Applications of Clustering Techniques in Data Mining: A Comparative Study

Muhammad Faizan[1], Megat F. Zuhairi[2*], Shahrinaz Ismail[3], Sara Sultan[4]
Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia[1, 2, 3]
College of Computing and Information Sciences, Karachi Institute of Economics and Technology, Karachi, Pakistan[4]

Abstract—**In modern scientific research, data analyses are often used as a popular tool across computer science, communication science, and biological science. Clustering plays a significant role in the reference composition of data analysis. Clustering, recognized as an essential issue of unsupervised learning, deals with the segmentation of the data structure in an unknown region and is the basis for further understanding. Among many clustering algorithms, "more than 100 clustering algorithms known" because of its simplicity and rapid convergence, the K-means clustering algorithm is commonly used. This paper explains the different applications, literature, challenges, methodologies, considerations of clustering methods, and related key objectives to implement clustering with big data. Also, presents one of the most common clustering technique for identification of data patterns by performing an analysis of sample data.**

Keywords—Clustering; data analysis; data mining; unsupervised learning; k-mean; algorithms

I. Introduction

Data mining is the latest interdisciplinary field of computational science. Data mining is the process of discovering attractive information from large amounts of data stored either in data warehouses, databases, or other information repositories. It is a process of automatically discovering data pattern from the massive database [1], [2]. Data mining refers to the extraction or "mining" of valuable information from large data volumes [3], [4]. Nowadays, people come across a massive amount of information and store or represent it as datasets[4], [5]. Process discovery is the learning task that works to the construction of process models from event logs of information systems [6]. Fascinating insights, observable behaviours, or high-level information can be extracted from the database by performing data mining and viewed or browsed from various angles. The knowledge discovered can be applied for process control, decision making, information management, and question handling. Decision-makers will make a clear decision using these methods to improve the real problems of this world further. In data mining, many data clustering techniques are used to trace a particular data pattern [2]. Data mining methods for better understanding are shown in Fig. 1.

Clustering techniques are useful meta-learning tools for analyzing the knowledge produced by modern applications. Clustering algorithms are used extensively not only for organizing and categorizing data but also for data modelling and data compression [7]. The purpose of the clustering is to classify the data into groups according to data similarities, traits, characteristics, and behaviours [8]. Data cluster evaluation is an essential activity for finding knowledge and for data mining. The process of clustering is achieved by unsupervised, semi-supervised, or supervised manner [2]. However, there are more than 100 clustering algorithms known and selection from these algorithms for better results is more challenging.

PyClustering is an open-source library for data mining written in Python and C++, providing a wide variety of clustering methods and algorithms, including bio-inspired oscillatory networks. PyClustering focuses primarily on cluster analysis to make it more user friendly and understandable. Many methods and algorithms are in the C++ namespace "ccore::clst" and in the Python module "pyclustering.cluster." Some of the algorithms and their availability in PyClustering module is mentioned in Table I [9].

A. Clustering in Data Mining

Data volumes continue to expand exponentially in various scientific and industrial sectors, and automated categorization techniques have become standard tools for data set exploration [10]. Automatic categorization techniques, traditionally called clustering, helps to reveal a dataset's structure [9]. Clustering is a well-established unsupervised data mining-based method [11], and it deals with the discovery of a structure in unlabeled data collection. The overall process that will be followed when developing an unsupervised learning solution can be summarized in the following chart in Fig. 2:

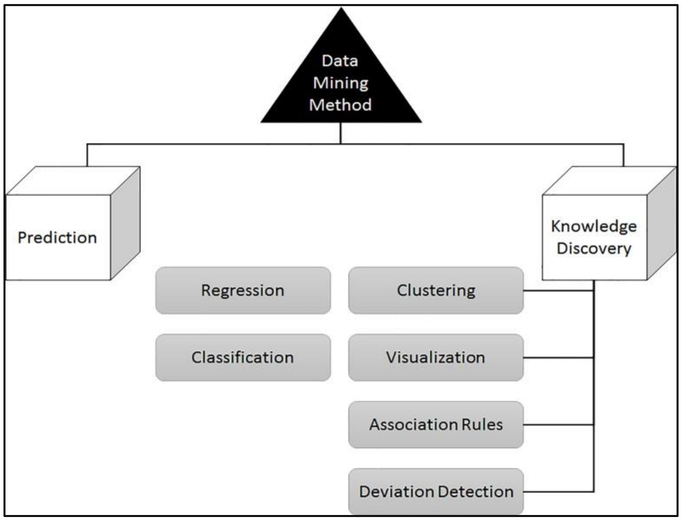


Fig. 1. Methods of Data Mining Techniques.

