SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

EFFICIENT CLUSTERING ALGORITHM DESIGNED FOR MANAGING LARGE DATA SET

S Archana[1], Dr. Neeraj Sharma[2], Dr. Pradosh chandra Patnaik[3]
[1]Research Scholar, Dept. of Computer Science and Engineering
Sri Satya Sai University of Technology and Medical Sciences,
Sehore Bhopal-Indore Road, Madhya Pradesh, India.
[2]Research Guide, Dept. of Computer Science and Engineering
Sri Satya Sai University of Technology and Medical Sciences,
Sehore Bhopal-Indore Road, Madhya Pradesh, India.
[3]Research Co-Guide, Professor & Principal . Dept. of Computer Science and Engineering
Aurora's PG College (MCA) Hyderabad

ABSTRACT
Clustering methods are particularly well-suited for identifying classes in spatial databases. However, when applied to large spatial datasets, the following needs for clustering algorithms become apparent: minimal domain knowledge is required to calculate the input parameters, clusters of any shape can be discovered, and huge databases demand high efficiency. The well-known clustering methods are incapable of meeting all of these requirements. Clustering is the process of grouping similar data from a population data set so that data points belonging to the same group have a higher degree of similarity than data points belonging to other groups. Data clustering enables academics to reduce the dimension of complex problems, create spam filters, detect fraudulent or illegal behaviour, analyse documents, classify network traffic, and aid in marketing or sales analysis. The data is compared to the numerous clusters that exist. The cluster with the greatest degree of proximity is picked to store the data. By utilising this algorithm, we can decrease access time and make data retrieval easier. Additionally, an iterative process is used to create such clusters within the data node itself, facilitating data access via parallelization.

Keywords : Clustering Algorithm, Data clustering, Spatial database.

INTRODUCTION
Data clustering is a type of data analysis that divides unlabeled data into distinct groups based on a measure of similarity. Each group is referred to as a "cluster" since the data contained inside it are comparable but distinct from the data contained within other clusters. Clustering is most frequently employed in areas that require multivariate data processing. Cluster analysis has played a key role in recent years in a variety of sectors, including engineering, life and medical sciences, earth sciences, and economics. The fundamental difficulty in clustering analysis is to precisely calculate the approximate number of clusters, as this number has a significant influence on the clustering outcomes. Clustering techniques are broadly categorised into two types: partitional and hierarchical clustering. Hierarchical clustering organises data points into a hierarchical tree structure based on their similarity. It is blind to the shape and size of the clusters generated. Additionally, because this clustering assigns a single cluster to a single

DOI: 10.5281/zenodo.7259992

data point at a time, the cluster structure remains static. In data, partitioning clustering analyses the dataset and groups data points based on their similarity. The partitioning clustering algorithm seeks to optimise a global criterion by minimising similarity between components inside a cluster and increasing dissimilarity between clusters. While both of these methods have demonstrated their utility and performance in a variety of domains, they both have certain major limitations. These algorithms are only effective if the number of clusters existing in the datasets is known in advance. Due to the fact that different datasets, particularly those from real-world applications, exhibit a wide variety of patterns, cluster analysts lack knowledge about the number of acceptable clusters present in the dataset. As a result, these techniques requiring the cluster number as an onset parameter are ineffective. Due to the fact that the majority of real-world datasets lack class labels, there are no explicit criteria for conducting clustering research. It is seen as a significant limitation of the dataset, making it difficult to discover an appropriate number of clusters. As a result, establishing the optimal number of clusters in a data set has emerged as a critical research subject for overcoming these constraints.

Following a time of handling data buildup difficulties, the topic has shifted to how to proceed with these massive amounts of data. Researchers and experts concur that one of the most vital topics in computer science nowadays is Big Data. For instance: Social networks such as Facebook and Twitter have billions of users and generate gigabytes of data every minute, retail locations constantly collect information about their customers, and You Tube has one billion unique users and delivers hundreds of hours of cinematic video each hour, while its substance ID service analyses over 400 years of video. Big data is a term that refers to massive amounts of information in the form of a mixture of relevant and irrelevant data for a specific application. By doing knowledge discovery in databases, data mining identifies intriguing patterns in large amounts of data. It depicts the Big data architecture in its simplest form.

