Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 6 August 2022

Air Pollution Prediction on Hybrid of Agglomerative and K-Means Clustering Algorithms

S.Suganya , Dr.T.Meyyappan

Department of computer Science
Alagappa University, karaikudi, India
Suganyasudhakar04@gmail.com

Abstract— **The current state of microarray technology allows for the simultaneous observation of a large number of genes' expression levels at many time points. Several clustering techniques have been applied to the learn of NO2 and SO2 data using microarrays. The plan of this do research is to develop the hybridization of the k-means technique and agglomerative technique. K-means is the for the most part extensively used centroid-based clustering algorithm. K-means clustering is one of the most popular clustering methods, and it's frequently the first thing people do when they're working on a clustering problem to gain sense of the dataset's structure. K-means outperforms than other clustering techniques such as density-based and expectation-maximization. Agglomerative, according to a few users, is a very simple and straightforward algorithm. Although the K-mean approach takes longer to execute, its performance is superior to that of the hierarchical algorithm. When compared to the k-mean approach, the hierarchical algorithm produces higher quality results. In universal, the k-mean technique is better for large datasets, whereas the hierarchical algorithm is better for small datasets. Since this reason this proposed work have been developed the hybridization of k-means and agglomerative clustering technique for make suitable for all kinds of dataset, and to produce higher accuracy. Finally carried experimentation of hybridization algorithms obtained better accuracy.**

Keywords— Clustering, Prediction, K-means, Air Pollution, Agglomerative, Hybridization.

I.INTRODUCTION

Pick k items at random as of the dataset to serve as cluster representatives. Using a Euclidean distance derived by a similarity function, link each remaining item in the collection to the closest cluster representative. Recalculate the representatives of the new clusters. Genetic information is stored in information units known as genes in every living species. The genome is a term that refers to an organism's entire set of genes. Microarray studies must be designed in order enable other researchers to acquire, interpret, save, and exchange genetic data. Researchers and medical specialists can use microarray technology to assess the expression of thousands of genes in a hankie model in a single experiment, which aids in the identification of the disease gene. The data collected by the studies, however, cannot be evaluated manually due to their large bulk and high complexity. Microarrays, in this scenario, have the capability of simultaneously observing the expression profiles of thousands of genes under various experimental settings.

Knowledge Discovery in Databases is how Data Mining is typically described (KDD). It refers to the significant ability to identify relevant, original, possibly useful, and finally understanding data patterns. Data mining is the development of generating patterns in a certain representational format, such as organization rules, resolution plants, cluster or regression models. Cluster analysis has become a critical component of gene expression analysis. Gene clustering is the process of group genes with similar patterns. For the investigation of microarray gene expression data, a number of clustering techniques have been used. As a result, the primary purpose of this research is to examine a more effective clustering method for microarray gene expression information.

II. LITERATURE REVIEW

Clustering is one of the well-known Data mining approaches for finding meaningful patterns from data in a large database (Fayyad, 1996), according to **Huda Hamden Ali** et al [1]. However, because data sets in data mining typically comprise categorical values, working primarily with numeric values limits its application in data mining. Clustering is useful in a variety of fields, including economics (particularly market research), document classification, pattern recognition, spatial data analysis, and image processing. K-means is a technique for detecting unknown attacks in intrusion detection systems. The use of data mining techniques has long been acknowledged to give support to in the detection of network intrusions.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 6 August 2022

Based on their basic approach, **Dr. Aishwarya Batra** et al [2] advised that the K-Means be evaluated and assessed. Based on their performance, the best algorithm in each category was discovered. The input data points are generated in one of two ways: with a normal distribution or with a uniform distribution. The most often used partition-based clustering algorithm is K-means. However, it is computationally expensive, and the quality of the generated clusters is highly dependent on the original centroid selection and data dimension. Several strategies for increasing the performance of the k-means clustering algorithmhavebeenproposedintheliterature

.

Santosh Nirmal et al [3] have offered an alternative to K-Means. According to the author, k-means achieves a higher accuracy easiness than k-medoids. The research is based on the application of disinterest measurements in two clustering methods. They proposed that for better results, they use the Euclidean and Manhattan distance metrics.

