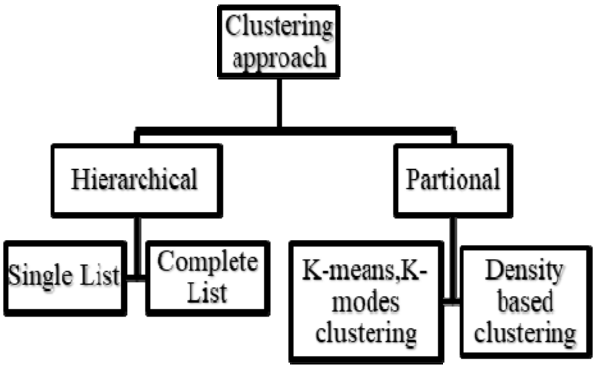

Fig. 2 Taxonomy of clustering approaches [6]

Hierarchical clustering

Hierarchical clustering is based on recursive partitioning where datasets is recursively divided by either top-down approach into finer granularity or bottom-up approach into higher granularity for making the cluster. This type of clustering can be represented by a tree data structure where each non-internal node or child nodes of given cluster represents the data points, which is also called siblings of the parent cluster and internal node represents the cluster. Meaningful information is being captured by this clustering on given set of data points. Ex: at various community level we can create the cluster of communities by using social networks, Digital images can be divided iteratively into distinct region of finer granularity [7]. One basic motivation of using Hierarchical clustering is that, it has large no of partition, all partition is linked with a level. Which is called as a binary tree data structure. Motivation of this clustering over K-means is that there is no need of prior knowledge about the no of cluster [8].

Agglomerative hierarchical clustering:

This type of clustering comes under greedy algorithm, in which no of steps is irreversible for constructing the required data structure. In this approach at each step of the algorithm every pair of cluster is merged or agglomerated. At the start there are n objects with n-1 fine partition and then ending with the trivial partition with one cluster [8].

Given below are some major steps which is being followed by agglomerative clustering.

a. Define the efficient matrix using traditional K-means clustering approach for the initial cluster of given set of points.
b. Searching the minimum distances among the set of points in given matrix.
c. Merge the two cluster which is having minimum distances among the objects of one cluster to objects of another cluster.
d. Update the efficient matrix and iterate the previous three steps until we get one cluster [9].

Given below are some approaches of agglomerative clustering: Single linkage or nearest neighbors(SLINK), Un-weighted pair-group method of average (UPGMA), Un-weighted pair-group method of centroids(UPGMC).

SLINK:

Distance between two cluster A and B with having ai and bj objects is the minimum distances between the all objects pair of one cluster with another cluster.

$D_{A B}$=min(D ai , bj) , where $D_{A B}$ is the distance between two cluster A and B and D ai , bj is the distance between object i and j. The min distance cluster will merge together [27].

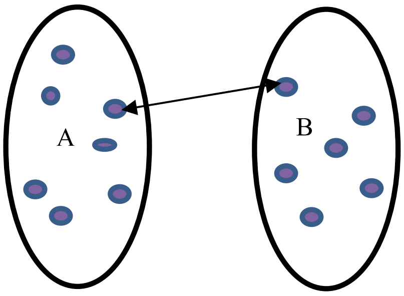

Fig. 3 Single Linkage Clustering

UPGMA:

We have to calculate the average distance between all pairs of objects of two cluster A and B. where A having ai objects and B having bj. After calculating the distances between all pairs, we have to merge the two clusters with minimum distances. This can be repeated till there will be one merged cluster.

$$\mathbf{D}_{A,B}=\frac{\sum Dai,bj}{|A||B|}$$

UPGMC:

In this case we have to calculate first the centroid of the two clusters A and B than calculate the distance between the two centroids.

$$\mathbf{D}_{A,B}= \mathbf{D}\ddot{a}\ddot{b}$$

Where a € A and b € B and a¨, b¨ are the centroids of the two clusters A and B [9].