Massive Data is inextricably linked to the transformation of unstructured, valuable, faulty, and difficult data into useable data. However, it becomes difficult to maintain a significant volume of data and information each day from numerous assets and administrations that were not accessible to human space just a few decades ago. To manage this torrential flow of data, it is critical to leverage extraordinary information discovery assets. Clustering is one of them. It is an approach in which data is split into groups so that items within each group share a greater degree of similarity than other articles in different collections do. Data clustering is a highly effective approach in a variety of areas of software engineering and related fields. While data mining can be considered the fundamental source of clustering, it is widely used in a variety of sectors of learning, for example, machine learning, bio informatics, networking, energy engineering, and pattern recognition, and a substantial amount of study has been conducted in these areas. From the start, researchers have managed clustering algorithms to minimise their complexity and computing cost, while increasing their adaptability and speed. Clustering is an unsupervised learning task in which one attempts to recognise a limited set of classifications referred to as groups to represent the data. Clustering is also defined as the accumulation or occurrence of identical or comparable components. Clustering separates the population data set into several clusters based on the degree of homogeneity among the various classes. Different

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

clustering algorithms employ a variety of metrics when clustering data.

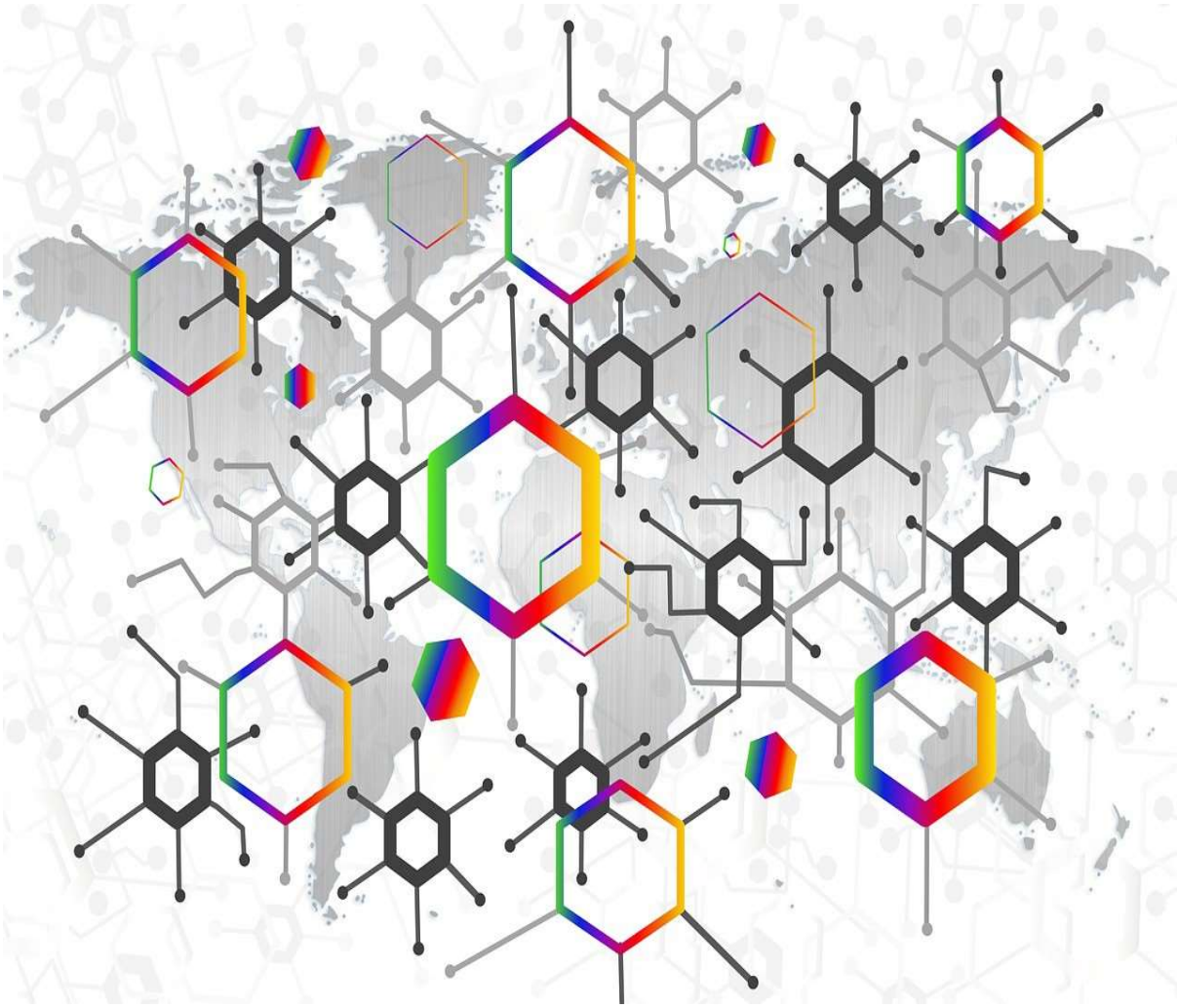

Figure 1 Cluster Analysis

Additionally, numerous terabytes of data are generated each second. Handling this data, analysing it, and getting usable information from it is a significant problem in and of itself. To manage this expansion, existing sequential algorithms and tools must be parallelized or enhanced. Additionally, new tools for analysing them are required. It requires a significant amount of computational power and distributed storage. Existing parallel processing architectures suffer from a number of limitations, including node failure, fault tolerance, network congestion, and data loss. Clustering is an efficient technique for managing large amounts of data. When dealing with enormous datasets, the standard techniques become inefficient. The primary reason for this is that the majority of algorithms are incapable of dealing with extremely large data sets and dimensions. Additionally, they are meant to work exclusively with structured data. Graph data has become ubiquitous. The majority of data comes via graph-based networks in which the nodes represent the data points and the edges reflect the relationships between the data points. Every second, data from multiple streams such as log files, social networks, and YouTube is ingested. As a result, existing algorithms should be fine-tuned to produce high-quality clusters regardless of the data size. This expanding

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992
