Improvement of Data Classification Based on K-Value Selection Clustering Algorithm with Incomplete Data Clustering

P.S.Deshmukh[1]
Ph. D. Scholar of Dept of CSE
SRK, University
Bhopal (M.P.), India
pank1980@gmail.com

Dr. M. Sivakkumar[2]
Associate Professor in Dept of CSE
SRK, University
Bhopal (M.P.), India
sivakum.slm09@gmail.com

Dr. Varsha Namdeo[3]
Professor in Dept of CSE
SRK, University
Bhopal (M.P.), India
varsha_namdeo@yahoo.com

***Abstract*-** in today's world, most of the data is generated through computer applications. These applications can be used to predict and analyses the future. To achieve this phenomenon, we train the machines to read data and accordingly predict the future which is called as Machine Learning. Machine learning is done by training the machine to react to different data inputs. Many unsupervised learning approaches and algorithms have been introduced since the last decade where are well-known and widely used algorithms of unsupervised learning. The growing interest in applying unsupervised learning techniques forms a great success in fields such as computer vision, natural language processing, speech recognition, developing autonomous self-driving cars. Unsupervised learning eliminates the need for labelled data and manual handcrafted feature engineering enabling general, more flexible and automated ML methods. In the current age of the Fourth Industrial Revolution (4IR or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. K-means algorithm is one of the well-known unsupervised machine learning algorithms. The algorithm typically finds out distinct non-overlapping clusters in which each point is assigned to a group. The minimum squared distance technique distributes each point to the nearest clusters or subgroups. One of the K-means algorithm's main concerns is to find out the initial optimal centroids of clusters. Evaluation is performed on the Iris and novel power plant fan data with induced missing values at missingness rate of 5% to 20%. We show that both miss Forest and the k nearest neighbour can successfully handle missing values and offer some possible future research direction.

Keywords: Data Mining Tools, Machine Learning, Unsupervised learning, Clustering algorithms, Neural Networks, Time Complexity, big data.

DOI Number: 10.48047/NQ.2022.20.12.NQ77740 **NeuroQuantology2022;20(12): 4129-4146**

I. INTRODUCTION

The concepts of Machine Learning (ML) come from the domains of computer science and Artificial intelligence (AL), ML deals with systems that can learn from data instead of only executing the programmed commands overtly [1,2,3]. Furthermore, ML is closely linked to optimization and statistics, which brought their theories and approaches to the field. ML is utilized in different computing missions where constructing and programming rule-based, overt algorithms is not feasible. In certain cases, ML, pattern recognition, and data mining share their background [4,5,6,7].

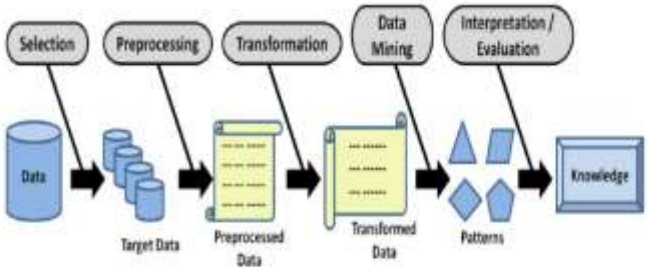

Fig1. The Knowledge Discovery in Databases (KDD) process [18]

Cluster analysis is one of the most important research directions in the field of data mining. "Things are clustered and people are grouped"; compared with other data mining methods, clustering can complete the classification of data without prior knowledge. Clustering algorithms can be divided into multiple types based on partitioning, density, and model [8].

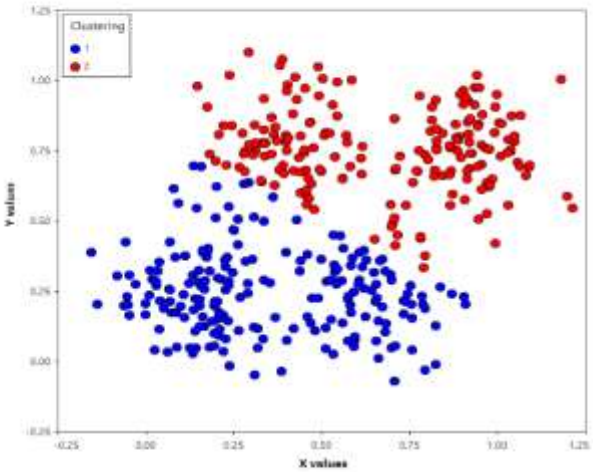

Fig 2. Here show two types of clusters [20]

A clustering algorithm is a process of dividing a physical or abstract object into a collection of similar objects. A cluster is a collection of data objects; objects in the same cluster are like each other and different from objects in other clusters [9]. For a clustering task, we want to get the objects as close as possible within the clusters: first cluster tends to sample or data point. However, the randomness of sample canter point selection tends to make cluster aggregation not converge. Cluster analysis is based on the similarity in clustering data sets, which is unsupervised learning. In the partition-based clustering algorithm, K-means algorithm has many advantages such as simple mathematical ideas, fast convergence, and easy implementation [10]. Therefore, the application fields are very broad, including different types of document classification, music, movies, classification based on user purchase behavior, the construction of recommendation systems based on user interests, and so on. With the increase of the amount of data, the traditional K-means algorithm has been difficult to meet the actual needs when analysing massive data sets. In view of the shortcomings of the traditional K-means algorithm, many scholars have proposed improvement measures based on K-means. For instance, in Reference [4], a simple and efficient implementation of the K-means clustering algorithm is presented to solve the problem of the cluster canter point not being well-determined; it built a kd-tree data structure for the data points. The algorithm is easy to implement and can effectively avoid entering the local optimal solution to some extent. For the problems of the traditional clustering algorithms having no way to take advantage of some background knowledge (about the domain or the data set), an Improved K-means Algorithm Based on Multiple Information Domains is presented in Reference [5]; they apply this method to six data sets and the real-world problem of automatically detecting road lanes from global positioning system (GPS) data. Experiments show that the improved algorithm is more correct when selecting K values when solving practical problems. Two algorithms which extend the k-means algorithm to categorical domains and domains are reported in Reference [11], through the pattern mixing algorithm, the combination of the effectiveness measure, in order to solve the problem of complex data and more noise in the real world. A principal Component Analysis (PCA) method is implemented in Reference [12]; they use the artificial neural network (ANN) algorithm and K-nearest neighbor (KNN) and support vector machine (SVM) classification algorithms to extract and analyze the features, which effectively realize the classification of malware. The clustering algorithm is also applied to the early detection of pulmonary nodules [13]; they propose a novel optimized method of feature selection for both cluster and classifier components. In the field of medical imaging, clustering and classification based on selection features effectively improve the classification performance of Computer-aided detection (CAD) systems. With the advent of deep learning methods in pattern recognition applications, some scholars have applied them to cluster analysis. For example, in Reference [14-18], by studying the performance of a CAD system for lung nodules in Computed tomography (CT) as a function of slice thickness, a method of comparing the performance of CAD systems using a training method using nonuniform data was proposed

II.RELATED WORK

In general, determining the optimal number of clusters and their centroids is left up to the researchers focus. There have been several attempts to find a solution for selecting the number of clusters and their centroids which gives the optimum solution in isolating a cluster. A general solution to determine the number of clusters is either to run the algorithm multiple times and select the desired number of clusters based on some validity criteria or determine automatically by some meaningful methods or criteria. In the same way the cluster centroids can be selected randomly and optimized by running the algorithm several times [21]. This paper focuses on the review works of automatic selection of optimal number of clusters and the cluster centroids along with the comparison.

Wang et. al. [21] we propose a novel K-means based clustering algorithm which unifies the clustering and imputation into one single objective function. It makes these two processes be negotiable with each other to achieve optimality. Furthermore, we design an alternate optimization algorithm to solve the resultant optimization problem and theoretically prove its convergence. The comprehensive experimental study has been conducted on nine UCI benchmark datasets and real-world applications to evaluate the performance of the proposed algorithm, and the experimental results have clearly demonstrated the effectiveness of our algorithm which outperforms several commonly-used methods for incomplete data clustering

M.R.Rezaee et. al. [22] proposed a segmentation technique for segmenting clinical images. Pyramidal approach is used to view the image at multilevel. Cluster validation indices such as partition coefficient and partition entropy are measured to find the number of clusters automatically. Then fuzzy c-means clustering is used for merging purpose. The proposed algorithm works well compared to the conventional algorithms.

Fahim A. M et. al. [23] proposed an efficient enhanced kmeans clustering for real dataset and synthetic dataset that improves the computational speed by reducing algorithm’s time complexity. In a large dataset the gravity center of the spherical shaped cluster is chosen as the centroids. As very few pixels are far away from the gravity center, the distance calculation becomes easy. The selection of best centroids leads to reduced number of iterations, which results in time complexity. The proposed algorithm is better than the existing conventional algorithms.