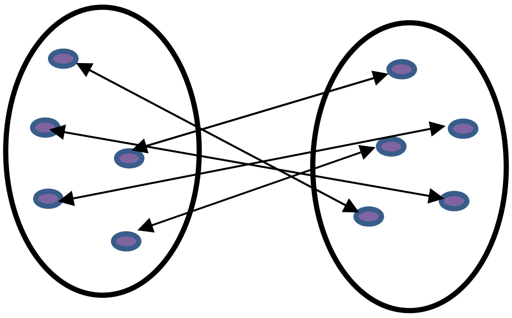

Fig.4.UPGMA

Divisive Hierarchical Clustering

Divisive clustering approach is reverse of agglomerative clustering. It is a top-down approach, which consider the entire sample as a whole cluster. Which then split the whole cluster into two sub-classes at each level and so on. The two new sub-classes at each level so called bi-partition of the former. Hence there are 2n-1 -1 combinations required in the first step for partitioning into two sub-sets. Therefor it is exponential time complexity algorithm whose performance is very poor for large number of items. Hence divisive procedure is not generally use in hierarchical clustering. That's why a new top-down Divisive hieRArchical maximum likelihood clustering procedure(DRAGON) has been introduced. In this procedure the entire sample is not split into all possible sub-classes. Instead it takes out one sample at a time, maximally growing the probability function. This procedure will be going on as usual till first cluster will obtain. This obtained cluster will not further division and will remove from the whole sample. Also n number of times searches is required for removal of n samples. Hence there is O(n2) time complexity is reduced in a top-down procedure.

Pros and Cons of Hierarchical Clustering:

Main pros of this clustering is it is appropriate for large datasets and no need of prior information about number of clusters. It is easy to use because it is based on recursive algorithm. This algorithm has no stochastic elements. It is robust for outlier detection also. The cons of this algorithm is space complexity which depends on the initialization of heaps as well as terminal condition of the algorithm must be satisfied. Also this algorithm extends poorly with respect to memory and computational time when the data-size will increase [11].

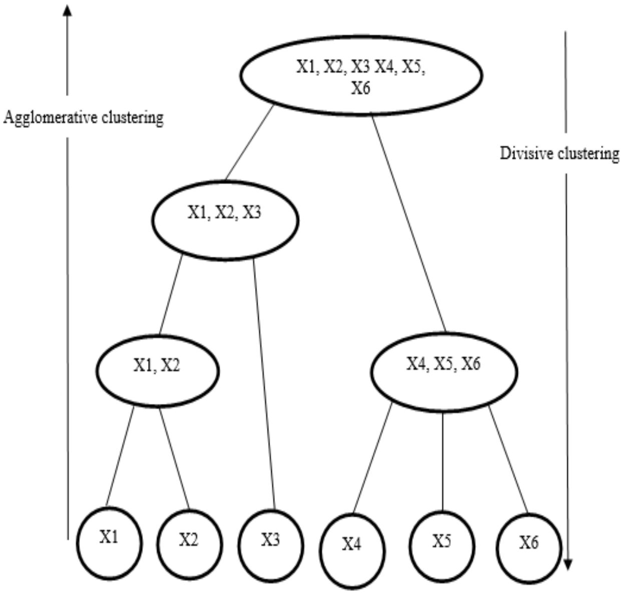

Fig. 5 Hierarchical Clustering [28].

