“Uncertain Data clustering in Peer to Peer Networks”

A PROJECT REPORT

SUBMITTED IN THE PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

SOFTWARE ENGINEERING

Submitted By

Divyanka

(2K19/SWE/05)

Under the supervision of

Dr. Shailender Kumar
Associate Professor
Department of Computer Science & Engineering
Delhi Technological University, Delhi

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2021

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Divyanka, 2K19/SWE/05 student of M.Tech (SWE), hereby declare that the project entitled **"Uncertain Data clustering in Peer to Peer Networks"** which is submitted by to the Department of Software Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology in Software Engineering, has not been previously formed the basis for any fulfilment of the requirement in any degree or other similar title or recognition.

This report is an authentic record of my work carried out during my degree under the guidance of Dr. Shailender Kumar.

Divyanka

Place: Delhi **Divyanka**

Date: 27th June, 2021 **(2K19/SWE/05)**

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the project entitled **"Uncertain Data clustering in Peer to Peer Networks"** which is submitted by Divyanka (2K19/SWE/05) to the Department of Software Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology in Software Engineering, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or diploma to this university or elsewhere.

Place: Delhi
Date: 28/07/2021

Dr. Shailender Kumar
SUPERVISOR
Associate Professor
Dept. of Computer Science & Engineering

ACKNOWLEDGEMENT

I am very thankful to **Dr. Shailender Kumar** (Associate Professor, Department of Computer Science) and all the faculty members of the Department of Software Engineering at DTU. They all provided us with immense support and guidance for the project.

I would also like to express my gratitude to the University for providing us with the laboratories, infrastructure, testing facilities, and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors, and our peer group who aided us with all the knowledge they had regarding various topics.

Divyanka

Divyanka
(2K19/SWE/05)

Abstract

A sensational rise of various applications and businesses has led to the rise of data being collected, it is getting harder to store this data on a single machine. It has become more feasible to manage dispersed data sets and organizations. Recently a new technique has been introduced called distributed data mining (DDM). It is becoming popular due to the fact that analysing huge data sets present at different locations, applying traditional centralized techniques could be inefficient. That is the reason distributed data mining methods have become popular. Perhaps the main hurdle of data mining is to perform data clustering. It is quite possible that data produced in distributed environments contain noise or obsolete information, this type of data is termed as uncertain data. This type of data is difficult to cluster and consequently it becomes difficult to make any business decision or inference. Uncertain data clustering has been perceived as a fundamental step in the analysis of data mining, or in other words it has become a part of pre-processing.

Many centralized clustering calculations have been modified by characterizing new estimations. While many clustering calculations have been introduced for certain and uncertain data sets, there is need of productive calculations for distributed data sets. A modified hierarchical clustering algorithm is proposed for uncertain dataset to improve timing in execution of the process. This algorithm does not require any pre-specified number of clusters. Efficiency of the algorithm is worth noting and is tested with real world data sets.