*Corresponding Author

TABLE I. ALGORITHMS AND METHODS IN "PYTHON MODULE PYCLUSTERING"

Algorithm	**Python**	**C++**
Agglomerative (Jain & Dubes, 1988)	✓	✓
BIRCH (Zhang, Ramakrishnan, & Livny, 1996)	✓	
CLARANS (Ng & Han, 2002)	✓	
TTSAS (Theodoridis & Koutroumbas, 2009)	✓	✓
CURE (Guha, Rastogi, & Shim, 1998)	✓	✓
K-Means (Macqueen, 1967)	✓	✓
BANG (Schikuta & Erhart, 1998)	✓	
ROCK (Guha, Rastogi, & Shim, 1999)	✓	✓
K-Medians (Jain & Dubes, 1988)	✓	✓
Elbow (Thorndike, 1953)	✓	✓
GA - Genetic Algorithm (Harvey, Cowgill, & Watson, 1999)	✓	✓
DBSCAN (Ester, Kriegel, Sander, & Xu, 1996)	✓	✓
X-Means (Pelleg & Moore, 2000)	✓	✓
K-Means++ (Arthur & Vassilvitskii, 2007)	✓	✓
Elbow (Thorndike, 1953)	✓	✓
BSAS (Theodoridis & Koutroumbas, 2009)	✓	✓
K-Medoids (Jain & Dubes, 1988)	✓	✓
Sync-SOM	✓	
SyncNet	✓	✓
SOM-SC (Kohonen, 1990)	✓	✓
OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999)	✓	✓
CLIQUE (Agrawal, Gunopulos, & Raghavan, 2005)	✓	✓
Silhouette (Rousseeuw, 1987)	✓	
MBSAS (Theodoridis & Koutroumbas, 2009)	✓	✓
HSyncNet (Shao, He, Böhm, Yang, & Plant, 2013)	✓	✓
EMA (Gupta & Chen, 2011)	✓	

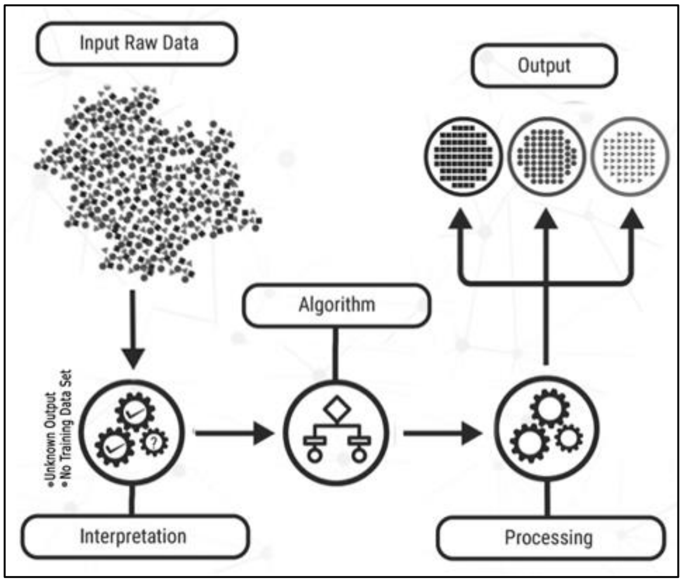


Fig. 2. Unsupervised Learning Model.

The main applications of unsupervised learning are:

- Simplify datasets by aggregating variables with similar attributes.
- Detecting anomalies that do not fit any group.
- Segmenting datasets by some shared attributes.

Clustering results in the reduction of the dimensionality of the data set. The objective of such a clustering algorithm is to identify the distinct groups within the data set [12]. There are different clustering objects, such as hierarchical, partitional, grid, density-based, and model-based [13]. The performance of various methods can differ depending on the type of data used for clustering and the volume of data available [14]. For example, Document clustering has been investigated for use in many different areas of text mining and information retrieval [15]. There are several different metrics of quality, relative ranking, and the performance of different clustering algorithms that can vary considerably depending on which measure is used. Two measures of "goodness" or quality of the cluster are used for clustering. One type of measure allows comparing different cluster sets without external knowledge and is called an "internal quality measure." The other form of measure is called an "external quality measure," which allows evaluating how well the clustering works by comparing the groups generated by the clustering techniques to the classes identified. Fig. 3 shows a simple example of data clustering based on data similarity.