Anindya Bhattacharya et. al. [24] proposed a divisive correlation clustering algorithm for genetic engineering. Initially the number of clusters is considered as one, then the

Pearson correlation coefficient for all pair of gene within a cluster. If a gene has negative correlation, then that gene is kept in another cluster. Thus, this algorithm produces the k number of cluster and centroids.

Mark Junjie Li et. al. [25] proposed an agglomerative fuzzy k-means Clustering algorithm for synthetic data and real data. a negative entropy term is introduced in the objective function of k-means clustering. The introduced negative entropy minimizes the objective function as well as cluster dispersion. Also, different values of are used in the algorithm to find the optimal number of true clusters as well as the true centroids. The mentioned validity indices are better for the proposed algorithm than classical algorithm.

Siti Noraini Sulaiman et. al. [26] proposed adaptive fuzzy k-means clustering for the consumer electronic related indoor and outdoor images taken on a digital camera. Initially the pixels are assigned to its cluster by calculating the Euclidean distance. A quantitative parameter called belongingness is introduced and the membership matrix is updated, the centroids are calculated from the updated membership matrix. Statistical evaluations prove the effectiveness of the algorithm compared with conventional algorithm.

Ujjwal Maulik et. al. [27] proposed an automatic fuzzy clustering algorithm based on modified differential evolution for satellite image segmentation. A fitness function called Xie-Beni index using Euclidean distance is measured for proposer clustering. Modified differential evolution selects the optimal number of clusters based on the fitness function measured for the proper clustering. The proposed algorithm shows good qualitative and quantitative results compared to existing algorithm.

Baolin Yi et. al. [28] proposed improved version of kmeans clustering for machine learning database. Density of the objects is found using Euclidean distance and Gaussian density function. The initial centroids are taken as the samples that have maximum density. The proposed algorithm shows high purity and less sensitive to initial centroids compared to existing algorithm.

Chen et. al. [29] The paper proposes an incomplete high-dimensional big data clustering algorithm based on feature selection and partial distance strategy. First, a hierarchical clustering-based feature subset selection algorithm is designed to reduce the dimensions of the data set. Next, a parallel k-means algorithm based on partial distance is derived to cluster the selected data subset in the first step. Experimental results demonstrate that the proposed algorithm achieves better clustering accuracy than the existing algorithms and takes significantly less time than other algorithms for clustering high-dimensional big data.

Chau et. al. [30] Hence, author define a robust and effective algorithmic framework for incomplete educational data clustering using the nearest prototype strategy. Within the framework, we propose two novel incomplete educational data clustering algorithms K_nps and S_nps based on the k-means algorithm and the self-organizing map, respectively. Experimental results have shown that the clusters from our proposed algorithms have better cluster quality as compared to the different existing approaches.

Yan et. al. [31] the clustering results use K-means algorithm as the initial scope of EM algorithm, according to the different choice of different characteristics of mining purposes, then use incremental EM algorithm (IEM) step by step EM iterative refinement repeatedly, it obtains the optimal value of filling missing data quickly and efficiently. it is concluded that the optimal value of filling missing data experimental results show that the algorithm of this paper to speed up the convergence rate, strengthened the stability of clustering, data filling effect is remarkable.

A et. al. [32] proposed a modified fuzzy cmeans clustering algorithms for background removal purpose. The co-occurrence matrixes, especially the diagonal elements are selected as the initial centroids. The qualitative results are better for the proposed algorithm than conventional algorithm.

Min Ren et. al. [33] therefore, propose Spectral Ensemble Clustering (SEC) to leverage the advantages of co-association matrix in information integration but run more efficiently. We disclose the theoretical equivalence between SEC and weighted K-means clustering, which dramatically reduces the algorithmic complexity. We also derive the latent consensus function of SEC, which to our best knowledge is the first to bridge co-association matrix-based methods to the methods with explicit global objective functions. Further, we prove in theory that SEC holds the robustness, generalizability, and convergence properties. We finally extend SEC to meet the challenge arising from incomplete basic partitions, based on which a row-segmentation scheme for big data clustering is proposed. Experiments on various real-world data sets in both ensemble and multi-view clustering scenarios demonstrate the superiority of SEC to some state-of-the-art methods. In particular, SEC seems to be a promising candidate for big data clustering.

Chau et. al. [34] Hence, incomplete data clustering has been considered in many research works with many different approaches based on the well-known existing clustering algorithms such as k-means, fuzzy c-means, the self-organizing map (SOM), mean shift, etc. However, few of them have examined both effectiveness and robustness of the incomplete data clustering algorithms. Some of them are not practical due to a lot of parameters in hybrid approaches and/or cannot handle incomplete data which appear in any object at any dimension. In contrast, this paper aims at a SOM-based incomplete data clustering algorithm, iS nps, which is a robust and effective solution to clustering incomplete data in a simple but practical approach. Is nps can do clustering on incomplete data as well as estimate incomplete data using the nearest prototype strategy in an iterative manner. As compared to several different existing approaches, our proposed algorithm can produce the clusters of good quality and a better approximation of incomplete data via the experiments on benchmark data sets.

Honda et. al. [35] author proposed, the PCA-guided k-Means procedure is extended to a situation in which some observations are missing. Principal component scores, which can be identified with a rotated solution of cluster indicators of k-Means clustering, are estimated in an iterative process without imputation. Besides solving the eigenvalue problem of covariance matrices, k-Means-like partitions are derived through lower rank approximation of the data matrix ignoring missing elements. Several experimental results demonstrate that the PCA-guided process is more robust to initialization problems even though it is based on iterative optimization, just as the k-Means procedure is.

Vauski et. al. [36] this paper author examines a comparative study of different methods with advantage and drawbacks. Performing spectral ensemble cluster (SEC) via weighted k-means are not efficient to handle incomplete basic partitions and big data problems. To overcome the problems in SEC, Greedy k-means consensus clustering is combined with SEC. By solving the above challenges, named spectral greedy k-means consensus clustering (SGKCC) is proposed. The proposed SGKCC efficient to handle incomplete basic partitions in big data which enhance the quality of single partition. Extensive evaluation NMI and RI used to calculate the performance efficiency compared with existing approach proving the result of proposed algorithm.

Pugazhenthi A. et. al. [37] In image segmentation, clustering is the process of sub dividing the whole image into the meaningful sub images. The most commonly used image segmentation algorithms such as K-means and Fuzzy c-means clustering face the specific important problem in selecting the optimal number of clusters and the corresponding cluster centroids. Plenty of research works have been done on the limitations of the said clustering algorithms to improve the efficient isolation of clusters. This paper enumerates the works done by different researchers in selecting the initial number of clusters and the centroids using K-means and Fuzzy c-means clustering. The limitations and applications of the above-mentioned clustering algorithms are explored

X. Liu et. al. [38], Multiple kernel clustering (MKC) algorithms optimally combine a group of pre-specified base kernel matrices to improve clustering performance. However, existing MKC algorithms cannot efficiently address the situation where some rows and columns of base kernel matrices are absent. This paper proposes two simple yet effective algorithms to address this issue. Different from existing approaches where incomplete kernel matrices are first imputed and a standard MKC algorithm is applied to the imputed kernel matrices, our first algorithm integrates imputation and clustering into a unified learning procedure. Specifically, we perform multiple kernel clustering directly with the presence of incomplete kernel matrices, which are treated as auxiliary variables to be jointly optimized. Our algorithm does not require that there be at least one complete base kernel matrix over all the samples. Also, it adaptively imputes incomplete kernel matrices and combines them to best serve clustering. Moreover, we further improve this algorithm by encouraging these incomplete kernel matrices to mutually complete each other. The three-step iterative algorithm is designed to solve the resultant optimization problems. After that, we theoretically study the generalization bound of the proposed algorithms. Extensive experiments are conducted on 13 benchmark data sets to compare the proposed algorithms with existing imputation-based methods. Our algorithms consistently achieve superior performance and the improvement becomes more significant with increasing missing ratio, verifying the effectiveness and advantages of the proposed joint imputation and clustering.

Bay et. al. [39] Advances in data collection and storage have allowed organizations to create massive, complex and heterogeneous databases, which have stymied traditional methods of data analysis. This has led to the development of new analytical tools that often combine techniques from a variety of fields such as statistics, computer science, and mathematics to extract meaningful knowledge from the data. To support research in this area, UC Irvine has created the UCI Knowledge Discovery in Databases (KDD) Archive (http://kdd.ics.uci.edu) which is a new online archive of large and complex data sets that encompasses a wide variety of data types, analysis tasks, and application areas. This article describes the objectives and philosophy of the UCI KDD Archive. We draw parallels with the development of the UCI Machine Learning Repository and its affect on the Machine Learning community.