CONTENTS

Candidate's Declaration 2

Certificate 3

Acknowledgement 4

Abstract 5

Contents 6

List of Figures 7

List of Tables 8

Chapter 1 Introduction 9

1.1 General 12

1.2 Problem Formulation 14

1.3 Objectives of the Project 15

Chapter 2 Literature Review 16

Chapter 3 Theoretical Concepts 159

3.1 Data analysis and clustering algorithms 19

3.2 Peer to peer Network (P2P) 20

3.3 Data uncertainty in P2P networks 21

Chapter 4 Proposed model for Uncertain clustering in P2P networks 18

4.1 Hierarchical clustering 21

4.2 Proposed algorithm 21

Chapter 5 Experimental Setup 24

Chapter 6 Results and Discussion on Results 30

Chapter 7 Conclusion and Future Scope 34

References 35

List of Figures

Fig. 1: Clustering Algorithm Procedure…………………………………………………...15

Fig. 2: Example of P2P………………………………………………………………18

Fig. 3: Data points in space…………………………………………………………...19

Fig. 4: Initial points…………………………………………………………………...19

Fig. 5: Clustering of initial points……………………………………………………….19

Fig. 6: Initial points combined in a single cluster ……………………………………………20

Fig. 7: Points in space………………………………………………………………....21

Fig. 8: Hierarchical Clustering Calculation……………………………………………...22

Fig. 9: GUI……………………………………………………………………………….24

Fig. 10: Upload dataset……………………………………………………………….25

Fig. 11: Calculation of weight screen ……………………………………………………26

Fig. 12: Attribute weight of loaded dataset………………………………………………26

Fig. 13: Selecting cluster size………………………………………………………………27

Fig. 14: DKmeans Output………………………………………………………………..27

Fig. 15: Input cluster size to WDUK-means……………………………………………..28

Fig. 16: Weighted Distributed Uncertain K means Output………………………………28

Fig. 17: Modified Hierarchical Weighted Clustering…………………………………………29

Fig. 18: Data Flow Diagram……………………………………………………………..30

Fig. 19: Execution Time result UI……………………………………………………….30

Fig. 20: Execution time for Glass & Haberman dataset………………………………………31

Fig. 21: Execution time for Heart & Ionosphere dataset………………………………...31

Fig. 22: Execution time for Heart & Ionosphere dataset………………………...…………32

List of Tables

Table I: Proximity Matrix……………………………………………………………..22

Table II: Result comparison………………………………………………………… 32

CHAPTER 1

INTRODUCTION

1.1 General

Data clustering is quite possibly the most popular data mining tasks, which helps data mining tasks to efficiently work on huge datasets & knowledge discovery. It works, through different means, to find already obscure gatherings inside the data collections. In the previous years, significant advancement has been made in this field prompting the improvement of inventive and promising clustering algorithm. These traditional clustering algorithms present some difficult issues with respect to the speed, the throughput, and the adaptability. Distributed data mining manages the issue of data analysis where the data is distributed among different peers with client. Parallel clustering algorithms is arising too, as another distributed processing, providing better solution for some novel applications that include trade of data among large number of peers with minimal shared synchronization among peers. Distributed data sharing, distributed electronic business, and shared checking systems are dependent on an organization of sensors are a few models. Distributed data sharing frameworks currently create a critical bit of Internet traffic. A decent comprehension of their jobs is significant to improve their flexibility, robustness and execution. Lately, with the expanding number of various applications on distributed (P2P) organizations, uncertain data analysis in huge powerful organizations is undoubtedly going to gain a significant position in the industry soon. For instance, in a lodge booking system, clients are approached to assess and give rating & reviews of the lodges through a series of various parameters, like office data, sterile condition, administration quality, and area data. Every lodging can be scored by numerous clients. All assessments to a lodge ought to be demonstrated as an uncertain item on the client score space. Practically, different lodges might be registered in the diverse destinations that give reservation administration. A significant analysis work is to be done to cluster the client’s data from all destinations.

As a significant tool for data analysis, cluster analysis inspects unlabelled data, by either building a progressive construction, or shaping a bunch of data as indicated by a pre-specified number. This step is followed by progression of steps, going from pre-processing and calculation improvements, to arrangement of clusters to minimize the distance. Every data point in the cluster is identified in with respect to centroid, till minimal distance is achieved. There are many approaches available in the existing literature. These algorithms have their pros and cons, and tackle various issues. There are numerous approaches to address primary, as well as secondary issues like performance, user friendliness, security, and robustness etc. Among the potential models for uncertain data, there is a pressure among basic and advanced models, which will in general be fragmented, and complete models, which will in general be non-intuitive and more perplexing than needed for some applications.

1.2 Problem Formulation

In real life applications an enormous amount of data is being produced daily and captured by sensors, servers or databases. It is quite possible that the captured data contains undesired values or noise. When data contain errors like incorrect measurements, outliers, obsolete data, and other errors, then this type of data is termed as uncertain data. Uncertain data cannot be presented as a simple coordinate point. Data collection is one such issue. Sources that collect data might not be in the same zone and collecting them for any analysis becomes difficult because of security & privacy reasons. Sometimes data produced becomes obsolete and will be considered as outlier in data analysis.