This proposed method, which is based on their centroids, combines the K-Means and agglomerative clustering techniques to create a new hybrid methodology model. Clustering is the most important unsupervised learning problem; like with any other problem of this type, it can be loosely defined as "the act of grouping items into groups whose members are familiar in some manner." Because of the ever-increasing volume of data and the exponential increase in computer processing rates, clustering has become widely used, and its relevance has expanded accordingly. Clustering is significant because it has a wide range of applications in areas such as education, business, agriculture, machine learning, pattern recognition, and economics. Modern artificial intelligence and pattern recognition systems have both benefited from this technique.

This proposed method integrated the K-Means and Agglomerative clustering technique. This hybridization method has been applied on the air pollution dataset to predict the early air pollution to take the necessary steps to avoid the pollution and to control the air pollution.

METHODOLOGY

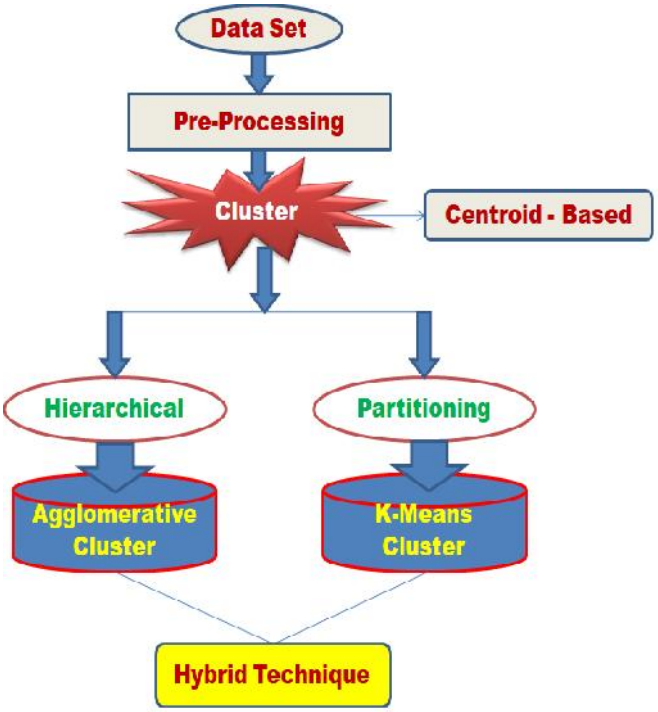

Fig.1. Architecture of the Proposed Work

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 6 August 2022

A. Data Set Used

The data set for the proposed study endeavor was obtained from the Government of India, National Centre for Medium Range Weather Forecasting, Ministry of Earth Science (NCMRWF), Delhi. This data set has a total of seven attributes, all of which were acquired in 2011. Only two attributes were chosen for this study from this data set. Air Quality (AQ) values of NO2, SO2 are examples of Air Quality (AQ) metrics. The remaining characteristics have a large number of null values. However, NO2 and SO2 are only found in this data set. As a result, NO2 and SO2 levels in the air are higher. For living things, NO2 and So2 are more hazardous.

B. ClusteringTechniques

Clustering techniques divide a large quantity of information into smaller groups called clusters. A cluster is a collection of statistics elements. Except for items in different clusters, they are comparable to one another in the same cluster. Clustering analysis' major goal is to partition a big quantity of data into homogeneous and distinct groups while also reducing data complexity. The clustering process is depicted in the diagram below.

Clustering technique have complete which is helpful to recognize the quality power, quality direction, cell forms and also subtypes of cells. Co-communicated behavior which mean qualities with same communication designs are clustered together with the same cell capacities. In this kind of come within reach of, it may helpful to recognize the process of many genes for which in order has not been previously obtainable. Clustering technique is helpful tool to determining structure and associate patterns in gene appearance data.

Data from big data sets can be organized, modeled, categorized, and compressed using clustering methods. The different cluster algorithms used in microarray gene expression data are discuss in this section.

B.a. Hierarchical Clustering

Clustering is an important tool in microarray analysis. Hierarchical clustering groups data objects into cluster and then find larger clusters from those clusters, resulting in a cluster pecking order. It creates a cluster tree out of the data objects. Agglomerative and disruptive various tiered clustering techniques are two types of hierarchical clustering algorithms. This classification is based on whether the gradual degradation is formed from the ground up or from the top down.