Yang et. al. [40] This paper presents an effective Bayesian network model for medical diagnosis. The proposed approach consists of two stages. In the first stage, a novel feature selection algorithm with consideration of feature interaction is used to get an undirected network to construct the skeleton of BN as small as possible. In the second stage for greedy search, several methods are integrated together to enhance searching performance by either pruning search space or overcoming the optima of search algorithm. In the experiments, six disease datasets from UCI machine learning database were chosen and six off-the-shelf classification algorithms were used for comparison. The result showed that the proposed approach has better classification accuracy and AUC. The proposed method was also applied in a real-world case for hypertension prediction. And it presented good capability of finding high risk factors for hypertension, which is useful for the prevention and treatment of hypertension. Compared with other methods, the proposed method has the better performance.

Table 1. Literature review on the selection of optimal number of clusters and ACC

Authors	Study Purpose	Methods / Algorithms	Outcome Measures
wang et al. [21]	K-means based clustering algorithm which unifies the clustering and imputation into one single objective function, f-score and accuracy	K-means clustering	Better estimation of optimal number of clusters and low ACC
Mahmoud Ramze Rezaee et al. [22]	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	Fuzzy c-means clustering	Detected ventricular volume in magnetic resonance images and low ACC
Fahim A. M et al. [23]	Real dataset and accurate data analysis and error minimization and accuracy	K-means clustering	Improved time complexity and low ACC
Anindya Bhattacharya et al. [24]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution	K-means clustering and fuzzy c-means clustering	Very high biological significance on clustering of gene expression dataset and low ACC
Mark Junjie Li et al. [25]	run more efficiently and Remove copy data in clustering and optimal solution and accurate data analysis	K-means clustering and fuzzy c-means clustering	Produced consistent clustering result with best determination of optimal number of centroids
Siti Noraini Sulaiman et al. [26]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution and accuracy	K-means clustering and fuzzy c-means clustering	Good segmentation results with better qualitative and quantitative results and low ACC
Ujjwal Maulik et al. [27]	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	Pyramidal image segmentation and fuzzy cmeans clustering	Better qualitative and quantitative results
Baolin Yi et al. [28]	run more efficiently and Remove copy data in clustering and optimal solution and accurate data analysis	K-means clustering	Clusters with high purity and less sensitive to initial assumption
Chen et al. [29]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution and accuracy	K-means clustering	Good segmentation results with less cluster variance
Chau et al. [30]	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	K-means clustering	Found better clusters in less time and low ACC
Yan et al. [31]	run more efficiently and Remove copy data in clustering and optimal solution and accurate data analysis and accuracy	K-means clustering	Effective segmentation of fish from complex background
Pugazhenthi A et al. [32]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution	Fuzzy c-means clustering	Segmentation of objects from complex background
Liu et al. [33]	run more efficiently and Remove copy data in clustering and optimal solution and accurate data analysis and accuracy	Fuzzy c-means clustering	Optimal or stable clustering result with less number of iterations and low ACC
Chau et al. [34]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution	K-means clustering	Improved time complexity and number of iterations required and low ACC
Honda et al. [35]	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	K-means clustering	Improved efficiency along with reduced time complexity and low ACC
Vauski et al. [36]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution and accuracy	K-means clustering	Optimized accuracy in text clustering and reduced time complexity and low ACC
Pugazhenthi A et al. [37]	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	K-means clustering and fuzzy c-means clustering	Segmented the satellite image into high level clouds, middle level clouds Optimized accuracy in text clustering and reduced time complexity
X. Liu et al. [38]	run more efficiently and Remove copy data in clustering and optimal solution and accurate data analysis	Multiple Kernel kk-Means with Incomplete Kernels	Optimal or stable clustering result with less number of iterations but low AC
Bay et. al. [39]	feature selection and incomplete data analysis, Remove copy data in clustering and optimal solution	The UCI KDD archive of large data sets for data mining research and experimentation	Optimized accuracy and number of iterations required more better estimation of optimal number of clusters
Yang et. al. [40]	resolves the incomplete data clustering task in feature selection and incomplete data analysis, remove copy data in clustering and optimal solution, Remove copy data in clustering and optimal solution	A consistency contribution-based Bayesian network model for medical diagnosis	Optimized accuracy and number of iterations required more and reduced time complexity and low ACC

Table 1 shows the literature review on the selection of optimal number of clusters and cluster centroids for K-means and fuzzy c-means clustering algorithm. The K-means and fuzzy c-means clustering algorithms have wide range of applications and applied for different types of images as well as data. The qualitative metrics used for analyses also vary with the applications. So, it is difficult to conclude or suggest a universal method to select the optimal number of clusters and their centroids. The optimal method for the selection of number of clusters and their centroids will be selected based on the qualitative metrics requirement of segmentation process and also based on applications.

The optimal choice of number clusters shows improvement in the quantitative parameters. Peak Signal to Noise Ratio, Structural Content, Means Squared Error, Structural Similarity Index, Universal Quality Index, Correlation Coefficient and Image Fidelity are some of the parameters that shown good improvement for optimal selection than random selection of number of centroids and low ACC [4], [22].

III.PROPOSED METHODOLOGY

(a) overview proposed work: Clustering is a standard procedure in multivariate data analysis. It is designed to explore an inherent natural structure of the data objects, where objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible. The equivalence classes induced by the clusters provide a means for generalizing over the data objects and their features. Clustering methods are applied in many domains, such as medical research, psychology, economics and pattern recognition. Human beings often perform the task of clustering unconsciously; for example, when looking at a two-dimensional map one automatically recognizes different areas according to how close to each other the places are located, whether places are separated by rivers, lakes or a sea, etc. However, if the description of objects by their features reaches higher dimensions, intuitive judgements are less easy to obtain and justify. The term clustering is often confused with a classification or a discriminant analysis. But the three kinds of data analyses refer to different ideas and are distinguished as follows: Clustering is (a) different from a classification, because classification assigns objects to already defined classes, whereas for clustering no a priori knowledge about the object classes and their members is provided. And a cluster analysis is (b) different from a discriminant analysis, since discriminant analysis aims to improve an already provided classification by strengthening the class demarcations, whereas the cluster analysis needs to establish the class structure first. Clustering is an exploratory data analysis. Therefore, the explorer might have no or little information about the parameters of the resulting cluster analysis. In typical uses of clustering the goal is to determine all of the following: The number of clusters, The absolute and relative positions of the clusters, The size of the clusters, The shape of the clusters, The density of the clusters. The cluster properties are explored in the process of the cluster analysis, which can be split into the following steps. 1. Definition of objects: Which are the objects for the cluster analysis 2. Definition of clustering purpose: What is the interest in clustering the objects 3. Definition of features: Which are the features that describe the objects 4. Definition of similarity measure: How can the objects be compared 5. Definition of clustering algorithm: Which algorithm is suitable for clustering the data 6. Definition of cluster quality: How good is the clustering result, What is the interpretation, Depending on the research task, some of the steps might be naturally given by the task, others are not known in advance. Typically, the understanding of the analysis develops iteratively with the experiments. The following sections define a cluster analysis with respect to the task of clustering verbs into semantic classes. Proposed unsupervised learning algorithm, Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is Proposed unsupervised learning algorithm m, how the algorithm works, along with the MATLAB implementation of k-means clustering. Proposed algorithm Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters

(i)Preprocessing: In the preprocessing phase, raw data is prepared for subsequent clustering and classification steps:
(ii)Data Cleaning: Missing values are addressed using mean imputation. For each feature with missing values, the missing entries are replaced with the mean value of the available entries in that feature.
(iii)Data Transformation: Feature scaling is applied to ensure that all features contribute equally to clustering and classification. Features are normalized to have zero mean and unit variance.
(iv) Data Objects, Clustering Purpose and Object Features:This work is concerned with inducing a classification of German verbs, i.e. the data objects in the clustering experiments are German verbs, and the clustering purpose is to investigate the automatic acquisition of a linguistically appropriate semantic classification of the verbs. The degree of appropriateness is defined with respect to the

ideas of a verb classification at the syntax-semantic interface in Chapter 2. Once the clustering target has been selected, the objects need an attribute description as basis for comparison. The properties are grasped by the data features, which describe the objects in as many dimensions as necessary for the object clustering. The choice of features is of extreme importance, since different features might lead to different clustering results. Kaufman and Rousseeuw (1990, page 14) emphasise the importance by stating that 'a variable not containing any relevant information is worse than useless, because it will make the clustering less apparent by hiding the useful information provided by the other variables. Possible features to describe German verbs might include any kind of information which helps classify the verbs in a semantically appropriate way. These features include the alternation behaviour of the verbs, their morphological properties, their auxiliary selection, adverbial combinations, etc. Within this thesis, I concentrate on defining the verb features with respect to the alternation behaviour, because I consider the alternation behaviour a key component for verb classes as defined in Chapter 2. So I rely on the meaning-behaviour relationship for verbs and use empirical verb properties at the syntax-semantic interface to describe the German verbs. The verbs are described on three levels at the syntax-semantic interface, each of them refining the previous level by additional information. The first level encodes a purely syntactic definition of verb subcategorization, the second level encodes a syntactico-semantic definition of subcategorization with prepositional preferences, and the third level encodes a syntactico-semantic definition of subcategorization with prepositional and selectional preferences. So the refinement of verb features starts with a purely syntactic definition and step-wise adds semantic information. The most elaborated description comes close to a definition of the verb alternation behaviour. I have decided on this three-step proceeding of verb descriptions, because the resulting clusters and even more the changes in clustering results which come with a change of features should provide insight into the meaning-behaviour relationship at the syntax-semantic interface. The exact choice of the features is presented and discussed in detail in the experiment setup in Chapter 5. The representation of the verbs is realised by vectors which describe the verbs by distributions over their features. As explained in Chapter 1, the distributional representation of features for natural language objects is widely used and has been justified by Harris (1968). The feature values for the distributions are provided by the German grammar, as described (i) real values f representing frequencies of the features, (ii) real values p representing probabilities of the features, and (iii) binary values ,Generally speaking, a standardisation of measurement units which converts the original measurements (such as frequencies) to unitless variables (such as probabilities) on the one hand may be helpful by avoiding the preference of a specific unit, but on the other hand might dampen the clustering structure by eliminating the absolute value of the feature