1) Types of clustering: Clustering can generally be broken down into two subgroups:

- Hard Clustering: In hard clustering, each data point is either entirely or not part of a cluster.
    - For example, each customer is grouped into one of 10 groups.
- Soft Clustering: In soft clustering, a probability or likelihood of the data point being in certain clusters is assigned instead of placing each data point into a separate cluster.
    - For example, each customer is assigned a probability to be in 10 clusters.

2) Clustering methodologies: Since the clustering method is subjective, it is the tool that can be used to accomplish plenty of objectives. Every methodology follows several sets of rules and regulations that describe the 'similarity' between data points. Cluster analysis is not an automated task, but an iterative information discovery process or multi-objective collaborative optimization involving trial and error [16]. There are typically more than 100 known clustering algorithms. But few of these algorithms are popularly used. Some of the clustering methodologies are mentioned below in Table II.

The best known and most widely used method of partitioning is K-means [17]–[19]. There are many clustering techniques from which K-means is an unsupervised and iterative data mining approach [11]. The standard approach of all clustering techniques is to classify cluster centres representing each cluster. K-means clustering is a method of cluster analysis aimed at observing and partitioning data point into k clusters in which each observation is part of the nearest mean cluster [7]. The most significant advantage of the K-means algorithm in data mining applications is its efficiency in clustering large data sets. K-means and its different variants have a computation time complexity that is linear in the number of records but is assumed to discover inferior clusters [15].

The K-means algorithm is a basic algorithm for iterative clustering. It calculates the distance means, giving the initial centroid, with each class represented by the centroid, using the distance as the metric and given the classes K in the data set. In the k-means partitioning algorithm, the mean value of objects within-cluster is represented at the centre of each cluster.

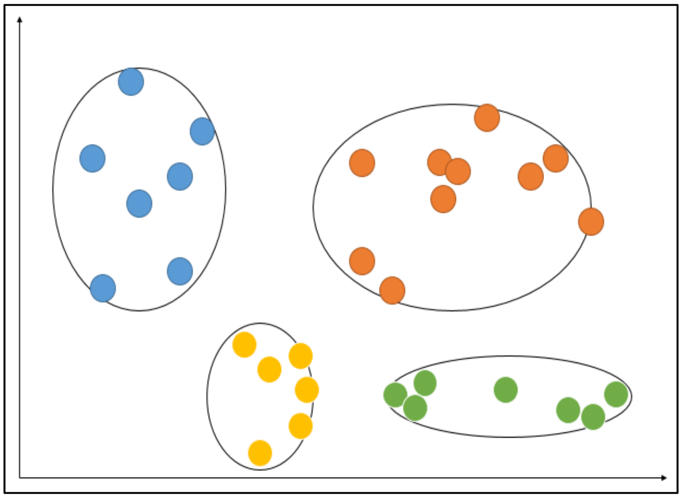


Fig. 3. Simple Clustering Example.

TABLE II. CLUSTERING METHODOLOGIES

Typical Clustering Methodologies	
Method	***Algorithm***
Distance-based method	• Partitioning algorithms "K-means, K-medians, K-medoids." • Hierarchical algorithms, "Agglomerative, Divisive method." These algorithms run iteratively to find the local optima and are incredibly easy to understand but have no scalability for handling large datasets.
Grid-based method	• Grid-base algorithm: Individual regions of the data space are formed into a grid-like structure. These methods use a single-uniform grid mesh to separate the entire problem domain into cells. The cell represents the data objects located within a cell using a collection of statistical attributes from the objects.
Density-based method	• Density-Based Spatial Clustering of Applications with Noise / DBSCAN • Ordering points to identify the clustering structure OPTICS These algorithms scan the data space for areas with different data points density within the data space. It isolates different density regions within the same cluster and assigns the data points within those regions.
Probabilistic and generative models	• Expectation-maximization algorithm: Modeling data from a generative process. Often these models suffer from over-fitting. A prominent example of such models is the Expectation-Maximization algorithm that uses normal multivariate distributions.

II. Background and Discussion of Clustering Applications and Approaches

Cluster analyses have lots of applications in different domains, e.g., It has been popularly used as a preprocessing step or intermediate step for other data mining tasks "Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, compression, reduction, and outlier detection, etc." Clustering analysis can also be used in collaborative filtering, recommendation systems, customer segmentation, multimedia data analyses, biological data analyses, social network analysis, and dynamic trend detection. Some of the clustering techniques and approaches are discussed in Table III.