Agglomerative Method

The agglomerative approach of clustering is additional prevalent than the disruptive method. The proper application of this technology aids in the processing of gene appearance data in microarray technique. This technique begins with a bottom-up approach. Each object forms a separate group in this procedure, which is then combined with objects or groups that are close to one another. This technique is repeat until only one cluster remains [8]. The evolution of agglomerative clustering is depicted in the diagram below.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 6 August 2022

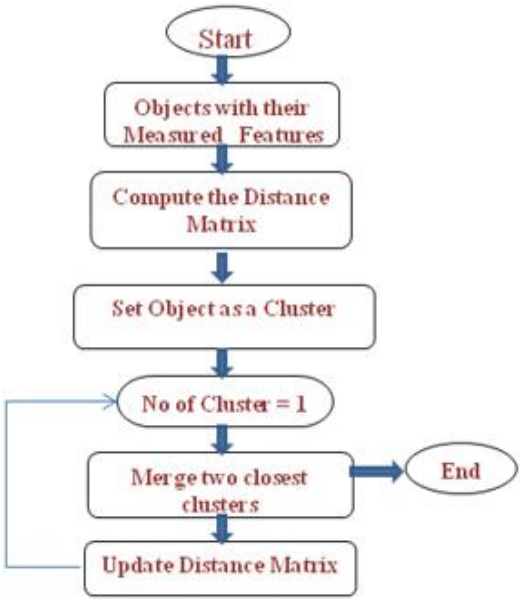

Fig.2: Flowchart of agglomerative clustering algorithm

B.b. Partition Based Clusterig

The Partition base Clustering algorithm reduces the size of a given clustering calculate by relocating data points among groups iteratively until the best partition is found. The K-means and K-Medoids methods are two partitioning approaches. The K-Means estimates the distance between groups using their centroids, while the K-Medoids uses their centre of gravity to calculate the distance between groups.

K-Means Clustering Method

Because of its convenience and efficiency, K-Means is a fairly successful method in partition-based clustering algorithms, as well as the most widely used among all clustering algorithms. This algorithm's goal is to find groups in the data based on the number of groups that correspond to the variable K. This approach works in an iterative fashion to assign each data object to one of K groups based on the features available. Feature similarity is used to group information objects. The function of the K-means clustering algorithm is depicted in the diagram below.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 6 August 2022

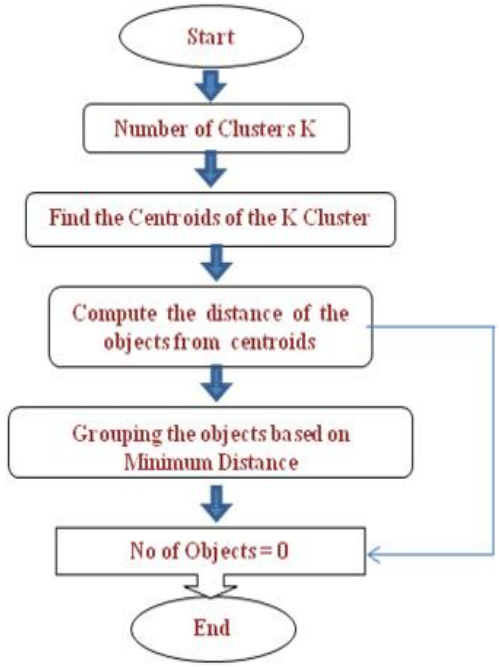

Fig.3: Flowchart of K-means clustering algorithm

II. PROPOSED MODIFIED CLUSTERING TECHNIQUE

It is proposed that a Modified Clustering Algorithm be used. The calculation's main purpose is to generate two easy data structures that can be used in the next focus to keep track of cluster titles as well as the distance in between data piece and the next group during each cycle. If the deliberate distance is less than or equal to the gap between the current data protest as well as the previous cluster focus, the data question stands inside the cluster to which it was assigned in the prior emphasis. As a result, there's no compelling reason to inference the division from opening data protest the previous k-1 clustering focus, which keep the getting to time to the k-1 clustering focuses at a minimum to the k-1 cluster focuses. Otherwise, it should calculate the separation from the current data query for each of the k cluster focuses and distinguish the neighboring cluster focus. This point is circulated to the closest cluster focus in this manner, which then records the separation to its interior independently. Every cycle, A small number of data focuses remain in the initial cluster, imply that a little portions of the data focuses will not be projected, saving time in influential the division and boost the calculation's success.