Partitional Clustering:

It is most popular non-hierarchical clustering algorithm which is also known as relocation algorithm. This algorithm iteratively minimizes the cluster criteria by relocating or by updating the centroids until data points are partitioned into optimal clusters. It takes the input datasets as a pattern matrix, then divide it into number of groups and levelling each group on certain criteria known as fitness measure. This fitness measure selected as optimization problem which makes the partition of a datasets D of having N objects in to K optimal cluster (K<=N), so that objects with same nature will be in one cluster whereas dissimilar objects will be in other clusters [12]. Currently these partitional clustering is very hot topic in research fields. Because they have the ability to cluster large datasets. Example: Image segregation clustering for signal and image processing, for clustering the sensor nodes to improve the durability and coverage area in a wireless sensor network, in Artificial intelligence robotics can be easily classified accordingly to the activities of humans, for web categorizing and pattern recognition in the field of computer science, in marketing we can cluster the customers according to their purchasing behavior, portfolio analysis in management science, analysis of high dimensional data, prediction of disease in medical sciences from their gene expression patterns and medical reports etc. In above cases the pattern of data is linked differently in different datasets by different nature.

Partitioning Algorithm [13]:

Below are the steps in partitioning algorithm for clustering the entire dataset D

```
Specify K data points t1, t2, …… tk from Dataset D
Execute loop form i=1 to K step by 1
    Initialize the cluster center Ci=ti
Iterate
        for all the points x in D
        If tj be the prototype that minimize $d_{min}(t,p)$
        Than allocate x in cluster Cj
Quality=clustering (c1, c2, c3, ……………. ck)
Update prototype while quality does not change.
```

If in an application number of cluster is known in advance than its better to use partitional clustering method. Some of the important partitioning clustering algorithms are K-means, partition around median, clustering large application (clara)etc.

K-means clustering

MacQueen has given the idea of K-means partitional clustering, which is one of the most important unsupervised learning partitional clustering. The logic behind this method is that dataset is classified into k centroid disjoint subsets, where K is prior known number of cluster. This algorithm iteratively calculates the distance of all objects from all the K centroid than put the object in the closest centroid cluster, after that all the K centroid will be again re-calculated and update. Like this all the centroid will change their location step by step in each iteration until no more change will done. Finally, the aims of this clustering algorithm is to minimize the objective function of squared error function. Objective function is given below [15].

$$W(S,C) = \sum_{k=1}^{K} \sum_{i \in S_k} \| y_i - C_k \|^2$$

Algorithm [16].

Below are the algorithm steps of K-means

```
    Input: S is a dataset points, No of clusters K
    Result: Cluster
1. First put k points in space that are being clustered
2. Initialize all the K clusters center
3. While termination condition is not specified
    a. Assign each objects to closest centroid
    b. Re-calculate the positions of all the k centroids
4. End while
```

This clustering is very simple but one drawback, it's difficult the determine number of cluster in advance (Elavarasi, Akilandeswari, & Sathiyabhama, 2011). The total time taken by this algorithm for making the cluster is O(nkr) where n is the total data points, k is number of cluster and r is number of iteration.

Partition around median:

Kaufman and Rousseuw has first demonstrated this algorithm, which is based on K-mediods for the all data points of the dataset. This is robust and efficient algorithm to noise and outlier detection. Mediods are the static points with small average dissimilarity of the mean from all other points. In this algorithm entire dataset is randomly partitioned into k subsets. Logic behind this algorithm is that the dataset is partitioned randomly into k subsets and then iteratively improve the cluster mediods so that the objective function can be minimized. In this algorithm number of partitioned is defined randomly (K partition) in which K number of data points for partition are chosen as a mediod. Rest of the non-mediods points are verified iterative in every steps so that mediod will also updated iteratively which will improve the quality of the cluster. This quality can be calculated by adding all the distances in between the mediods and non-mediods data points [14]. The total time taken by this algorithm for making the cluster is O(n(n-k)2) which is a quadratic time complexity which will take more and more time and n is increasing. So its performance is poor than k-means, because k-means has linear time complexity.