volume and variety of data on the web inspires us to develop scalable and effective clustering algorithms for Big Data.

LITERATURE REVIEW

Khan et al., (2020) Clustering is a data mining technique used to discover interesting patterns within a given dataset. Using the K-Means algorithm, a huge dataset is partitioned into clusters of smaller sets of related data. When employing the K-Means clustering algorithm, initial centroids are required as input parameters. There are numerous ways for selecting initial centroids from a dataset's real sample data points. These methods are frequently implemented using intelligent agents, which are quite popular in distributed networks due to their little impact on network traffic. Additionally, they are capable of overcoming network latency, operating in heterogeneous environments, and exhibiting fault-tolerant behaviour. In this research, a Multi Agent System (MAS) is used to generate initial centroids from actual sample data points. This multiagent system consists of four KMeans clustering agents that generate initial centroids using a variety of approaches, including Range, Random number, Outlier, and Inlier. The modified initial condition enables the iterative algorithm to reach a "better" local minimum. The approach can be used with a broad variety of clustering algorithms for both discrete and continuous data. The method is scalable and can be used in conjunction with a scalable clustering algorithm to solve data mining's large-scale clustering difficulties.

Singh et al. (2017) employ the Swarm Optimization technique to achieve efficient clustering. The clustering procedure and cluster head selection are initiated by the BS. The cluster head is chosen based on the relay node that is the next hop from the BS. Member nodes are assigned to their corresponding CHs using the ceiling function, and the range is demonstrated using an example. The cluster head with the least leftover energy is intended to extend the network's lifetime. To extend the lifetime of the least residual energy node, a first fitness function has been devised. The node's lifetime is a function of the remaining energy and the total energy consumed by a CH per round. As a result, the fitness function is precisely proportional to the CH lifespan. The second fitness function is defined in terms of the average distance between sensor nodes and the CH. When the average distance is the shortest possible, the longevity is increased. The fitness function does not have a direct relationship with the average distance. The first and second fitness functions work in tandem to prolong the life of CH. When compared to LEACH and HEED, the proposed algorithm exhibits significant improvements. However, the energy of nodes was not taken into account when choosing the CH. For network longevity, the proposed protocol can be compared to a recent protocol. The author began by calculating the distances between each pair of data points; next he sought out similar data points; and finally, he constructed initial centroids based on these discovered data points. Diverse initial centroids produce disparate results. The more initial centroids that are congruent with the data distribution, the more accurate clustering can be obtained.