(v) Data Similarity Measures: With the data objects and their features specified, a means for comparing the objects is needed. The German verbs are described by features at the syntax-semantic interface, and the features are represented by a distributional feature vector. A range of measures calculates either the distance d or the similarity sim between two objects x and y. The notions of 'distance' and 'similarity' are related, since the smaller the distance between two objects, the more similar they are to each other All measures refer to the feature values in some way, but they consider different properties of the feature vector. There is no optimal similarity measure, since the usage depends on the task. Following, I present a range of measures which are commonly used for calculating the similarity of distributional objects. I will use all of the measures in the clustering experiments and using Datasets they evaluate the proposed algorithm on several UCI and several large benchmark datasets. They are Iris, Wine, Glass, Breast Cancer, Mice Protein,1 Ovarian Cancer Dataset, Pen Digits [41], Avila and Sensorless Drive

(b)Working Process: Improvement performance of data analysis using Proposed algorithm unsupervised. Machine learning algorithms are classified into supervised / unsupervised clustering techniques in which a supervised algorithm uses its trained experiences while unsupervised methods utilize the visible similarity / differences of current objects. The goal of unsupervised methods needs to ensure that the objects assembled in a similar group are comparable and predictable as indicated by explicit constraints. It is hard to put on information or data mining grouping measures in Big Data due to the new problems. As the amount of data / information increases, the computational complexity, handling and examination costs of clustering algorithms increase. Further, these complex situations with high volume of information increase the burden of clustering algorithms to produce effective clusters in limited time. These clustering produce effective clusters in limited time. These clustering strategies can be divided into: partitioning based, model All things are considered from the start as a lone collection. The things are sheltered into number of assignments by iteratively discovery the areas between the segments. Partitioning technique behaviours one -level partitioning on informational collection, first it makes starting arrangement of k segment, where parameter k is the quantity of segments to build. It at that point utilizes an iterative movement strategy that endeavours to enhance the partitioning by moving items starting with one group then onto the next group. Usually, partitioning technique incorporates two prominent algorithms called, Proposed algorithm in many real time situations, the information regarding the optimum number of clusters will be unavailable and hence clustering algorithm requires additional pre-processing step to determine the number of clusters. When these inputs are online with no prediction of future data, k means clustering are not practical. Data Mining (or information Discovery in Databases) describes the big idea of finding "interesting" patterns in large collections of information. There's a large quantity of data available within the data business. This knowledge is of no use till its regenerate into helpful information. It’s necessary to research this large quantity of data and extract helpful information from it. Extraction of knowledge isn't the only method we need to perform; data mining so describes the abstract goals of what must be done, and depends upon a large range of various techniques to achieve them, like artificial neural networks, cluster analysis different processes such as data cleaning, data Integration, information Transformation, data mining, Pattern analysis and knowledge Presentation

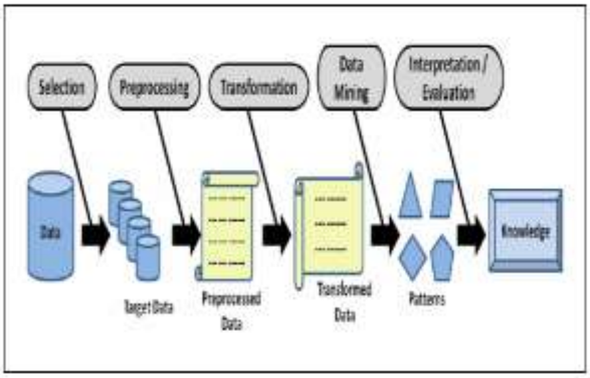

Fig3. block diagram of information detection Process

In these processes are overcome problem data error, they might be able to use this data in several applications like error detection in large data set, market dataset analysis research, Production control, Science Exploration, etc. clustering could also be defined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than individual data point. Basically, clustering is justified as a method used for grouping a large variety of knowledge into significant teams or clusters supported similarity types of objects data. Clusters area unit the teams that have knowledge similar on basis of common options and dissimilar to knowledge in different clusters Cluster analysis teams’ knowledge objects primarily based solely on data found in knowledge that describes the objects and their relationships.

(c)Proposed Enhancement Clustering Algorithm

(i)Proposed Algorithm:PA cluster is an unsupervised hard partitioning cluster technique. the target is to search out k clusters from the information based on the target function J given in eq. (1).J=∑i=1k∑j=1Nd2 (Ci−Xj) Where d2 (Ci-Xj) is that the square Euclidian distance between ith cluster centroid and jth information N is that the total variety of information points supported the gap obtained, the points are appointed to the cluster with minimum distance from the centroid. Once the points are clustered the mean of all points belonging to the cluster is found. Then mean is assigned because the new cluster centroid for future iteration. This method is perennial till the centroid obtained is same as that of the previous iteration. The aim of proposed algorithm rule is to attenuate the target perform however suboptimal resolution. proposed algorithm was introduced by European nation and additional represented by Goldberg as improvement approaches to search out a worldwide or near-global optimum resolution. proposed algorithm starts with a group of potential solutions. Next, genetic operators (selection, mutation and crossover) are applied one once another to get a brand-new generation of chromosomes. This method is perennial till the termination criterion is met. Algorithmically, the fundamental steps of PA are made public as below.

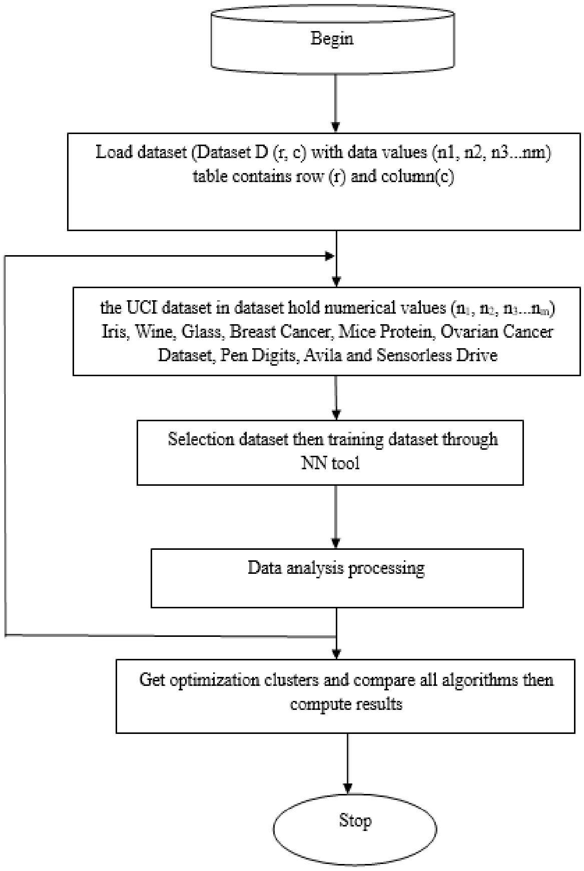
Begin
Load dataset (Dataset D (r, c) with data values (n1, n2, n3...nm) table contains row (r) and column(c)
the UCI dataset in dataset hold numerical values (n_1, n_2, n_3...n_m) Iris, Wine, Glass, Breast Cancer, Mice Protein, Ovarian Cancer Dataset, Pen Digits, Avila and Sensorless Drive
Selection dataset then training dataset through NN tool
Data analysis processing
Get optimization clusters and compare all algorithms then compute results
Stop

Fig4. Flow Diagram PA

Classification using Enhanced Clusters: The enhanced clusters generated through the integrated approach are utilized to improve classification performance: Cluster Assignment: Each data point is assigned to the cluster obtained from the proposed algorithm process. Feature Engineering: Features are engineered based on cluster properties: distance to centroid: Calculate the distance of each data point to the centroid of its assigned cluster. Cluster Density: Compute the inverse of the average intra-cluster distance to capture cluster density. Classifier Training: Employ a machine learning classifier, such as a Random Forest classifier, using the enriched features for precise data classification 3. Experimental Setup: A comprehensive experimental setup is crucial to validate the proposed approach:\Datasets: Select diverse real-world datasets containing varying degrees of missing values. This ensures the evaluation's thoroughness and applicability. Performance Metrics: Adopt standard performance metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve to quantitatively measure the effectiveness of the proposed algorithm. Baseline Comparison: Compare the performance of the proposed algorithm against conventional classification techniques and existing methods for handling incomplete data.