Algorithm [17]:

```
1.First of all, we have to choose k objects as an initial mediods.
2.While loop
    i. Put all the remaining objects according to their nearest representative objects
    ii. Choose a non-representative object randomly(Oran)
    iii. Calculate swapping cost C of representative object Oi and Oran
    iv. Swap Oi with Oran if C<0, so that we can design new set of K nearest objects
3. End loop
```

CLARA(Clustering Large Application)

Performance of K-means as well as PAM are not much good even they are not much practical in large dataset due to their prior determination of fixed number of cluster. As the possible number of dataset point increases the rate of number of cluster increases exponentially. This problem we can solved using CLARA. CLARA follow the PAM algorithm for large dataset application for clustering K number of subsets from given datasets. CLARA follow the PAM algorithm for large dataset application for clustering k number of subsets from given datasets. CLARA implements so-many (5) samples, each of them with 40+2k points and all are related to PAM. After this next step is to find all the objects which not belongs to initial sample, that must be equally distributed to the nearest representative object. After that whole datasets will assigned to resulting sample, the resulting sample is compared with n other sample from the entire dataset application. From all these sample the best clustering will be selected by the algorithm [17].

Table 1. Clustering Methods, Complexity & Performance

Name	Outliers or Noise	Time complexity	Performance
K-means	Not reliable for outlier	O(n)	Time increasing linearly as accordingly data points increasing
PAM	More Reliable to Noise than K-means	$O(n^2)$	Time increasing qudratically as accordingly data points increasing
CLARA	Sensitive to outliers	$O(n^2)$	Time increasing qudratically as accordingly data points increasing

PROS AND CONS OF PARTITIONING CLUSTERING

Partitioning clustering algorithm is easy to understand, implement and scalable because it will take less to execute compare to another algorithm. It works good for Euclidian distance data. Drawback of this algorithm is that it gives poor result when the data points is close to the another cluster which lead to the overlapping of data points. User must have to pre-define number of cluster K. Even it is not robust for noisy data and it works only for well-shaped data [18].

4. CLUSTERING ON HIGH-DIMENSIONAL DATA

In Previous section we have discussed lot of clustering techniques, methods and their algorithms all of them are related to 1 to 3 dimensional data and their results also in two dimension. But in many cases like medicine, text-document mining, gene-expression in biological data, DNA microarray data they have from few dozens to thousands dimension in their data. In this scenario clustering formulation is very complex due to its high-dimension and lot of irrelevant features in their data objects. For making cluster if we use feature selection, the biggest challenges will be finding of relevancy features among cluster in high-dimensional data. To overcome this challenges is the reduction in dimension then after we can apply clustering techniques on reduced dimensional dataset. Given below Fig 6. showing the main objectives of clustering in high dimension data [21][30].

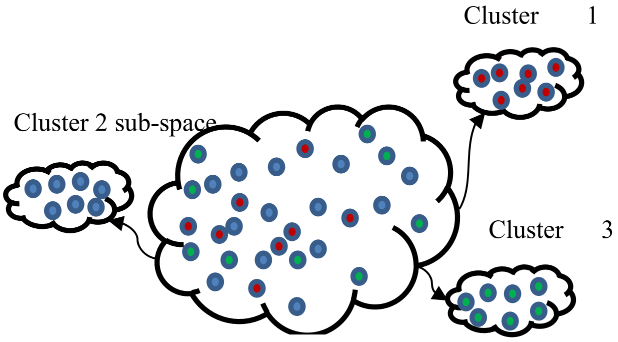

Fig. 6 High dimension data

Problem in Formulation of Cluster in High-dimensional Data

If the dimension of datasets increasing, then the complexity of clustering relationship will increase exponentially which is known as “curse of dimensionality”. In this “curse of dimensionality” there is a decreasing of distances among the data points as there is an increase of space dimension. i.e. there will be no any consequences remains for clustering distance measurement in high dimensional-spaces [21].