Jirong Gu et al., (2019) investigate the K-Means clustering algorithm and one of its upgrades. Clustering is the process of classifying items into distinct groups, or more precisely, the partitioning of a data collection into subsets (clusters), with the data in each subset (ideally) sharing some common attribute - frequently proximity according to some defined distance

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992
metric. A prominent clustering technique is based on K-Means, which partitions the data into K clusters. The number of clusters is predefined in this procedure, and the strategy is extremely dependent on the initial identification of pieces that adequately represent the clusters. If the sample data set is too large, it may cause the cluster members to become unstable. Another issue is seed point selection, as clustering results are always dependent on the original seed points and partitions. To address this issue, a refining beginning points algorithm is offered; it can significantly reduce execution time and enhance solutions for huge data sets by specifying initial conditions refinement. Following that, it applies this strategy to the well-known K-Means clustering algorithm, demonstrating that revised initial starting points do actually result in improved solutions. The time necessary to conduct the refinement is significantly less than the time required to cluster the entire database.

RESEARCHMETHODOLOGY

The current research suggests that by locating the initial cluster centres, the performance of K-Means clustering can be improved. A variety of techniques is used to determine the first cluster centroid. In this approach, the initial centroids are computed using spectral biclustering, which entails data normalisation, bistochastization, seeded region growth, and clustering. Two parts comprise the proposed semi-unsupervised centroid selection method: data ranking and data combination selection. This strategy outperforms the usual K-Means algorithm and other established clustering techniques.

Description of the concept The term "concept description" refers to a type of summary description that seeks to condense data by comparing it to other things or by comparing it to other concepts. By summarising the data, you can gain a general understanding of it. The simplest concept is the use of statistics in the traditional approach to calculate the various data items in a database, such as total, mean, variance, and so on, or the use of OL "(On Line Processing, online analytical processing) to perform multi-dimensional query and calculation of data. Analysis of Correlation Correlation analysis revealed that a substantial number of data elements from the collection of contacts had an interesting link or correlation. With a significant number of people collecting and storing data on a continual basis and many people in the sector using their database for mining association rules, the subject becomes increasingly interesting. Correlations discovered between data from a vast variety of company services could aid in the formulation of numerous business decisions.

Prediction and Classification Classification and prediction are two data of data analysis that can be used to extract models characterising significant data classes or forecasting future trends. Classification and prediction are applicable to a wide variety of applications; for instance, you can develop a classification model. On the bank's loan clients to classify, in order to mitigate loan risk; and also on the factory machines to classify, in order to forecast the occurrence of machine failure, through the construction of a classification model. Analysis of Clusters According to the principle of maximum similarity and minimum between-cluster similarity, the same class of items with a high degree of resemblance to other classes of things is extremely similar. Each cluster formation is unique. Class can be thought of as an object class from which rules can be exported. Clustering also makes it simple to observe the organization's contents in

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

a hierarchical framework and group related events together. Outlier Analysis Databases may contain data items that exhibit inconsistent behaviour with the data or with the model. These data objects are considered outliers.