(ii)Algorithm: K-Value Selection Clustering with Incomplete Data Clustering

Input: Data set with incomplete data Number of clusters, K
Output: Cluster assignments for each data point

(iii)Procedure: Data Preprocessing: Handle missing or incomplete data points using appropriate techniques such as imputation or data completion. Initialization: Randomly initialize K cluster centroids. These centroids can be selected from the available data points. Repeat Until Convergence:

Step I a. Assignment Step: For each data point, calculate the distance to each cluster centroid. Assign the data point to the cluster with the nearest centroid using a distance metric such as Euclidean distance., initiates load dataset in mat lab tool (Dataset D (r, c) with N instance values (n_1, n_2, n_3..n_m) .

Step II Update Step: Recalculate the centroids of each cluster by taking the mean of all data points assigned to that cluster. Datasets they evaluate the proposed algorithm on several UCI and several large benchmark datasets. the UCI dataset in dataset hold numerical values (n1, n2, n3...nm), They are Iris, Wine, Glass, Breast Cancer, Mice Protein,1 Ovarian Cancer Dataset, Pen Digits [41], Avila and Sensorless Drive

Step III K-Value Selection: Evaluate the clustering quality using a criterion like the silhouette score, Davies-Bouldin index, or within-cluster sum of squares (WCSS). Increment or decrement the value of K and repeat steps (a) and (b) for different K values. Select the K value that yields the best clustering quality based on the chosen criterion. an analysis or evaluated to dataset is created by repeating the following steps:

Step IV Convergence Check:

Check if the cluster centroids have converged. Convergence can be determined by assessing whether the centroids have significantly changed between iterations or by setting a maximum number of iterations.

Accepting new offspring based on vector space model is placed in optimal values (the new population). Go to stepII.

Step V Generated output based on minimization error values in dataset. The objective is to find k clusters from the data based on the objective function J given in Eq. (1).

$$J=\sum i=1k\sum j=1Nd2\ (Ci-Xj)$$

Where

d^2 $(C_i\text{-}X_j)$ is the squared Euclidean distance between ith cluster centroid and jth data point N is the total number of data points Based on the distance obtained, the points are assigned to the cluster with minimum distance from the centroid. After the points are clustered the mean of all points belonging to the cluster is found.

Step VI Output: Return the cluster assignments for each data point, as well as the final cluster centroids. If the end condition is satisfied, return the best solution in current dataset and go to step next step.

Step VII Stop.

IV. EXPERIMENTS SETUP

(a)Simulation Tool: MATLAB (matrix laboratory) is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization a programming. MATLAB is developed by MathWorks. It allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications. It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods. MATLAB's Power of Computational Mathematics MATLAB is used in every facet of computational mathematics. Following are some commonly used mathematical calculations where it is used most commonly, Dealing with Matrices and Arrays,2-D and 3-D Plotting and graphics, Linear Algebra, Data Analysis etc.

(b)Experiments Used Datasets: Datasets they evaluate the proposed algorithm on several UCI and several large benchmark datasets. They are Iris, Wine, Glass, Breast Cancer, Mice Protein, Ovarian Cancer Dataset, Pen Digits [41], Avila and Sensorless Drive. The detailed information of these datasets is listed, randomly generate the incompleteness by the original complete data matrix. The first five datasets, including Iris, Wine, Glass, Ovarian and Breast Cancer And three 3 are the most commonly-used benchmarks for incomplete data clustering

(i)Mice Protein is a dataset that consists of the expression levels of 77 proteins measured in the cerebral cortex of eight classes of control and trisomic mice. Differently

(ii)Pen Digits is a hand-written digit with 10992 samples for 10 classes. Moreover.

(iii)Avila and Sensorless Drive are downloaded from the UCI Machine Learning Repository. The Avila dataset4 has been extracted from 800 images of the 'Avila Bible', an XII century giant Latin copy of the Bible, which has 20871 samples in 12 classes.

(iv)The Sensorless Drive dataset 5 extracts the features from electric current drive signals, resulting in 37715 samples with 8 classes

Experimental result our main goal is to improve data analysis using the proposed clustering algorithm's performance. In this sub-section, they will show the experimental results with different datasets (Iris, Wine, Glass, Breast Cancer, Mice Protein,1 Ovarian Cancer Dataset, Pen Digits, Avila and Sensorless Drive), they have used high dimension datasets, large instances with relatively low dimensions and very high instances. The first five datasets, including Iris, Wine, Glass, Ovarian and Breast Cancer 3 are the most commonly-used bench marks for incomplete data clustering. Mice Protein is a dataset that consists of the expression levels of 77 proteins measured in the cerebral cortex of eight classes of control and trisomic mice. Differently, Pen Digits is a hand-written digit with 10992 samples for 10 classes. Moreover, Avila, and Sensorless Drive are downloaded from the UCI Machine Learning Repository. The Avila dataset4 has been extracted from 800 images of the 'Avila Bible', an XII century giant Latin copy of the Bible, which has 20871 samples in 12 classes. The Sensorless Drive dataset 5 extracts the features from electric current drive signals, resulting in 37715 samples with 8 classes[21]. Description of UCI Data Sets In this section, we tested the performance of our method in different real data sets: Iris, Glass, Wine, and Wisconsin Breast cancer, as shown in Table 3. These data sets are referenced in Benchmark of UCI Repository of Machine Learning Data base [31].

V. EXPERIMENTS RESULT ANALYSIS

(a) Iris Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 1 Iris dataset analysis. Iirs data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure3 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

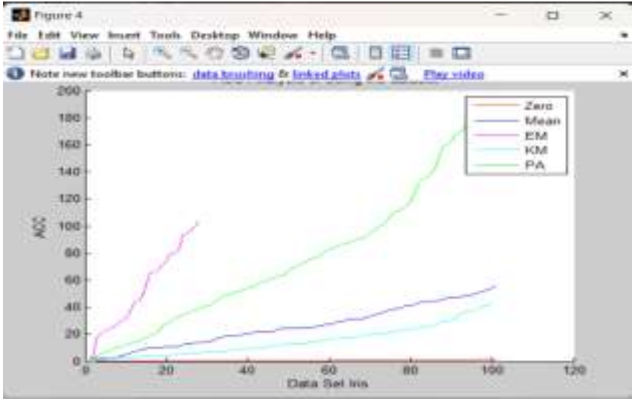

Fig3.Iris dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 1 Iris dataset analysis. Iirs data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure4 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

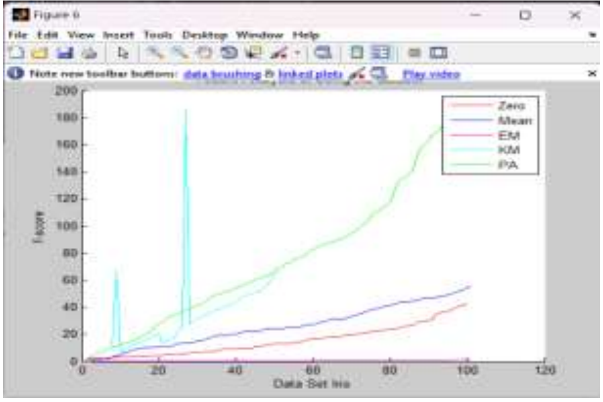

Fig4. Iris dataset f-score analysis between existing algorithm and proposed algorithm

(b) Glass Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system

(PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 2 glass dataset analysis. glass dataset analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure5 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

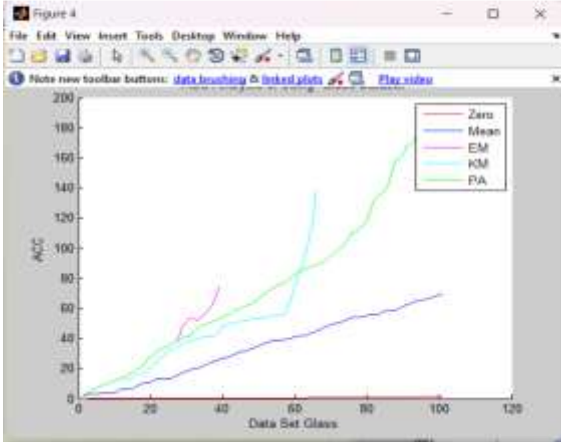

Fig5. Glass dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 2 glass dataset analysis. glass dataset analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure5.16 below the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