Based on this problem following questions has been identified:

1. What is the current status of evaluation of uncertain data?
2. How do we determine the appropriate number of clusters in distributive environment?
3. Which algorithms are available for data clustering of uncertain items particularly for peer to peer networks?
4. Which datasets are available for data clustering?
5. What is the execution time achieved by existing data clustering algorithms for uncertain items?
6. How to improve execution time while retaining accuracy of the algorithm?

1.3 Objectives of the Project

In this project, the following objectives need to be achieved:

1. Collecting dataset for the simulation of peer to peer networks which would be suitable for cluster analysis.
2. Building an algorithm for uncertain data in the network.
3. Comparing the execution time of different clustering techniques in the existing literature.
4. Establish a baseline for improvement of the clustering techniques in the future.
5. Develop a new model that combines the capabilities of top-performers based on the baseline developed.
6. Comparing the performance of the developed model with other models.

CHAPTER 2

LITERATURE REVIEW

A lot of research is being conducted to counter this problem.

- *K. M. Hammouda and M. S. Kamel* [1]: A Hierarchically-distributed Peer-to-Peer (HP2PC) architecture and clustering algorithm is proposed by Khaled et al, to address the problem of modularity, flexibility and scalability. It used peer to peer architecture with multiple layers. Super nodes, which work as representatives, are clustered repeatedly to create a multilevel neighborhoods. All the peers communicate and coordinate with their corresponding neighbors to perform clustering at a certain stage of the hierarchy. Clustering can be done separately with neighborhoods using this model, and solve each component separately with the help of distributed K-means variant, then merge each component in the hierarchy to obtain optimal global result. It requires a hierarchy of peer-to-peer neighbourhoods, each peer develops a model on the basis of the data available to them using P2P communication. Groups from lower levels are integrated into the hierarchy as we escalate the hierarchy. This model is different from the traditional networks because HP2PC model relies on a pre-designed fixed hierarchical structure upon which the peer network is built.

- *S. Datta et al* [2]: In a research conducted by S. Datta et al, authors introduced two algorithms which approximates standard K-means algorithm. The first algorithm is proposed for dynamic networks and works in a distributed manner to achieve synchronization with their topological neighbors. The second algorithm employs uniformly sampled peers and provides empirical guarantees for grouping precision on a peer-to-peer network. This algorithm is suitable for dynamic networks where data keeps on changing for example the joining and dropping fashion of the nodes. Experiments have shown that accuracy is quite decent and the algorithms are robust enough to the changes in network topology.

- *M. Chau et al* [3]: Some of the research includes new distance similarity measures among uncertain items like ED- based similarity, proposed by Chau et al named UK-means (uncertain K-means). This algorithm uses expected distance(ED), as distance based similarity. Then other improvements are done via algorithms that run on top of it to reduce the complexity. Authors proposed UK-means clustering, which modifies K-means algorithm to solve the issue of data uncertainty. It is applied to the moving-object uncertainty. This algorithm yields good results and accuracy while considering uncertainty. The significant difference between UK-mean clustering and the traditional K-means can be noted during the computation of distance and clusters. Based on the data uncertainty model it calculate expected distance and cluster centroids. This algorithm is also applicable to moving entities.

- *Zhou et al* [4]: Apart from the distance-based similarity, another category of cluster algorithm is distribution-based similarity. Distribution-based algorithms consider divergence like Kullback–Leibler divergence to cluster the object with known distribution to measure similarity between data points. Each uncertain data point is exhibited as random variable which follows a probability distribution. KL divergence measures how two distributions are different. It is used to measures difference between uncertain objects by using probability distribution of each object. The Kullback Leibler divergence is termed as distance between two distributions. The Kullback Leibler divergence can be used in two cases:

a) Discrete: when probability mass function is taken as a discrete domain with finite number values and,

b) Continuous: when probability density functions is defined in a continuous domain.

- *Patil et al* [5] conducted experiments on real data sets to proof that using probability distribution for clustering uncertain is useful and considering Kullback Leibler divergence technique is efficient. KL divergence is applied prior to the start of the actual clustering technique (DBSCAN in this case).