A. Requirement and Challenges

Despite recent efforts, the challenge of clustering on "mixed and categorical" data in the sense of big data remains, due to the lack of inherently meaningful similarity measurement between the high computational complexity of current clustering techniques and categorical objects [18]. For cluster analysis, there are several items to be considered. Some of them are mentioned in Table IV.

Typically, there are multiple ways to use or apply clustering analysis; some advantages and limitations of clustering techniques are mentioned in Table V.

- As a stand-alone tool to get insights into data distribution.
- As a preprocessing (or intermediate) steps for other algorithms.

According to [24], [25], parallel classification is a better approach for big data, but due to its implementation's complexity remains a significant challenge. However, the framework of MapReduce can be suitable for implementing parallel algorithms, but still, there is no algorithm to handle all Challenges of big data. In [26], the authors proposed a novel Spark extreme learning machine "SELM" algorithm based on a spark parallel framework to boost the speed and enhance the efficiency of the whole process. SELM gives the highest speed and minimal error in all experimental results compared to Parallel Extreme Learning Machine (PELM) and an improved Extreme Learning Machine (ELM*). Table VI presents the pros and cons of different clustering algorithms with real-world applications.

TABLE III. Clustering Techniques and Approaches with Benefits

Ref.	**Author**	**Year**	**Technique / Algorithm**	**Approach**	**Outcomes**
[19]	Chunhui Yuan and Haitao Yang	2019	K-Means Clustering Algorithm	Different methods applied to each dataset to determine the optimal selection of K-Value.	Concluded that these four methods (Elbow methods, silhouette coefficient, gap statistics, and canopy) satisfy the criteria for clustering small data sets. In contrast, the canopy algorithm is also the best choice for large and complex data sets.
[20]	Tengfei Zhang, Fumin Ma	2015	Rough k-means clustering	Improved rough k-means clustering with Gaussian function based on a weighted distance measure	An improved rough k means algorithm based on weighted distance measure with Gaussian function handles the objects which are wrongly assigned to clutters, also handles vulnerable sets while distributing *overlapping* objects in different clusters by rough k means with the same weighted distance for both upper and lower bounds.
[21]	Lior Rokach, Oded Maimon	2015	Clustering methods: Hierarchical- based, Model-based, Grid-based, Partitioning based, Density-based.	Different clustering methods/techniques are used to determine clustering efficiency in large data sets and explain how the number of clusters can be calculated.	For large dataset concluded that "K-means clustering is more efficient in terms of its time, space complexity, and its order-independent" and "Hierarchical clustering is more versatile, but it has the following disadvantages: Time complexity O($m^2 * logm$) and space complexity of a hierarchical agglomerative algorithm is O(m^2).
[22]	Zengyou He, Xiaofei Xu, Shengchun Deng, Bin Dong	2015	K-mean, K-modes, K-Histogram	Compare different clustering algorithms to determine an efficient clustering algorithm for the categorical dataset.	K-Histogram is the enhanced version of K-means to categorical areas by substituting means of clusters with histograms. In general, K-Histogram is almost similar to the K-modes algorithm, but as compared to k-modes, k-histogram algorithms are more stable, and the algorithm will converge faster.
[16]	M.Venkat Reddy, M. Vivekananda, RUVN Satish.	2017	Divisive, and Agglomerative Hierarchical Clustering with K-means.	Discover an efficient clustering by comparing Divisive and Agglomerative Hierarchical Clustering with K-means.	To obtain high accuracy, Agglomerative Clustering with k-means will be the practical choice. Divisive clustering with K-means also works efficiently where each cluster can be taken fixedly.
[23]	Ahamed Al Malki, Mohamed M. Rizk, M.A. El-Shorbagy, A. A. Mousa	2016	K-means, Genetic algorithm	For solving the clustering problems, introduced a hybrid approach of the Genetic algorithm with K-means.	A hybrid approach of K-means with a Genetic algorithm efficiently solves all the problems of the k-means, e.g., K-mean will produce empty clusters with initial centre vector and converge to non-optimal value, etc.