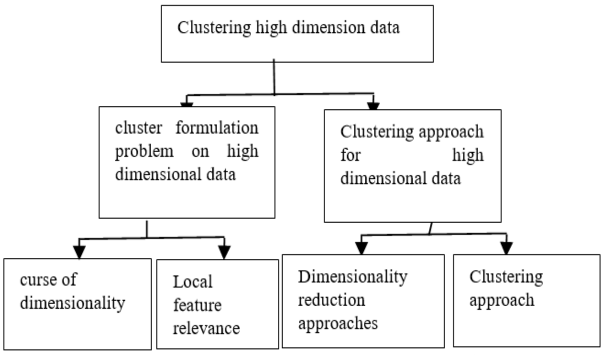
Clustering high dimension data
cluster formulation problem on high dimensional data
Clustering approach for high dimensional data
curse of dimensionality
Local feature relevance
Dimensionality reduction approaches
Clustering approach

Fig. 7. Clustering High-Dimensional Data

5. CLUSTERING APPROACHES FOR HIGH DIMENSIONAL DATA

The Traditional and old algorithm like PAM, CLARA, Agglomerative, Divisive are the basic algorithm for clustering in Data warehouse. A new hybrid clustering approach has been developed which deals with high-dimensional data.

a. Subspace clustering:

In this clustering with the help of integral feature space, they find the cluster in all sub-spaces with the help of axis parallel grid. In this case whole data space is partitioned into equal size unit. Each unit has specified number of points which is called as dense. Each set of dense point is considered as cluster [19].

b. Subspace search methods:

For finding the cluster the entire data spaces will be searched for subspaces recursively. This search method follows both top down as well as bottom up approach. If there is a probability of occurring of cluster in a high dimension than its better to follow up top down approach.

c. Dimensionality Reduction Methods:

Dimensionality reduction is more suitable for constructing a new data space than to adapting the original data sub-spaces. Example – If we project any sub-spaces as a clustering in x-y plane than all the 3 dimension will not be projected in x-y plane, clusters will overlap which is show in given below figure. If we construct another dimension as a dashed than all the three points will be visible. This dimensionality reduction is achieved by mathematical transformation, which shown in given below figure x. Some of the methods explained below [21].

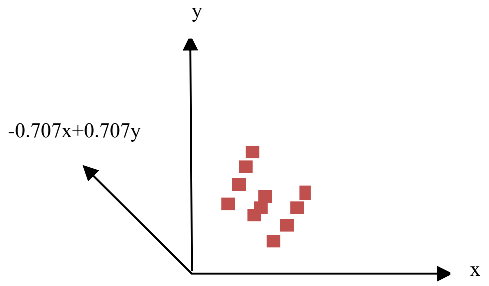

Fig. 8. Dimensionality Reduction

d. Non-negative Matrix Factorization (NMF)

The high-dimensional matrix is being split by NMF into its low rank matrix with the help of linear algorithm and multivariate algorithm analysis. Example: "V" is factorized in to small matrices "W" and "H" with having no negative elements. This method is being applied in document clustering, audio signal processor, recommended system, computer vision.

e. Adaptive Dimension Reduction (ADR) [20]

In K-means clustering we can implement ADR method for dimension reduction in a new data sub-space. Suppose P=[p1, p2, ………… pk] are the set of data points, For K number of clusters K-means will create K centroids C=[c1, c2, c3, …………… cn] for minimize the distance.

$$S_d(P,C)=\sum_{k=1}^{K}\sum_{i\in c_k}\|p_i - c_k\|^2_d$$

Theorem: Let's assume if we know somehow the correct r-dimension relevant sup-spaces which is explained by Rr let Y=R_r^t=R_r^(t)=(p1, p2, p3 …………..,pn) and C=[c1, c2, c3, ……………… ck] be K centroids in r-dim sub-space. Than K-means in r-dim sub-space,

$$\min_{c} J_r(Y, C)$$

6. CONCLUSION AND FUTURE SCOPE

Clustering is an unsupervised machine learning data mining techniques, which grouping the data into different groups according to their features and classes. We have seen that there are different types of clustering among which we observed that K-means clustering is more common in health-care sector for disease prediction especially in a high-dimensional data.

In the field of yield management, fraud detection, crime detection there is huge scope for example investigation for frequent sub-structure pattern on large data using clustering technique.