TABLE I. COMPARATIVE OF PROPOSED CLUSTERIG ALGORITHMS

Factors	Agglomerative	K-means	Hybrid Algorithm
Execution Time(ms)	1430	1279	997
Error Rate	0.7385	0.7465	0.6564
No of Clusters	6	6	6

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 6 August 2022

In the table above, the results of the traditional Agglomerative, K-means clustering, and Hybrid techniques are shown. Because the data is reassigned multiple times in each emphasis, the traditional calculation is extremely difficult to do. It reduces the accuracy of traditional clustering calculations. For six numbers of clusters, K-Means technique acquired execution time is 1279 mille seconds, error rate is 0.7465. Similarly, the Agglomerative approach achieved execution time of 1430 milliseconds with an error rate of 0.7385 for six clusters. However, for six clusters, the proposed Hybridized approach achieved an execution time of 997 and an error rate of 0.6564. As a result of this outcome, the hybrid strategy outperformed the other two techniques.

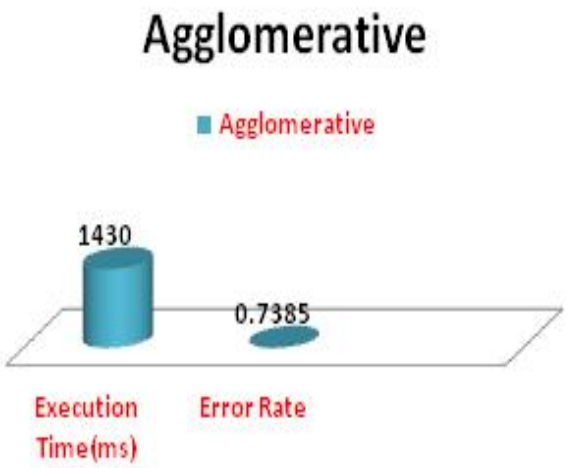

Fig.4: Performance of Agglomerative Algorithm

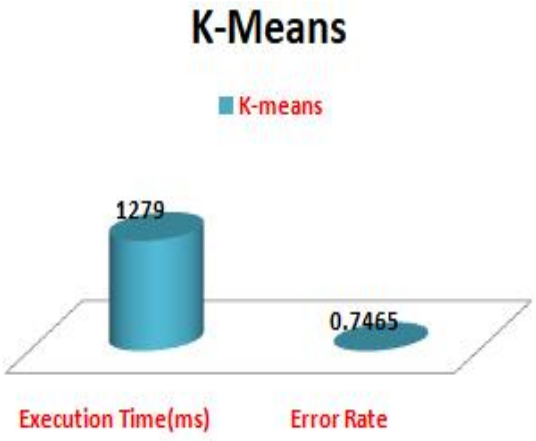

Fig.5: Performance of K-means Algorithm

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 6 August 2022

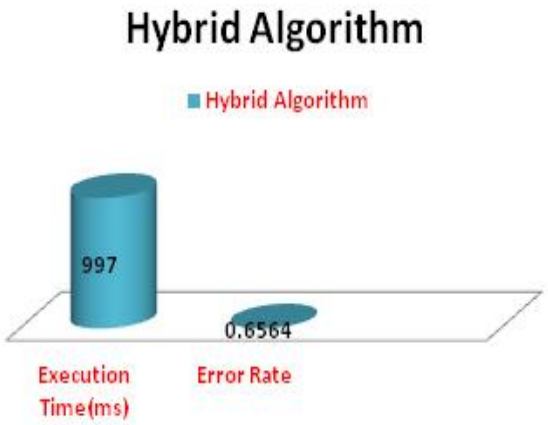

Fig.6: Performance of Hybrid Clustering Algorithm

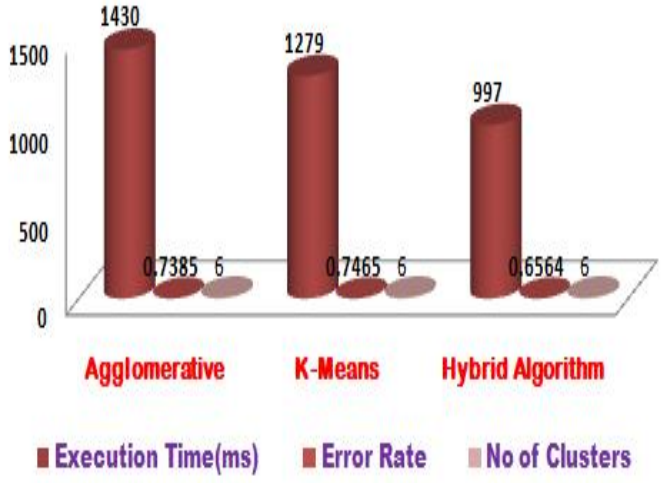

Fig.7: Comparison between Agglomerative, K-means and Proposed Hybrid Clustering Algorithm

From the above three graphical representation of k- means, agglomerative and hybrid of mechanism the depicted values and graphs says that proposed hybrid clustering mechanism is well performed than the other two clustering mechanisms.