- *Azimi et al* [6] proposed a new algorithm called GBDC-P2P. (Gossip Based Distributed Clustering algorithm for P2P networks). It extracts representative data using 3 methods. Peers perform data clustering with their neighbors in a distributive manner. This algorithm modifies K-means and K-medoids algorithms. It extracts the representative data and discover the final clustering results. Then another algorithm called CYCLON algorithm is used to establish connection between peers. During the establishment of connection between peer say p1 and p2, data packets are exchanged as well as the gossip message, between p1 and p2. Mainly two data packets are transmitted. First packet contains peer’s local data termed as internal data and another packet contains data received from its neighbors termed as external data, before the current round of gossip operation. As there is no data received from the neighboring peer initially, thus the size of external data is zero. All the peers find their respective final centroids. The data available locally on one peer is known as internal data, and data which is communicated among the peers is known as external data. Peers exchanges their internal and external data to the final centroids when they need clustering results. Internal data is hence clustered independently. Authors have also proposed Persistent K-means which gives better results as compared to traditional centralized clustering.

Following inferences are made after reviewing above papers:

1. Any traditional clustering algorithm was not able to solve the problem of uncertainty in the data. However they perform well when certain items were the agenda.
2. Most of the papers use multiple variants of K-means algorithm. The results are of great generality but the execution time achieved was quite high.
3. The peer networks being simulated are small in size because of hardware limitations.
4. The size of the datasets considered for testing the model performance is quite small, upto 300 instances only, otherwise clustering time will increase. To check the performance of any model it is sufficient.

CHAPTER 3

THEORETICAL CONCEPTS

This section presents the basic theoretical concepts required to understand the key processes and working of the experiment studied in this project. This section familiarizes the concept of data clustering, peer to peer networks, and uncertainty present in the data, produced by data sources. It also induces the idea of working on different kinds of clustering algorithm. The concepts introduced in this section will help to understand the proposed algorithm for uncertain data clustering in peer to peer networks.

3.1 Data analysis and Clustering algorithms

Data analysis plays a crucial part for understanding different aspects. Clustering is a task that divides data sets into a different clusters, such that the data having similar properties are put into same cluster. There are various algorithms available according to our need. We review clustering algorithms for various fields like software engineering, AI, sales, and bioinformatics. A few firmly related subjects like distance measure, and cluster approval, are also surveyed.

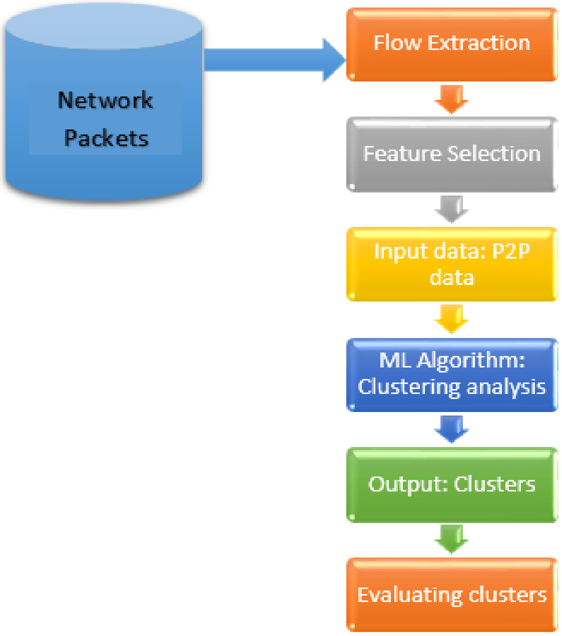


Fig. [1] Supervised Learning Procedure

Fig. 1 shows learning procedure basics in clustering algorithms.