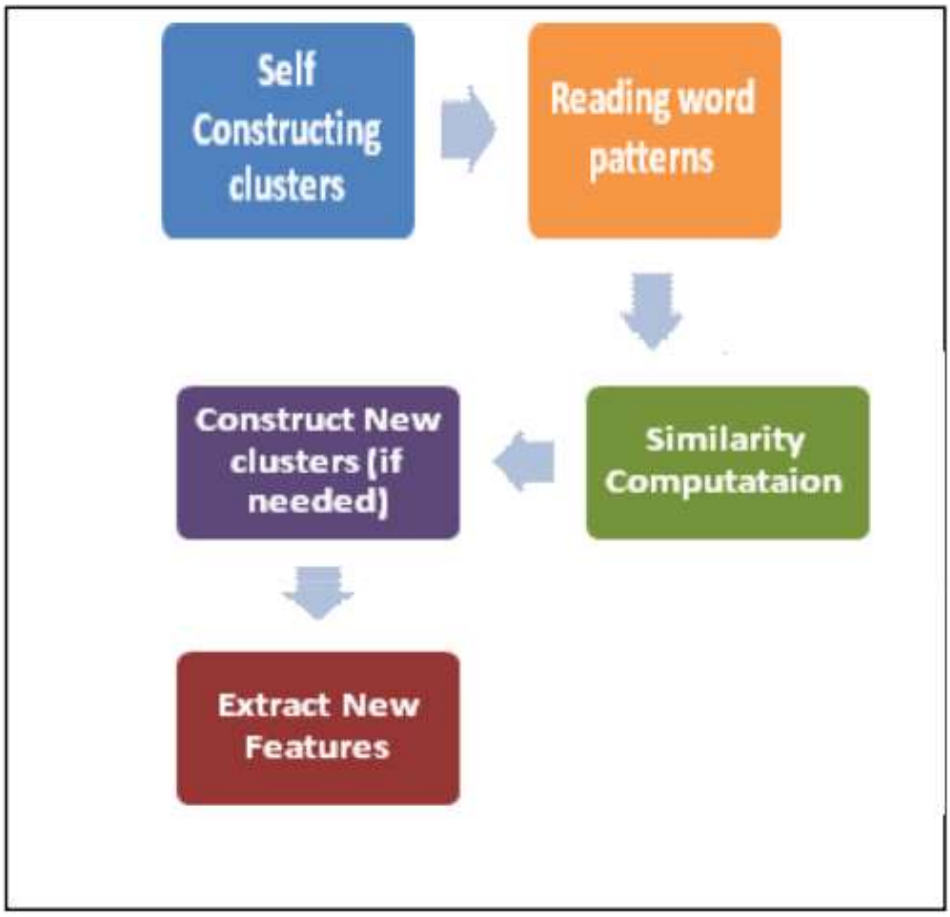

Figure 2 Framework

Numerous data mining methods make an effort to mitigate the effect of outliers, or rows. Additionally to them, in some situations, they may be an isolated point of a critical message. For instance, solitary points may signal fraud in fraud detection. Analyze Time Series In time series analysis, the value of the data attribute changes over time. These data are normally obtained at equal time intervals, but cannot be obtained at equal time intervals. Through time, a series map can be used to visualise time series data. Time series analysis consists of three fundamental functions: To begin, excavation in the dig mode, that is, by examining time series of historical patterns in order to ascertain the behavioural aspects of events. Second, trend analysis, which is the process of anticipating future values using previous data for time series. Thirdly, similarity search, which makes use of distance metrics to determine the similarity of two time series.

By definition, data clustering is an exploratory and descriptive data analysis technique that has attracted widespread interest in fields such as statistics, data mining, and pattern recognition. It is an exploratory technique for examining multivariate data sets that may comprise a variety of various data kinds. These data sets differ in size in terms of a variety of objects and dimensions,

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

or they comprise a variety of various data kinds. Without a doubt, data clustering is a critical component of data mining, which focuses on enormous data sets with an unknown underlying structure. The purpose of this study is to provide an overview of specific components of the cluster analysis approach. So-called partitioning-based clustering methods are adaptable clustering techniques based on recurrent data point relocation between groups. A clustering criterion is used to determine the quality of the solutions. Iterative relocation techniques reduce the value of the criterion function with each iteration until convergence. By modifying the clustering criterion, it is feasible to develop robust clustering systems that are less sensitive to incorrect and missing data.

DISCUSSION & RESULTS

We assessed the proposed approach using data sets from the University of California, Irvine's machine learning repository. We compared the clustering results obtained using k-means, PCA+k-means with random initialization, and the proposed algorithm's initial centres. Due to the information contained in the data structure, this function is able to minimise the amount of distance calculations required to assign each data object to the nearest cluster, which results in a quicker enhanced k-means algorithm than the regular k-means algorithm.