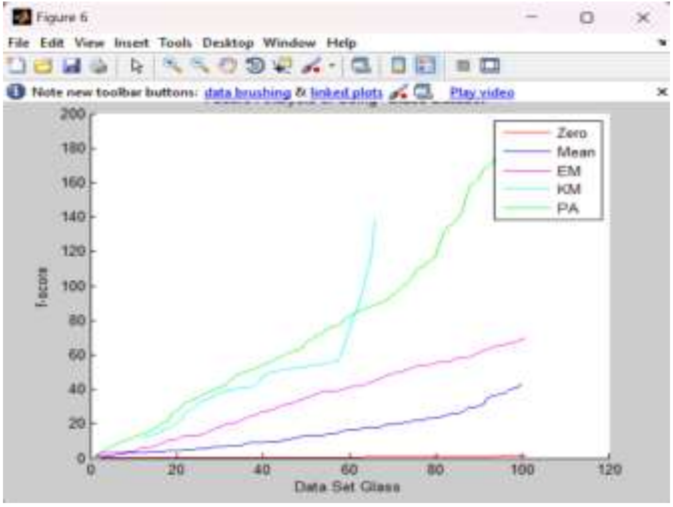

Fig6. Glass dataset f-score analysis between existing algorithm and proposed algorithm

(c) Wine Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 3 wine dataset analysis. wine data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure7below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

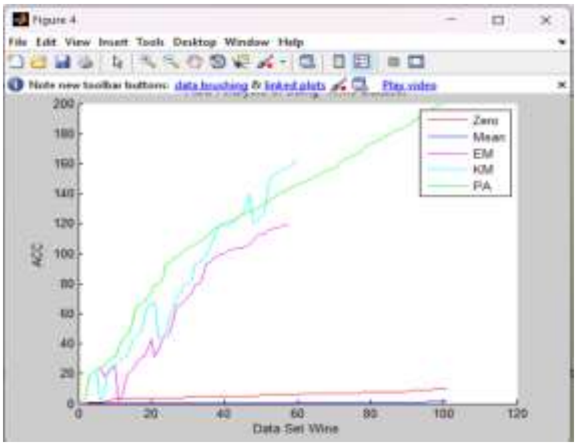

Fig7. Wine dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 3 wine dataset analysis. wine data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure8 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

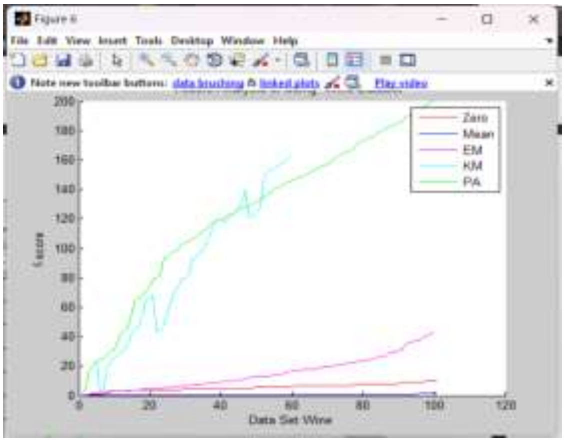
Figure 6
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
f-score
0
20
40
60
80
100
120
140
160
180
200
Data Set Wine
Zero
Mean
EM
KM
PA

Fig8. Wine dataset f-score analysis between existing algorithm and proposed algorithm

(d) Breast cancer Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 14 breast cancer dataset analysis. breast cancer data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure9 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

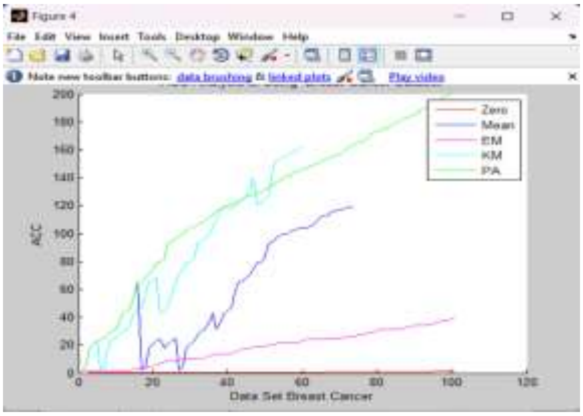
Figure 4
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
ACC
0
20
40
60
80
100
120
140
160
180
200
Data Set Breast Cancer
Zero
Mean
EM
KM
PA

Fig9. Breast cancer dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 4 breast cancer dataset analysis. breast cancer data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure10 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

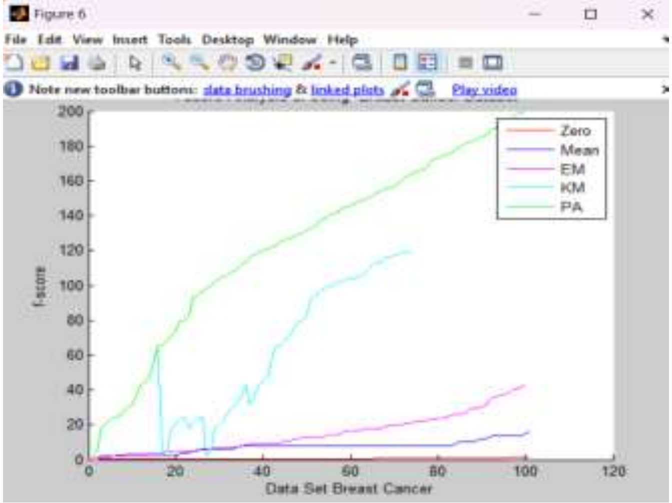
Figure 6
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
f-score
0
20
40
60
80
100
120
140
160
180
200
Data Set Breast Cancer
Zero
Mean
EM
KM
PA

Fig10. Breast cancer dataset f-score analysis between existing algorithm and proposed algorithm

(e) Ovarian cancer Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 5 Ovarian cancer dataset analysis. Ovarian cancer data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure11 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

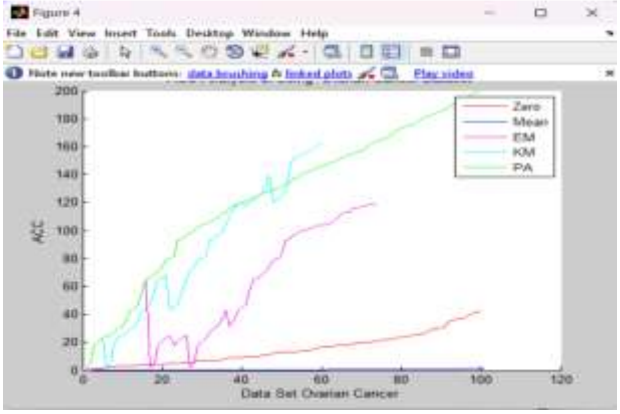

Fig11. Ovarian cancer dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM) using case 5 Ovarian cancer dataset analysis. Ovarian cancer data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure12 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

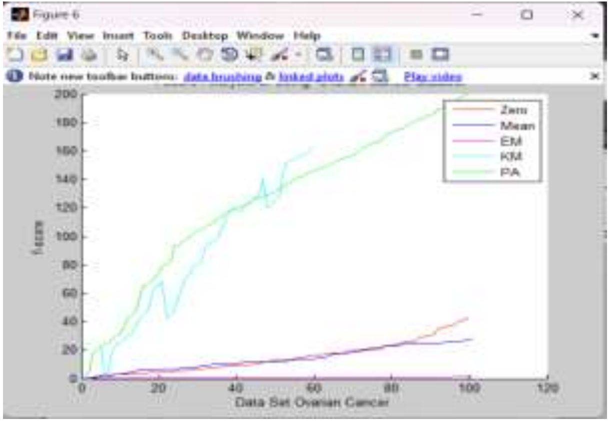

Fig12. Ovarian cancer dataset f-score analysis between existing algorithm and proposed algorithm

(f) Mice portion Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 6 Mice portion dataset analysis. Mice portion data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure13 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

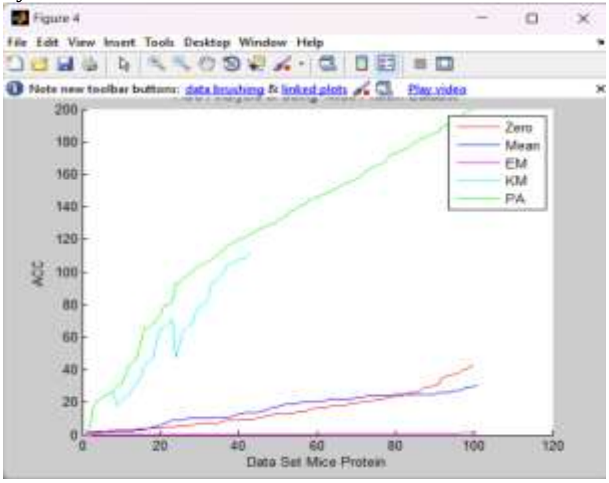

Fig13. Mice portion dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM) using case 6 Mice portion dataset analysis. Mice portion data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure14 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