[7]	Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta.	2012	Hierarchical, K-Means, DB Scan, OPTICS, Density-Based Clustering, EM Algorithm	A comparison was made between different clustering techniques to measure the best performing algorithm.	K-means is faster than all the algorithms that are discussed in this paper. When using a huge dataset, K-means and EM will the best results than hierarchical clustering.
[11]	Karthikeyan B., Dipu Jo George, G. Manikandan, Tony Thomas.	2020	K-means, Agglomerative Hierarchical Clustering	Comparative research to determine the best-suited algorithm on K-Means and Agglomerative Hierarchical Clustering.	The k-means is best suited for larger datasets in term of minimum execution time and rate of change in usage of memory. It is also concluded that agglomerative clustering is best suited for smaller datasets due to the overall minimum consumption of memory.

TABLE IV. CONSIDERATIONS OF CLUSTERING ANALYSIS

Considerations for Clustering Analysis		
Considerations	***Options***	***Examples***
Similarity measure	Distances-based / Connectivity- based	Euclidean, road network, vector / Density, Contiguity
Partitioning criteria	Single level / Hierarchical partitioning	Often / Multi-level
Cluster space	Full space / Subspace	Low-dimensional / High-dimensional
Separation of clusters	Exclusive / Non-exclusive	Datapoint belongs to only region / Data point belongs to multiple regions

TABLE V. ADVANTAGES AND LIMITATIONS OF CLUSTERING TECHNIQUES

Clustering techniques "Advantages & Limitations"		
Clustering Techniques	***Advantages***	***Limitations***
Data-mining clustering algorithms	• Implementation is simple. • Compromises on user's privacy.	• Do not deal with a large amount of data
Dimension reduction	• It is very fast, reduces the dataset, and the cost of the treatment will be optimized.	• It must be applied before the classification algorithm. • It cannot provide an efficient result for the high dimensional dataset. • It may lose some amount of data.
Parallel classification	• It gives minimal execution time and more scalable.	• Difficult to implement.
MapReduce framework	• Flexibility, scalability, security and authentication, batch processing, etc.	• It does not do best for graphs, iterative, and incremental, multiple inputs, etc.

TABLE VI. CLUSTERING ALGORITHMS PROS AND CONS

Algorithm Name	**Pros**	**Cons**	**Applications in Real World**
K-means	➢ Handles large amounts of data. ➢ Minimum execution time. ➢ Simple to implement, etc.	➢ Manually choose the K value. ➢ Clustering outliers. ➢ Dependent on starting point/value. ➢ Handle empty clusters, etc.	➢ Wireless networks. ➢ System diagnostic. ➢ Search Engine. ➢ Document Analysis. ➢ Fraud detection. ➢ Call record Analysis.
Hierarchical Clustering	➢ Do not need to specify the initial value. ➢ Easy to implement, scalable and easy to understand, etc.	➢ Cannot handle a large amount of data with different sizes. ➢ No backtracking. ➢ No swapping between objects. ➢ More space and time complexity.	➢ Humans skin analysis [27] ➢ Generating a portal site. ➢ Web usage mining.
Genetic Algorithm	➢ Easily understandable and converge with different problems. ➢ It cannot always give the best result for all problems but provide the optimum solution. ➢ It cannot search for a single point, search from a population of a point.	➢ It is computationally expensive, e.g., time-consuming. ➢ It may lose data in a crossover.	➢ Engineering Designs. ➢ Robotics. ➢ Telecommunications, traffic, shipments routing. ➢ Virtual Gaming. ➢ Marketing.
DBSCAN	➢ The number of clusters does not need to be defined. ➢ Handle outliers.	➢ Unable to handle datasets with distinct densities. ➢ Struggles to work with High Dimensionality Data.	➢ Satellite pictures, etc.

III. Running Example with K-mean

A car manufacturer company wants to identify the purchase behaviours of its customers to view which product is getting more sales and what is the procedure of our customers. They are currently looking at each customer's details based on this information, decide which product manufacturing should be increased and what are the behaviour of customers which helps the company to monitor sales for other products by starting a promotional campaign or increase the availability of resources.

Recently, the company can potentially have millions of customers. It is not possible to look at each customer's data individually and then make a decision. A manual process will take a huge amount of time. This is when K-means Clustering assists in a convenient way to analyze data automatically. The K-mean clustering algorithm utilizes a fixed number of clusters for optimum clustering [12], [28]. Initially, start partitioning with the chosen number of clusters next to improve the partitions iteratively to find the patterns in data. Let D= {D1, D2, …, Dn} be the set of data points and Y= {Y1, Y2, …, Yt} be the set of centers. This clustering technique is implemented and analyzed using a k-mean clustering tool WEKA. In the following steps, the K-means algorithm can be implemented:

K-mean Clustering Algorithm

1. The first step in k-means is to pick the number of clusters, k.
2. Randomly select k number of clusters centre.
3. Find the distance between each data point and each cluster centre.
4. In contrast with other cluster centres, assign the data point to the nearest cluster centre.
5. Discover the new Cluster Centre again $Y_i = \sum_{i=0}^{ki} D_i$ by where k_i represents number of data points in i^{th} cluster.
6. Once again, find out the distance between each data point and the new cluster centre.
7. If no data points were reassigned then stop, otherwise back to step 3.