Algorithm 1 OptiGrid algorithm Given: number of projections k

Given: number of projections k, number of cutting planes q, min cutting quality min_c_q, data set X
Compute a set of projections P = {P1, ..., Pk}
Project the dataset X wrt the projections {P1(X), ..., Pk(X)}
Best Cuts ← ∅, Cut ← ∅
for i ∈ 1..k do Cut ← Compute Cuts(Pi(X)) for c in Cut do if Cut Score(c)> min_c_q then
Best Cuts. append(c)
End
end

if Best Cuts.isEmpty()
then return X as a cluster BestCuts ← KeepQBestCuts(BestCuts,q)
Build the grid for the q cutting planes Assign the examples in X to the cells of the grid Determine the dense cells of the grid and add them to the set of clusters C
foreach cluster cl ∈ C **do** apply OptiGrid to cl
end

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

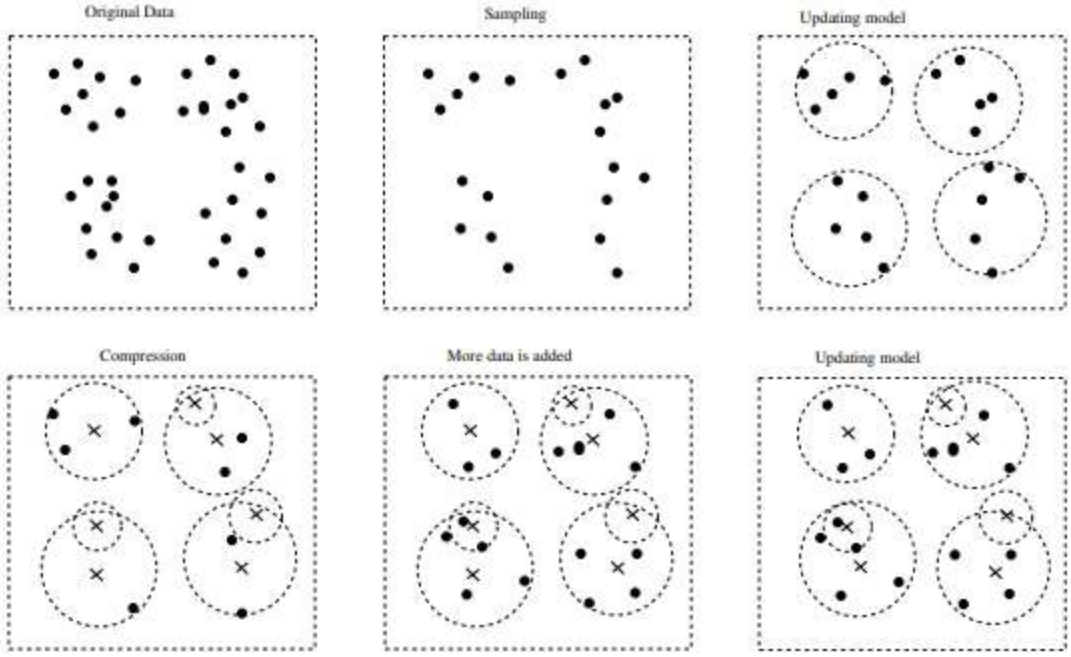

Figure 3: Scalable K Means

Algorithm 2 Mini Batch K-Means algorithm

Given: k, mini-batch size b, iterations t, data set X
Initialize each c ∈ C with an x picked randomly from X
v ← 0

for i ← 1 to t do
M ← b examples picked randomly from X
for x ∈ M do d[x] ← f(C,x)
end
for x ∈ M do
c ← d[x] v[c] ← v[c] + 1 η ← 1 v[c] c ← (1-η)c+ηx
End
End