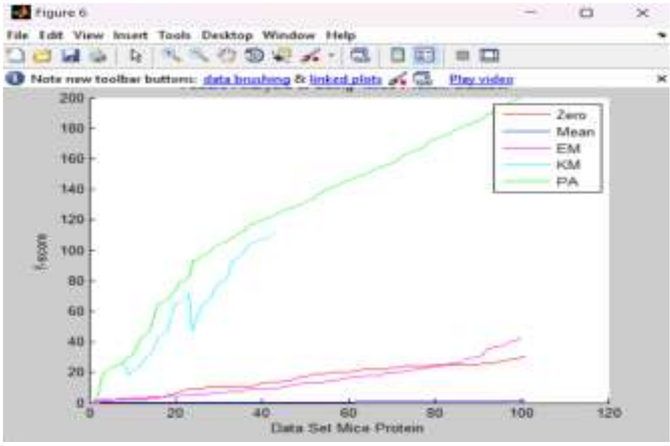

Fig14. Mice portion dataset f-score analysis between existing algorithm and proposed algorithm

(g) Pen digits Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 7 pen digits dataset analysis. pen digits data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure15 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality

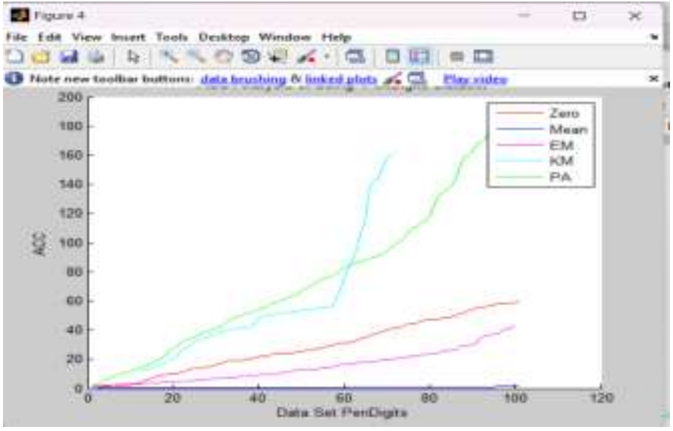

Fig15. Pen digits dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 7 pen digits dataset analysis. pen digits data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure16 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

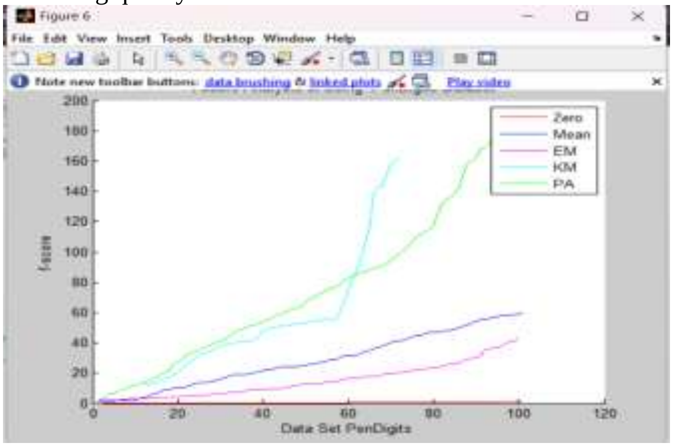

Fig16. Pen digits dataset f-score analysis between existing algorithm and proposed algorithm

(h) Avilla Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system

(PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 8 Avilla dataset analysis. Avilla data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure17 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

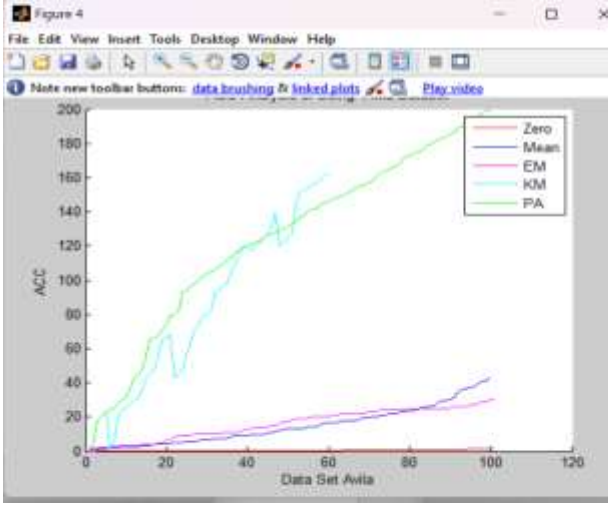

Fig17. Avilla dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 8 Avilla dataset analysis. Avilla data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure5.57 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

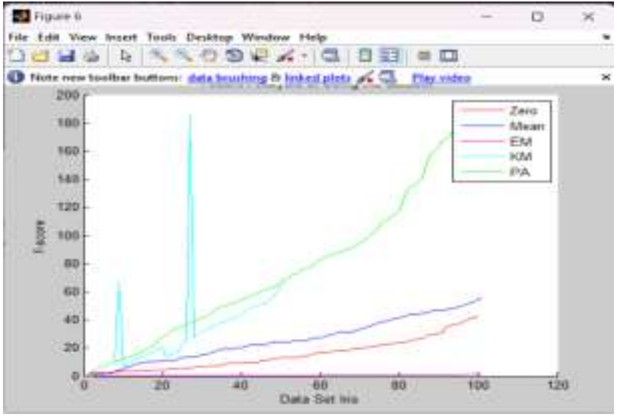

Fig18. Avilla dataset f-score analysis between existing algorithm and proposed algorithm

(i) Sensorlessdirve Dataset Result Analysis

(i)ACC (Accuracy of Clustering): Accuracy analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 9 sensorlessdirve dataset analysis. sensorlessdirve data analysis through proposed algorithm more accurate data, minimization error and also high accuracy but existing algorithms are more error and also low accuracy, in show figure19 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification accuracy. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

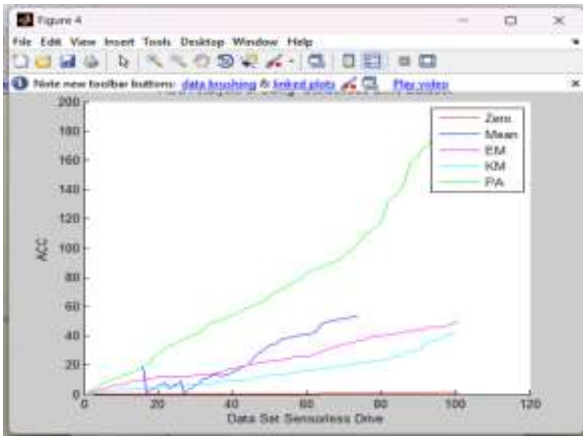

Fig19.sensorlessdirve dataset accuracy analysis between existing algorithm and proposed algorithm

(ii)F-score: F-score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset analysis. F-score analysis between proposed algorithm using smart information retrieval system (PASIRS) and existing algorithms (Zero Filling (ZF), Mean Filling (MF), Expectation Maximum (EM), k-mean clustering algorithm (KM)) using case 9 sensorlessdirve dataset analysis. sensorlessdirve data analysis through proposed algorithm correct prediction in clustering and also high correct prediction in clustering but existing algorithms are incorrectly prediction, in show figure20 below, the results obtained from the experiments: visualizations utilize visually appealing graphs. the resulting impact on classification normalized mutual information. effect of incomplete data handling: discuss how the specialized approach for handling incomplete data contributes to enhanced cluster quality and improved classification performance, the results effectively. impact of k-values examines the influence of different K-values on clustering quality.

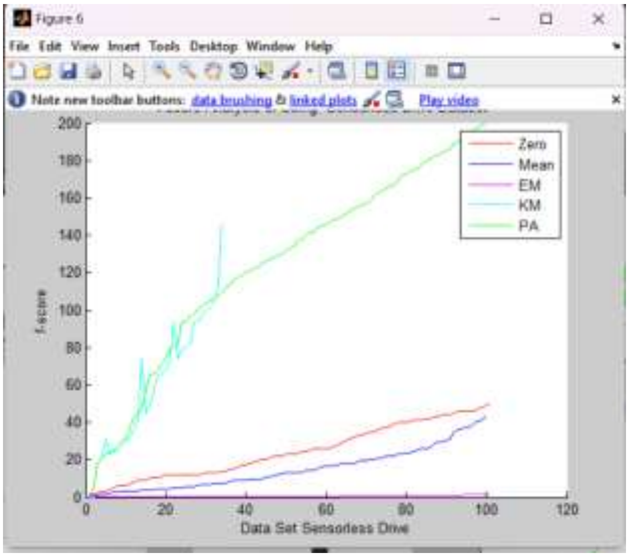

Fig20.sensorlessdirve dataset f-score analysis between existing algorithm and proposed algorithm