The data set used for the K-mean clustering example will focus on a fictional car dealership. The dealership is starting a promotional campaign for slow-selling units, whereby it is trying to push resources to its valuable customers. Table VII shows the sample dataset, which is used for the analysis.

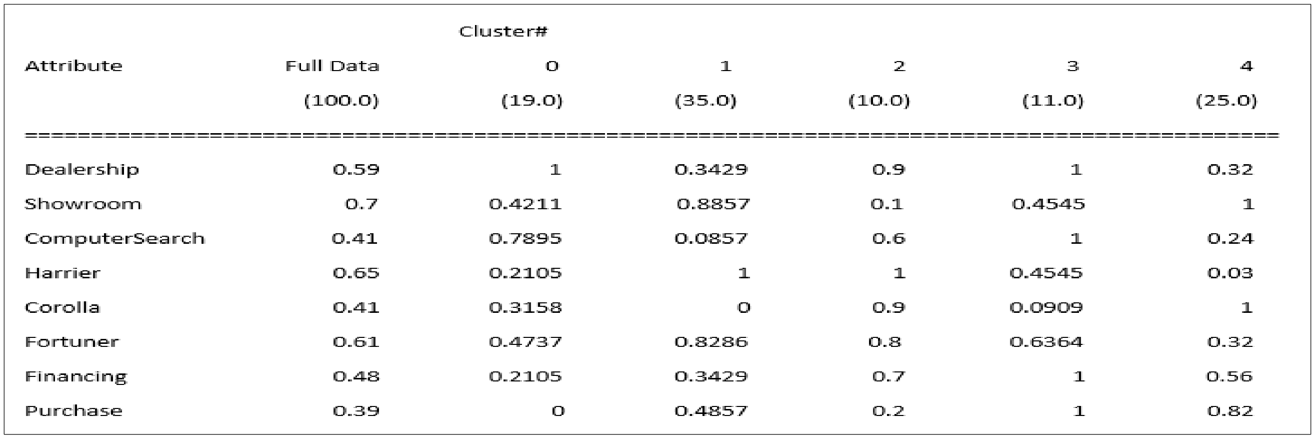
In Table VII, every row shows the purchase behaviour of customers, e.g., Customers went to the dealership without going on a showroom and done some computer search mostly interested in Toyota Harrier without financing they purchased it. These types of behaviour understandings about customers help Toyota to manage their sales. K-mean clustering allows the company to perform analysis without any efforts by finding patterns in a given dataset, shown in Fig. 4 and Fig. 5.

Fig. 4 explains that based on cluster 3, 100% of customers went for the dealership, whereas 45% went to the showroom too, and 100% of the customers also did computer searching. The majority of the customer that is 63% have shown interest in Fortuner, whereas 45% had shown interest in Harrier, and the least interest was found to be 9% in Corolla. These customers who 100% end up financing and purchasing a product consistently went to the dealership and done computer searching before buying an SUV car.

Meanwhile, based on cluster 4, only 32% of customers went to the dealership, whereas 100% went to the showroom, and 24% also did computer searching. Majority of the customers that are 100% interested in Corolla whereas 32% had shown interest in Fortuner, and the least interest was found to be 3% in Harrier; out of all these, 56% of the customers went for the financing details whereas 82% ends up purchasing a product. These are the customers looking for a small family car, i.e., Corolla, mostly approaching the showrooms.