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

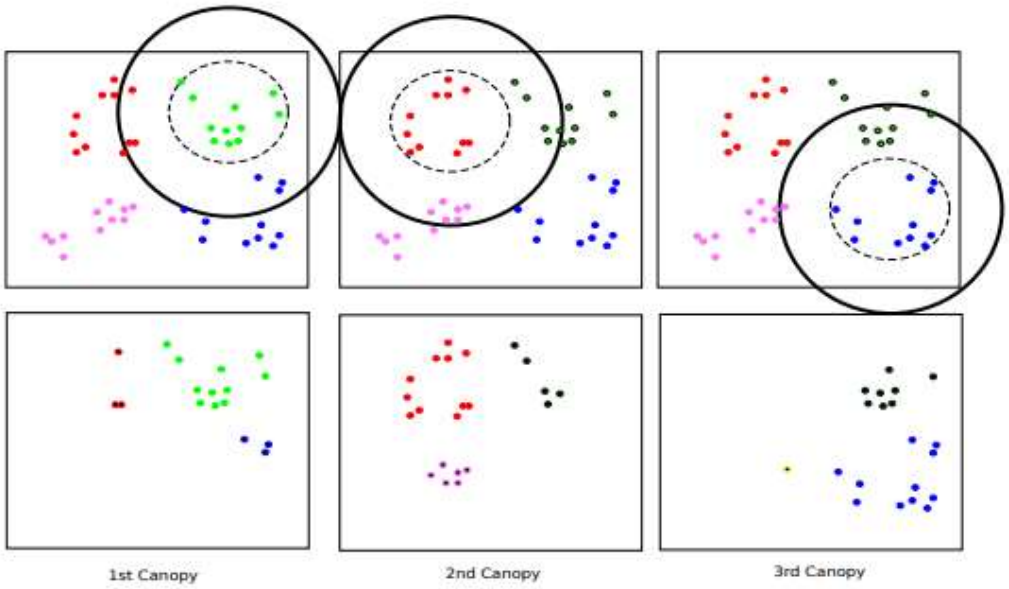

Figure 4:Canopy Clustering

The following data sets are used to evaluate the proposed method's accuracy and efficiency.

Table 1 Dataset Description

Data Set	Samples	Dimensions	No. of clusters
Iris	140	3	2
Wine	168	12	2
Glass	204	8	5
Imgseg	2320	18	6

The results in Figure 3 illustrate that the proposed method outperforms existing methods in terms of cluster accuracy when using k-means and modified k-means. The clustering results for random initial centre are averaged over seven runs, as each run produces unique results. It demonstrates that the suggested algorithm outperforms the random initialization algorithm significantly.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7259992