CONCLUSION

In the field of data analysis, managing large amounts of data is a key difficulty. Clustering algorithms are useful for dealing with large-scale data, such as microarray gene expression datasets, in this type of study. Microarrays enable simultaneous observation of the expression profiles of thousands of genes under a variety of experimental settings. Some clustering techniques available can produce higher accuracy on small volume of dataset but not for huge volume of dataset. Therearenumerousclusteringstrateg to choose from, but determining which method is best for a given dataset is difficult to predict. According to the findings of the experiments, the updated or hybridized clustering algorithm reduced the execution time and error rate, while the k-means and agglomerative approach produced higher execution time and error rat, this can improve the work efficiency of prediction on air pollution dataset.

ACKNOWLEDGMENT

This article has been written with the financial Support of RUSA-Phase 2.0 grant sanctioned vide Letter NO.F,24-51/2014-U,Policy (TN Multi-Gen),Dept of Edn. Govt of India, Dt. 09.10.2018

REFERENCES

[1] Kohonen T, 'The self-organizing map', *Proc. IEEE*, Vol.78, No.9, (1990), pp.1464–1480.
[2] Vesanto J & Alhoniemi E, 'Clustering of the Self Organizing Map', *IEEE Transactions on Neural Networks*, Vol.11, (2000), pp.586–600.
[3] S.Suganya[1], T.Meyyappan[2] ,S.Santhosh kumar[3] "Performance Analysis of KMeans and KMediods Algorithms in Air Pollution Prediction" International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-5, DOI:10.35940/ijrte.E6495.018520, Published By: Blue Eyes.
[4] Kaufman L & Rousseeuw PJ, Finding Groups in Data, Wiley, (1990).
[5] Zahra Z, Amirhossein H & Ali MN, "Computational methodologies for analyzing, modeling and controlling gene regulatory networks", Biomedical Engineering and Computational Biology, Vol.2, (2010), pp.47–62.
[6] Dey L & Mukhopadhyay A, 'Microarray Gene Expression Data Clustering using PSO based K-means Algorithm', Proceedings of the International Conference Advanced Computing, Communication and Networks, (2011), pp.587-591.
[7] Zhang T, Ramakrishnan R & Livny M, "BIRCH An efficient data clustering method for very large databases", SIGMOD International Conference on Management of Data, (1996), pp.103-114
[8] Dey L & Mukhopadhyay A, 'Microarray Gene Expression Data Clustering using PSO based K-means Algorithm', Proceedings of the International Conference Advanced Computing, Communication and Networks, (2011), pp.587-591.
[9] Zhang T, Ramakrishnan R & Livny M, "BIRCH An efficient data clustering method for very large databases", SIGMOD International Conference on Management of Data, (1996), pp.103-114

AUTHORS PROFILE

S. Suganya M.sc., B.Ed., PhD Research scholar, Department of computer science, Alagappa University, Karaikudi, Tamilnadu. India. Her research area are Big Data Analysis and Data Mining.

Dr. T. Meyyappan M.Sc., M.Tech., M.Phil., Ph.D. currently, Professor, Department of Computer Science , Alagappa University ,Karaikudi ,Tamilnadu. India. He has organized conferences, Workshops at national and international levels. He has published 90 numbers of research papers in National and International journals and conferences. He has developed Software packages for Examination, Admission Processing and official Website of Alagappa University. As a Co-Investigator, he has completed Rs.1crore project on smart and secure environment funded by NTRO, New Delhi. As principal Investigator, he has completed Rs. 4 lakhs project on Privacy Preserving Data Mining funded by U.G.C. New Delhi. He has been honoured with Best Citizens of India Award 2012. His research areas include Operational Research, Digital Image Processing, Fault Tolerant computing, Network security and Data Mining.