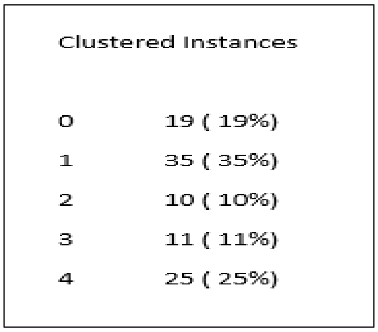
TABLE VII. Sample Dataset

No.	Dealership	Showroom	Computer Search	Harrier	Corolla	Fortuner	Financing
1	1	0	1	0	1	0	0
2	0	1	0	0	0	1	0
3	1	0	0	1	0	0	1
4	1	0	1	1	0	0	1
5	0	1	1	0	1	1	1
6	0	0	1	0	1	0	0
7	0	0	1	1	0	1	1
8	1	1	0	0	0	1	0



```
                              Cluster#
Attribute         Full Data          0          1          2          3          4
                    (100.0)     (19.0)     (35.0)     (10.0)     (11.0)     (25.0)
==================================================================================
Dealership             0.59          1     0.3429        0.9          1       0.32
Showroom                0.7     0.4211     0.8857        0.1     0.4545          1
ComputerSearch         0.41     0.7895     0.0857        0.6          1       0.24
Harrier                0.65     0.2105          1          1     0.4545       0.03
Corolla                0.41     0.3158          0        0.9     0.0909          1
Fortuner               0.61     0.4737     0.8286        0.8     0.6364       0.32
Financing              0.48     0.2105     0.3429        0.7          1       0.56
Purchase               0.39          0     0.4857        0.2          1       0.82
```

Fig. 4. Customers Purchase Behaviours.



```
Clustered Instances

0      19 ( 19%)
1      35 ( 35%)
2      10 ( 10%)
3      11 ( 11%)
4      25 ( 25%)
```

Fig. 5. Clustered Instances based on Customers Behaviour.

IV. Conclusion

This paper describes the different algorithms and methodologies used to handle large and small sets of data. The process of clustering is to group data based on their characteristics and similarities. Previously described the clustering models, many clustering techniques used to partition the data into a set of clusters. Algorithm selection should depend on the properties and the nature of the data collection because each algorithm has its pros and cons. This shows that there is no algorithm to manage all the clustering challenge. However, there are some algorithms to provide an optimist solution based on their sufficiency to face the challenges of the problem. To achieve high accuracy in terms of time and space, K-means would be the best choice for large and categorical data. However, we need to reduce their time and memory's complexity by upgrading Clustering Algorithms. However, a combined approach of the Genetic Algorithm with K-means can almost resolve all the issues of K-means. Genetic K-means Algorithm (GKA) speeds up the convergence to a globally optimum, and it concludes that GKA is faster than evolutionary Algorithms.

V. Future Directions and Open Issues

To date, Data Mining and information disclosure are advancing an essential innovation for businesses and scientists in numerous domains. Although information mining is extremely powerful, it faces innumerable difficulties during its usage. The problems could be identified with performance, data, strategies, and procedures utilized. The information mining measure becomes effective when the challenges or issues are distinguished accurately and sifted through appropriately.

Some of the following challenges and future directions are:

- Efficiency and Scalability of Algorithms: The data mining algorithms must be proficient and adaptable to extricate data from gigantic sums of information within the database. So, as a future direction, develop a parallel formulation of an Improved rough k-means algorithm to enhance the efficiency of an algorithm.
- Privacy and Security: Information mining ordinarily leads to genuine issues in terms of information security, protection, and administration. For case, when a retailer reveals his clients purchasing details without their permission. So, as a future direction, there needs to develop a single cache system and DES (Data Encryption Standard) techniques in any Clustering Algorithm to improve the privacy and security of data in the cloud.
- Complex Data Types: Complex data elements, objects with graphical data, temporal data, and spatial data may be included in the database. Mining of these types of data isn't practical to be done one device.
- Performance: The execution of the data mining framework depends on the proficiency of calculations and procedures are utilizing. The calculations and strategies planned are not up to the marked lead to influence the performance of the data mining process.

Therefore, as a future direction, we need to introduce a new hybrid approach of an Improved Rough k-means Algorithm, and the Genetic Algorithm will improve the performance and handles the complex data. The combination of Partitioning Clustering and Hierarchical Clustering Algorithms will also increase the accuracy of data analysis.

References

[1] S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma, "Machine learning techniques for data mining: A survey," 2013 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013, no. I, 2013.

[2] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, “A dynamic K-means clustering for data mining,” Indones. J. Electr. Eng. Comput. Sci., vol. 13, no. 2, pp. 521–526, 2019.

[3] Jiawei Han and M. Kamber, Data Mining: Concepts and Techniques Second Edition. 2013.

[4] D. Patel, R. Modi, and K. Sarvakar, “A Comparative Study of Clustering Data Mining: Techniques and Research Challenges,” Int. J. Latest Technol. Eng. Manag. Appl. Sci., vol. 3, no. 9, pp. 67–70, 2014.