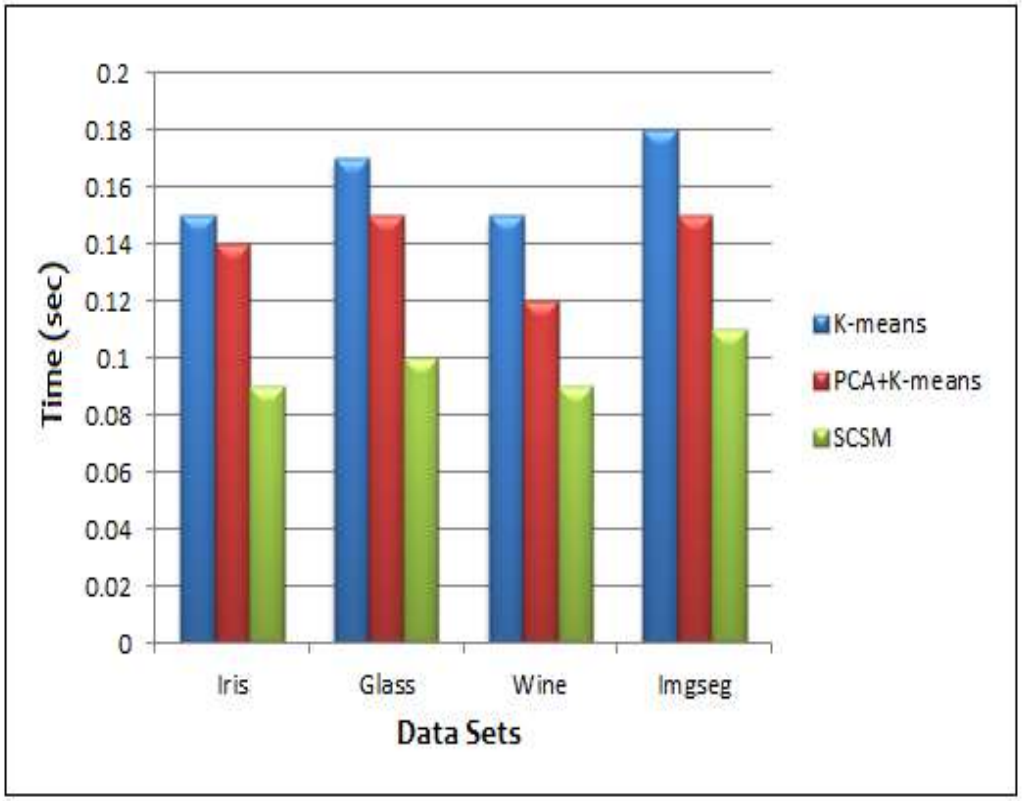

Figure 3 Execution time results on data sets

The experimental data demonstrates the efficacy of our technique. This could be because the first cluster centres formed by the proposed algorithm are very close to the optimal solution and it also discovers clusters in low-dimensional space, so overcoming the curse of dimensionality. To shed light on future paths for algorithm development and to aid in the selection of algorithms for big data, we suggested a framework for categorising a variety of clustering algorithms. The classifying framework is built from a theoretical perspective with the goal of automatically recommending the most appropriate algorithm(s) to network professionals while concealing all technical information unrelated to the application. Thus, subsequent clustering algorithms could be added into the framework based on the criteria and attributes proposed. Additionally, the best representative clustering methods for each category were empirically examined across a broad range of assessment criteria and traffic datasets.

CONCLUSION

SCSM is applied to original data prior to clustering with the primary goal of obtaining reliable findings. However, the clustering results are dependent on how the centroid is initialised. In this article, we propose a novel approach for initialising the centroid and reducing the dimension using principal component analysis to improve the accuracy of the cluster results, as well as a modified version of the standard k-means algorithm to increase efficiency by reducing the algorithm's computational complexity. The experiment findings demonstrate a significant improvement in clustering performance and accuracy by lowering the dimension and selecting the initial centroid using SCSM. While the proposed method outperformed random

DOI: 10.5281/zenodo.7259992

initialization methods in all circumstances, there is a limitation, namely the number of clusters (k) required as input. Future study should focus on developing some statistical methods for calculating the value of k that are dependent on the data distribution. We intend to use this technology to microarray datasets in the future.

REFERENCES:

1) Singh, S., Malik, A. and Kumar, R., 2017. 'Energy Efficient Heterogeneous DEEC Protocol for Enhancing Lifetime in WSNs', Engineering Science and Technology, an International Journal- Elsevier, Vol- 20(1), ISSN: 22150986, pp.345-353.

2) Khan D.M and Mohamudally N, "A multiagent system (MAS) for the generation of initial centroids for k-means clustering data mining algorithm based on actual sample datapoints", 2nd International Conference on Software Engineering and Data Mining (SEDM), Pp. 495–500, 2020.

3) Jirong Gu, Jieming Zhou and Xianwei Chen, "An Enhancement of K-means Clustering Algorithm", International Conference on Business Intelligence and Financial Engineering (BIFE '09), Pp. 237–240, 2019

4) The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Hey, T. , Tansley, S. and Tolle, K.. Microsoft Corporation, October 2009. ISBN 978-0- 9825442-0-4.

5) Demchenko, Y., Membrey, P., Grosso, C. de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.

6) A. Abbasi and M. Younis, ''A survey on clustering algorithms for wireless sensor networks,'' Comput. Commun., vol. 30, nos. 14–15, pp. 2826–2841, Oct. 2007.

7) C. Aggarwal and C. Zhai, ''A survey of text clustering algorithms,'' in Mining Text Data. New York, NY, USA: Springer-Verlag, 2012, pp. 77–128.

8) Almalawi, Z. Tari, A. Fahad, and I. Khalil, ''A framework for improving the accuracy of unsupervised intrusion detection for SCADA systems,'' in Proc. 12th IEEE Int. Conf. Trust, Security Privacy Comput. Commun. (TrustCom), Jul. 2013, pp. 292–301.

9) Almalawi, Z. Tari, I. Khalil, and A. Fahad, ''SCADAVT-A framework for SCADA security testbed based on virtualization technology,'' in Proc. IEEE 38th Conf. Local Comput. Netw. (LCN), Oct. 2013, pp. 639–646.

10) A. Fahad, Z. Tari, A. Almalawi, A. Goscinski, I. Khalil, and A. Mahmood, ''PPFSCADA: Privacy preserving framework for SCADA data publishing,'' Future Generat. Comput. Syst., vol. 37, pp. 496–511, Jul. 2014.

11) S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, ''Labelled data collection for anomaly detection in wireless sensor networks,'' in Proc. 6th Int. Conf. Intell. Sensors, Sensor Netw. Inform. Process. (ISSNIP), Dec. 2010, pp. 269–274.