- Feature selection: Data is preprocessed on the basis of dimensionality reduction. This step extract or select the important features. Other irrelevant features are discarded because it increases unnecessary complexity in the clustering process.
- ML algorithm: An appropriate cluster algorithm is chosen which fits our usecase, for generating better results. Then the algorithm is executed to discover the clusters within our data set.
- Evaluating Clusters: The identified clusters are evaluated to examine whether they are properly identified or not. There are different types of cluster evaluation techniques:

  1) internal, 2) external and 3) relative.

  After evaluation of the clusters, the clusters are analyzed and interpreted, appropriate conclusion is made.

3.2 Peer to peer Network (P2P)

It is an architecture suitable for distributed applications that partitions or distribute the jobs or among the peers. This technology is a big advancement in the IT world since it enhances companies' efficiencies. Peer-to-peer networking has also made communication far easier than it used to be, which is why many businesses have quickly adapted to this technology.

Another primary benefit of this software is that it allows companies to stay connected within the workplace, increasing a business's overall efficiency. Additionally, once your entire computer system has synced, you may be able to keep your data collected and checked in one place, advancing the overall security of your data.

There are two types of P2P networks: Unstructured P2P networks and Structured P2P networks. Unstructured networks are framed arbitrarily by joining and dropping connections over the long run, and they experience the downsides of flooding of traffic. Structured networks, analyses the network architecture and implement its protocol using that information. Structured networks, have a topological structure and generally a distributed hash table (DHT) is used to form and manage the network.

3.3 Data uncertainty in P2P networks

When data contain errors like incorrect measurements, outliers, obsolete data, and other errors, then this type of data is termed as uncertain data. It is the data that containing noise which consequently gets deviated from original or correct values. This type of data is present in large amounts over internet, sensor networks etc. Data analysis is an important activity when it comes to decision making. So analysis on uncertain data will be a contributing factor towards the quality of decision making, thus discrepancies in the uncertain data must be analysed thoroughly. An estimation is required to get a correctness check on the data.

Uncertain data clustering has been perceived as a fundamental step in the analysis of data mining, or in other words it has become a part of preprocessing. Many incorporated clustering calculations have been modified by characterizing new estimations. With the quick advancement of organization applications, data is now stored in distributed data centers rather than a centralized machine. Storing data on a centralized machine have limitations like security, data protection and specific requirements brought by distributive environment. Recently a new technique has been introduced called distributed data mining (DDM). It is becoming popular due to the fact that analyzing huge data sets present at different locations, applying traditional centralized techniques could be inefficient. With the advancements achieved in the field of distributed infrastructures, grid computing platforms, DDM is going to be in demand. It’s quite obvious to assume that data is distributed in various locations in DDM. This assumption is appropriate for us to build a model with the help of data mining algorithms that reveals the features of the entire data set. Data analysis is essential for understanding the characteristics of a business in different domains. There are numerous methods for data analysis. But choosing the right technique is very important. We shall discuss the proposed algorithm in the next section.

CHAPTER 4

PROPOSED ALGORITHM FOR UNCERTAIN DATA CLUSTERING IN P2P NETWORKS

This section presents the proposed model, in which we have considered a distributed P2P network and peers are evenly distributed over a 500 m × 500 m region. The peer communication range has been set at 100m. Figure 1 illustrates an example of distributed P2P network for 50 peers.