[5] P. Indirapriya and D. D. K. Ghosh, “A Survey on Different Clustering Algorithms in Data Mining Technique,” Int. J. Mod. Eng. Res., vol. 3, no. 1, pp. 267–274, 2013.

[6] J. De Weerdt, S. Vanden Broucke, J. Vanthienen, and B. Baesens, “Active trace clustering for improved process discovery,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2708–2720, 2013.

[7] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, “A Comparative Study of Various Clustering Algorithms in Data Mining,” Int. J. Eng. Res. Appl. www.ijera.com, vol. 2, no. 3, pp. 1379–1384, 2012.

[8] V. W. Ajin and L. D. Kumar, “Big data and clustering algorithms,” in International Conference on Research Advances in Integrated Navigation Systems, RAINS 2016, 2016.

[9] A. Novikov, “PyClustering: Data Mining Library,” J. Open Source Softw., vol. 4, no. 36, p. 1230, 2019.

[10] D. Xu and Y. Tian, “A Comprehensive Survey of Clustering Algorithms,” Ann. Data Sci., vol. 2, no. 2, pp. 165–193, 2015.

[11] B. Karthikeyan, D. J. George, G. Manikandan, and T. Thomas, “A comparative study on k-means clustering and agglomerative hierarchical clustering,” Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5, pp. 1600–1604, 2020.

[12] P. K. Jain and R. Pamula, “Two-Step Anomaly Detection Approach.”

[13] A. Saxena et al., “A review of clustering techniques and developments,” Neurocomputing, vol. 267, pp. 664–681, 2017.

[14] S. Rashid, A. Ahmed, I. Al Barazanchi, and Z. A. Jaaz, “Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set,” Period. Eng. Nat. Sci., vol. 7, no. 2, pp. 448–457, 2019.

[15] M. S. Michael Steinbach George, Vipin Kumar, “A Comparison of Document Clustering Techniques,” TextMining Work. KDD2000, pp. 75–78.

[16] M. V. Reddy, M. Vivekananda, and R. U. V. N. Satish, “Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,” Int. J. Comput. Sci. Trends Technol., vol. 5, no. Sep-Oct, pp. 5–11, 2017.

[17] H. H. Ali and L. E. Kadhum, “K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition,” Int. J. Sci. Res., vol. 6, no. 8, pp. 1577–1584, 2017.

[18] T. H. T. Nguyen, D. T. Dinh, S. Sriboonchitta, and V. N. Huynh, “A method for k-means-like clustering of categorical data,” J. Ambient Intell. Humaniz. Comput., no. Berkhin 2002, 2019.

[19] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” J, vol. 2, no. 2, pp. 226–235, 2019.

[20] T. Zhang and F. Ma, “Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function,” Int. J. Comput. Math., vol. 94, no. 4, pp. 663–675, 2017.

[21] O. M. Lior Rokach, “Clustering methods,” Adv. Inf. Knowl. Process., no. 9781447167341, pp. 131–167, 2015.

[22] S. D. Bin Dong, Zengyou He, Xiaofei Xu, “K-Histrograms: An Efficient Clustering Algorithm for Categorical Dataset*,” no. 1, pp. 6–8, 2003.

[23] A. Al Malki, M. M. Rizk, M. A. El-Shorbagy, and A. A. Mousa, “Hybrid Genetic Algorithm with K-Means for Clustering Problems,” Open J. Optim., vol. 05, no. 02, pp. 71–83, 2016.

[24] B. Zerhari, A. A. Lahcen, and S. Mouline, “Big Data Clustering : Algorithms and Challenges,” Proc. Int. Conf. Bihree Charact. Call. 3Vs (Volume, Veloc. Var. It Ref. to data that are too large, Dyn. complex. this Context. data are difficult to capture, store, Manag. Anal. using Tradit. data Manag., no. May, pp. 1–7, 2015.

[25] C. C. Aggarwal, Data classification: Algorithms and applications. 2014.

[26] M. Duan, K. Li, X. Liao, and K. Li, “A Parallel Multiclassification Algorithm for Big Data Using an Extreme Learning Machine,” IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 6, pp. 2337–2351, 2018.

[27] H. Azzag, G. Venturini, A. Oliver, and C. Guinot, “A hierarchical ant based clustering algorithm and its use in three real-world applications,” Eur. J. Oper. Res., vol. 179, no. 3, pp. 906–922, 2007.

[28] C. C. Aggarwal and P. S. Yu, “Outlier detection for high dimensional data,” SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data), 2001.