VI.CONCLUSION

The development of the economy, continuously produce all kinds of data mining algorithm, clustering algorithm is a kind of important analysis method of cluster analysis at the same time can get useful information from without prior knowledge, however, in most of the clustering algorithm, usually can only process data be inherent in certain numerical attributes and even is the classification of the single attribute, our proposed algorithm. Then, a discussion about the challenges of learning with big data and the corresponding possible solutions in recent researches was given. In addition, the connection of machine learning with modern signal processing technologies was analysed through studying several latest representative research papers in The algorithm solves the dilemma of over-dependence on the initial centre and local optimization, and the experimental results also prove the superiority of the algorithm, The results show that our algorithm makes it possible to estimate accurately the distribution of imprecisely known data. In particular, taking into account all the available information on the data uncertainty makes it possible to compute robust estimates of the parameters in presence of noisy attributes and corrupted labels. Finally, these segregated data have been used to select our initials centroids. Thus, we have successfully minimized the number of iterations. As the complexity of the traditional K-means clustering algorithm is directly related to the number of iterations. finding meaningful and useful results depends on the selection of the appropriate technique and proper tuning of the algorithm via the input parameters. In order to do this, one must understand the dataset in a domain specific context in order to be able to best evaluate the results from various approaches. One must also understand the various strengths, weaknesses, and biases of the potential clustering algorithm. our proposed approach outperformed compared to the existing methods. We believe this method could play a significant role for data-driven solutions in various real-world application domains

For future research, each ML model discussed can be stress tested on verification environments involving complex designs such as system on chips SoCs on both block and system level verification in each of the mentioned areas. Furthermore, incorporating ML technology to be a native feature in hardware verification methodologies, such as UVM, will significantly speed up coverage closure by converging to the planned coverage metrics faster. the proposed technique may also be implemented on other application datasets that also suffer from missing data. in future research. Also, most reviewed works show different domain datasets that are not as big as real-world datasets, which often contain a very large number of diverse features. Therefore, further work is needed to explore the possibilities of new methods of handling missing data in real world big data.

REFERENCE

[1] Jahwar, Alan Fuad, and Adnan Mohsin Abdulazeez. "Meta-heuristic algorithms for K-means clustering: A review." PalArch's Journal of Archaeology of Egypt/Egyptology 17, no. 7 (2020): 12002-12020.

[2] Ackerman, M. S. (2000). The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. Human–Computer Interaction15(2-3): 179-203.

[3] Abdulqader, D. M., Abdulazeez, A. M., & Zeebaree, D. Q. (2020). Machine Learning Supervised Algorithms of Gene Selection: A Review. Machine Learning, 62(03).

[4] Sulaiman, M. A. (2020). Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. Journal of Soft Computing and Data Mining, 1(2), 11-25.

[5] Chakrabarti, S. (2003). Mining the Web: Discovering knowledge from hypertext data. Morgan Kaufmann.

[6] Zebari, D. A., Zeebaree, D. Q., Saeed, J. N., Zebari, N. A., & Adel, A. Z. (2020). Image Steganography Based on Swarm Intelligence Algorithms: A Survey. people, 7(8), 9. Meta PJAEE, 17 (7) (2021) -Heuristic Algorithms for K-means Clustering: A Review 13

[7] Zeebaree, D. Q., Haron, H., Abdulazeez, A. M., &Zebari, D. A. (2019, April). Machine learning and Region Growing for Breast Cancer Segmentation. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 88-93). IEEE.

[9]Zhai, D.; Yu, J.; Gao, F.; Lei, Y.; Feng, D. K-means text clustering algorithm based on centers selection according to maximum distance. Appl. Res. Comput. 2014, 31, 713–719.

[10] Sun, J.; Liu, J.; Zhao, L. Clustering algorithm research. J. Softw. 2008, 19, 48–61.

[11] Li, X.; Yu, L.; Hang, L.; Tang, X. The parallel implementation and application of an improved k-means algorithm. J. Univ. Electron. Sci. Technol. China 2017, 46, 61–68.

[12] Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 0–892.

[13] Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 577–584.

[14] Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. 1998, 24, 283–304.

[15] Narayanan, B.N.; Djaneye-Boundjou, O.; Kebede, T.M. Performance analysis of machine learning and pattern recognition algorithms for Malware classification. In Proceedings of the 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), Dayton, OH, USA, 25–29 July 2016; pp. 338–342.

[16] Narayanan, B.N.; Hardie, R.C.; Kebede, T.M.; Sprague, M.J. Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities. Pattern Anal. Appl. 2019, 22, 559–571.

[17] Narayanan, B.N.; Hardie, R.C.; Kebede, T.M. Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses. J. Med. Imag. 2018, 5, 014504.

[18] https://www.researchgate.net/figure/The-Knowledge Discovery-in-Databases-KDD-process_fig1_274425359

[19]Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." J 2, no. 2 (2019): 226-235.

[20] http://ichrome.com/blogs/archives/221

[21] Wang, Siwei, Miaomiao Li, Ning Hu, En Zhu, Jingtao Hu, Xinwang Liu, and Jianping Yin. "K-means clustering with incomplete data." IEEE Access 7 (2019): 69162-69171.

[22]M. R. Rezaee, P. M. J. van der Zwet, B. P. E. Lelieveldt, R. J. van der Geest and J. H. C. Reiber, "A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering," IEEE Transactions on Image Processing, vol. 9, no. 7, July 2000, pp. 1238-1248.

[23]A. M. Fahim, A. M. Salem and F. A. Torkey, "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang UniversitySCIENCE A, vol. 7, no.10, 2006, pp. 1626–1633.

[24]A. Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," Bioinformatics,vol.24, no.11, 2008, pp. 1359 1366.

[25]M. J. Li, M. K. Ng, Y. Cheung and J. Z. Huang, "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, Nov. 2008, pp. 1519-1534.

[26]S. N. Sulaiman and N. A. Mat Isa, "Adaptive fuzzy-K-means clustering algorithm for image segmentation," IEEE Transactions on Consumer Electronics, vol. 56, no. 4, Nov 2010, pp. 2661-2668.

[27]U. Maulik and I. Saha, "Automatic Fuzzy Clustering Using Modified Differential Evolution for Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 9, Sep 2010, pp. 3503-3510.

[28]B. Yi, H. Qiao, F. Yang and C. Xu, "An Improved Initialization Center Algorithm for K-Means Clustering," 2010 International Conference on Computational Intelligence and Software Engineering, Wuhan, 2010, pp. 1-4.

[29] F. Bu, Z. Chen, Q. Zhang and X. Wang, "Incomplete Big Data Clustering Algorithm Using Feature Selection and Partial Distance," 2014 5th International Conference on Digital Home, Guangzhou, China, 2014, pp. 263-266.
doi: 10.1109/ICDH.2014.57

[30] V. T. N. Chau, N. H. Phung and V. T. N. Tran, "A robust and effective algorithmic framework for incomplete educational data clustering," 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 2015, pp. 65-70.doi: 10.1109/NICS.2015.7302224\

[31] S. Hua-Yan, L. Ye-Li, Z. Yun-Fei and H. Xu, "Accelerating EM Missing Data Filling Algorithm Based on the K-Means," 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC), Wuhan, China, 2018, pp. 401-406.
doi: 10.1109/ICNISC.2018.00088

[32] V. T. N. Chau, P. H. Loc and V. T. N. Tran, "A Robust Mean Shift-Based Approach to Effectively Clustering Incomplete Educational Data," 2015 International Conference on Advanced Computing and Applications (ACOMP), Ho Chi Minh City, Vietnam, 2015, pp. 12-19.doi: 10.1109/ACOMP.2015.14

[33] H. Liu, J. Wu, T. Liu, D. Tao and Y. Fu, "Spectral Ensemble Clustering via Weighted K-Means: Theoretical and Practical Evidence," in IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 5, pp. 1129-1143, 1 May 2017.doi: 10.1109/TKDE.2017.2650229

[34] V. T. Ngoc Chau, "A Robust Self-Organizing Approach to Effectively Clustering Incomplete Data," 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh City, Vietnam, 2015, pp. 150-155.doi: 10.1109/KSE.2015.11

[35] K. Honda, R. Nonoguchi, A. Notsu and H. Ichihashi, "PCA-guided k-Means clustering with incomplete data," 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan, 2011, pp. 1710-1714.doi: 10.1109/FUZZY.2011.6007312

[36] M. Vasuki and S. Revathy, "Efficient Handling of Incomplete basic Partitions by Spectral Greedy K-Means Consensus Clustering," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 299-305.doi: 10.1109/ICCMC48092.2020.ICCMC-00056

[37] A. Pugazhenthi and L. S. Kumar, "Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9276978.

[38]X. Liu et al., "Multiple Kernel kk-Means with Incomplete Kernels," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1191-1204, 1 May 2020, doi: 10.1109/TPAMI.2019.2892416.

[39] Bay, Stephen D., Dennis Kibler, Michael J. Pazzani, and Padhraic Smyth. "The UCI KDD archive of large data sets for data mining research and experimentation." ACM SIGKDD explorations newsletter 2, no. 2 (2000): 81-85.

[40] Yang, Yan-Ping. "A consistency contribution based bayesian network model for medical diagnosis." Journal of Biomedical Science and Engineering 3, no. 05 (2010): 488.