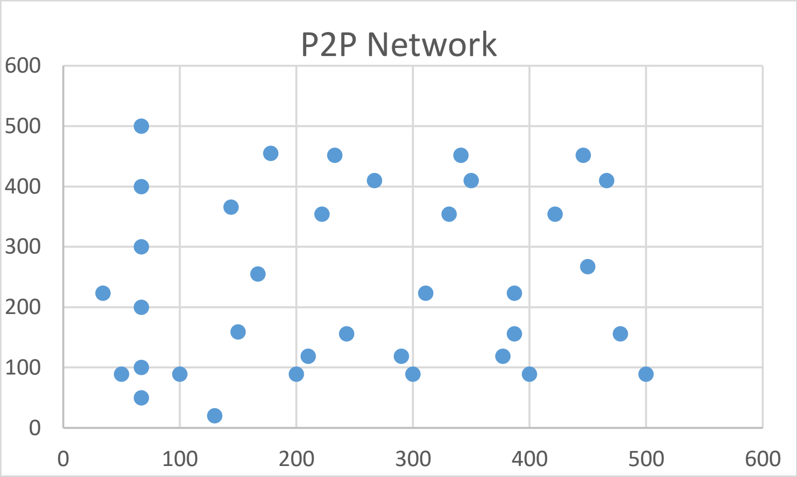


Fig. 2 Example of P2P

Here we introduce a new algorithm, we term it as, Modified Hierarchical Weighted Clustering algorithm, or Modified HWC algorithm. Hierarchical Algorithm is the clustering algorithm which requires no prior K value and provide full freeness to data to arrange itself in matching cluster which combine with processing on uncertain data. We have implemented a weighted hierarchical clustering algorithm on uncertain data clustering.

4.1 Hierarchical clustering concept

A hierarchical clustering approach operates in a cluster tree by grouping results. It recursively creates a hierarchy of clusters called dendrogram. Hierarchical clustering algorithms can be further divided into two algorithms on the basis of cluster decomposition.

A. What is Hierarchical Clustering?

Suppose we have certain data points as shown in Figure 2.

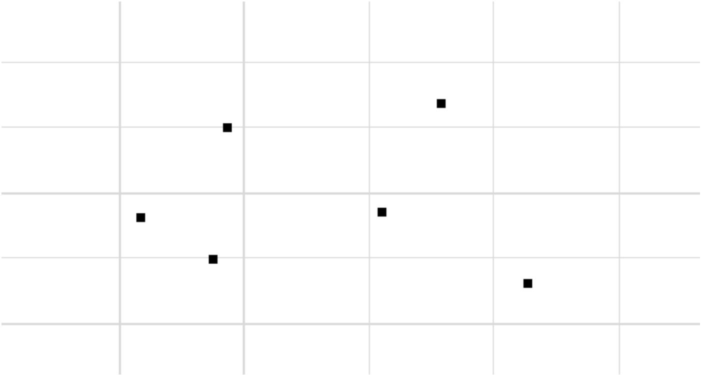


Fig. 3 Data points in space

So initially they will be clustered in different groups as initial points as shown in Figure 3.

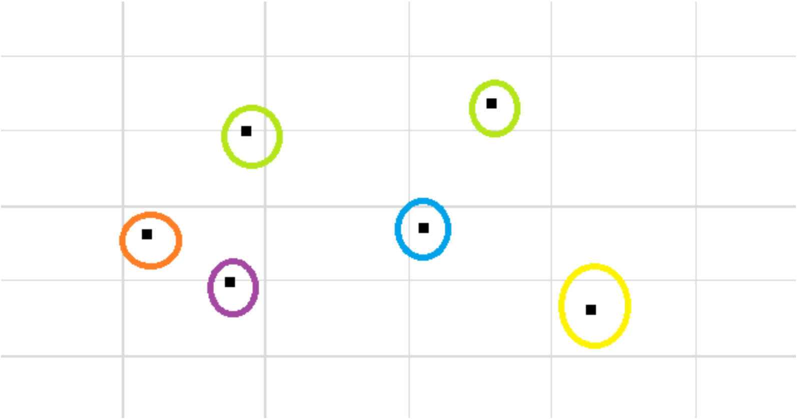


Fig. 4 Initial points

Our aim is to cluster them on the basis of similarity.

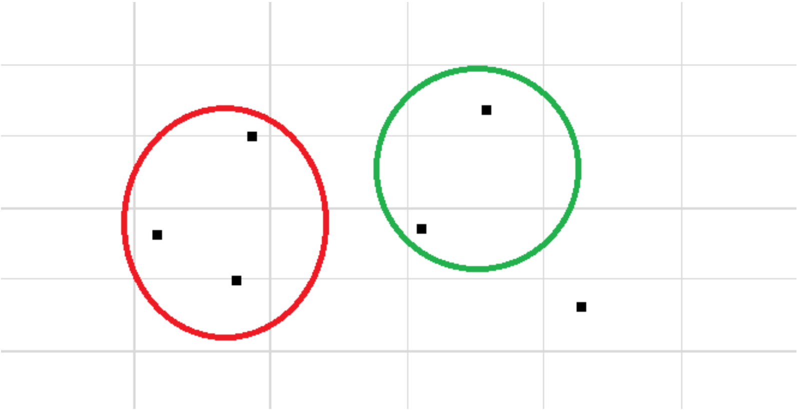


Fig. 5 Clustering of initial points

The data points possessing same behavior will be grouped in a separate cluster and this process is repeated until only one cluster is obtained, as shown Figure 5.