In the field of international super-market prediction of frequent item-sets selling, big Spenders customers, characteristics of software products that either increase or decrease according to their demands and need. To measure, monitor, characterize and discriminate these predictions we need suitable data-mining techniques like k-means or improved k-means clustering which still not solved these problem completely. We need some improvement and exploration on these techniques.

ACKNOWLEDGMENTS

The authors would like to thank the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for the assistance.

References

[1] Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. ACM Sigmod record, 27(2), 73-84.

[2] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664-681.

[3] Bansal, A., Sharma, M., & Goel, S. (2017). Improved K-mean clustering algorithm for prediction analysis using classification technique in data mining. International Journal of Computer Applications, 157(6), 0975-8887.

[4] Pavithra, M., & Parvathi, R. M. S. (2017). A survey on clustering high dimensional data techniques. International Journal of Applied Engineering Research, 12(11), 2893-2899.

[5] Han, J.,Pie, J., & Kamber, M. (2010). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2010.

[6] Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal, 41(8), 578-588.

[7] Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C. (2019). Hierarchical clustering: Objective functions and algorithms. Journal of the ACM (JACM), 66(4), 1-42.

[8] Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1219.

[9] Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. Expert Systems with Applications, 42(5), 2785-2797.

[10] Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2), 1-68.

[11] Kameshwaran, K., & Malarvizhi, K. (2014). Survey on clustering techniques in data mining. International Journal of Computer Science and Information Technologies, 5(2), 2272-2276.

[12] Popat, S. K., & Emmanuel, M. (2014). Review and comparative study of clustering techniques. International journal of computer science and information technologies, 5(1), 805-812.

[13] Shakeel, P. M., Baskar, S., Dhulipala, V. S., & Jaber, M. M. (2018). Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. Health information science and systems, 6(1), 1-7.

[14] Mohammed, N. N., & Abdulazeez, A. M. (2017, June). Evaluation of partitioning around medoids algorithm with various distances on microarray data. In 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 1011-1016). IEEE.

[15] Elavarasi, S. A., Akilandeswari, J., & Sathiyabhama, B. (2011). A survey on partition clustering algorithms. International Journal of Enterprise Computing and Business Systems, 1(1).

[16] Makwana, T. M., & Prashant, R. (2013). Partitioning Clustering algorithms for handling numerical and categorical data: a review. arXiv preprint arXiv:1311.7219.

[17] Shah, M., & Nair, S. (2015). A survey of data mining clustering algorithms. International Journal of Computer Applications, 128(1), 1-5.

[18] Zafar, M. H., & Ilyas, M. (2015). A clustering based study of classification algorithms. International journal of database theory and application, 8(1), 11-22.

[19] Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery, 11(1), 5-33.

[20] Ding, C., He, X., Zha, H., & Simon, H. D. (2002, December). Adaptive dimension reduction for clustering high dimensional data. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 147-154). IEEE.

[21] Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-dimensional and large datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2), 1-68

[22] Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. Expert Systems with Applications, 67, 12-18.

[23] Fu, X., Zeng, X. J., Feng, P., & Cai, X. (2018). Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China. Energy, 165, 76-89.

[24] Nanda, S. J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation, 16, 1-18.

[25] Torabi, M., Hashemi, S., Saybani, M. R., Shamshirband, S., & Mosavi, A. (2019). A Hybrid clustering and classification technique for forecasting short-term energy consumption. Environmental progress & sustainable energy, 38(1), 66-76.

[26] Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal, 41(8), 578-588.

[27] Sneath, P. H., & Sokal, R. R. (1973). Numerical taxonomy. The principles and practice of numerical classification.

[28] Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. The computer journal, 26(4), 354-359.

[29] Assent, I. (2012). Clustering high dimensional data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(4), 340-350.

[30] A. E. M. Eljialy, Sultan Ahmad,"Errors Detection Mechanism in Big Data",IEEE, Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019) on 27-29 November, 2019