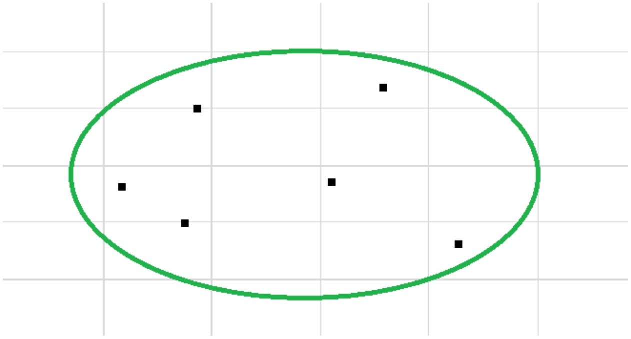


Fig. 6 Initial points combined in a single cluster

Hierarchy of clusters is built in above Figure 5. This process is called hierarchical clustering.

There are two types of hierarchical clustering:

a) Agglomerative hierarchical clustering

b) Divisive Hierarchical clustering

a) Agglomerative Hierarchical Clustering

In this technique, each point is grouped in a separate cluster. Assume there are five data points. Each of these points is assigned to a cluster, resulting in five clusters initially. Then, closest pair of clusters are repeatedly merged till one cluster is obtained. It is also termed as additive hierarchical clustering because of the repetitive merging or adding.

b) Divisive Hierarchical Clustering

The algorithm starts with assigning all data points to a single cluster rather than n separate data points assigned to different clusters. So, irrespective of the number of data points, all the points reside in the same cluster initially. The data points are split into separate cluster which have the maximum distance from other points present in the cluster. This process is repeated till each cluster contains only a single point. This splitting or division of points is termed as divisive hierarchical clustering.

Hierarchical Clustering Algorithm

In hierarchical clustering similar data is grouped together and dissimilar data is separated out. But the most important question is to determine which points are similar and which are dissimilar. To calculate similarity, distance between the centroids is a good choice to consider. Those data points having the minimal distance falls into the category of similar points and can be merged, hence it is a distance-based algorithm. In hierarchical clustering, proximity matrix is used to calculate the distance. It stores the distances between each point.

Creating a Proximity Matrix

1) Consider 4 points in the space for the simplicity

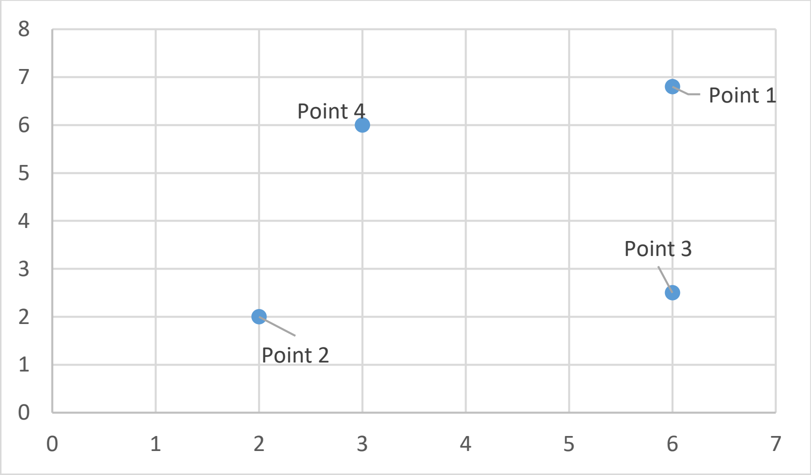


Fig. 7 Points in space

2) Build a proximity matrix which stores the distance between the data points shown in Figure 6. As the distance of a point to itself zero thus diagonal values will be zero always. To calculate the distance we are considering Euclidean distance. We have 2 points say, point 1 & point 2, distance between point 1 (2, 2) and point 2 (6, 6) will be:

√ (6-2) ^2 = √4 = 2

3) Calculate the distance of each point with respect to every other points, and fill the square matrix as shown below:

Following matrix is obtained:

Table 1. Proximity Matrix

	Point 1	Point 2	Point 3	Point 4
Point 1	0	4.12	4.33	3.1
Point 2	4.12	0	4.03	4.25
Point 3	4.33	4.03	0	4.12
Point 4	3.1	4.25	4.12	0

The proximity calculation is shown in the Figure 7, in terms of hierarchical tree

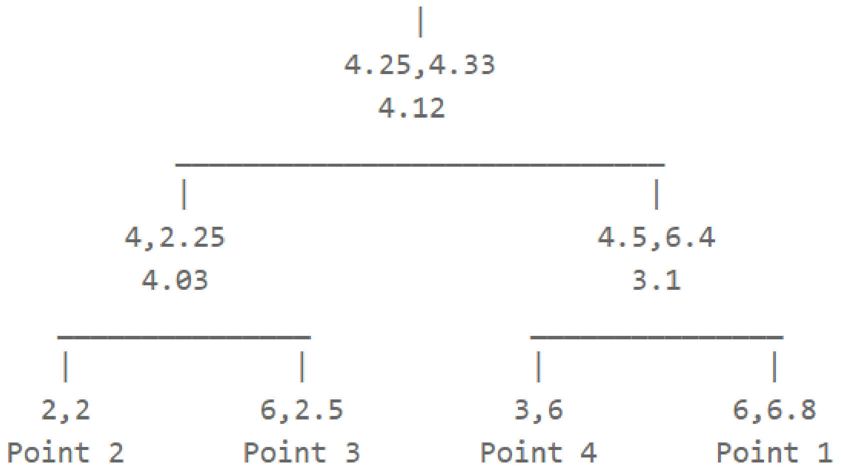


Fig. 8 Hierarchical Clustering Calculation

Hierarchical clustering starts with each data point being treated as a single cluster. Then the following procedures are consistently carried out: List the 2 clusters nearest to each other and merge the 2 equivalent maximum clusters. These procedures must be continued before all of the clusters are combined.

4.2 Proposed Algorithm

Introducing an algorithm Modified Hierarchical Weighted Clustering (MHWC), which combines the properties of hierarchical clustering and k-means.

Proposed Algorithm steps are as follows-

1. Calculate the relation between one cluster and all other (calculate proximity matrix)
2. Consider each data point as a single cluster
3. Fuse clusters that are very close to each other or very near.
4. Calculate each cluster's proximity matrix
5. Repeat steps 3 and 4 until there is just one cluster left.
6. This is then combined with the weighted attribute calculations

Weighted attribute is decided by similarity threshold given in algorithm below

```
if(similarity > 40){
                    data.add(vector.get(i))
                    weighted.add(i)
            }
```

The vector weights of the clustering variable are measured by the algorithm. The algorithm can also be used as a variable collection in the framework for data mining where massive and complex real data are often involved. For strong clustering outcomes, choose higher weighted values and exclude lower weighted variables.

CHAPTER 5

EXPERIMENTAL SETUP

To assess the performance of DK-means, WDK-means, Modified HWC algorithms, experiments are performed on real world data set. Experimental data chosen in this experiment is integer, real. We have taken seven real world dataset which are available on UCI Machine Learning repository which include:

1) Glass (N=214, K=2)

2) Heart disease (N=271),

3) Ionosphere (N=351),

4) Iris (N=150),

5) SomervilleHappinessSurvey2015 (N=144),

6) Wine (N=178), and

7) Haberman (N=306), where N indicates number of instances or data objects.

Figure. 9 shows the main GUI screen for the implementation of the proposed work.

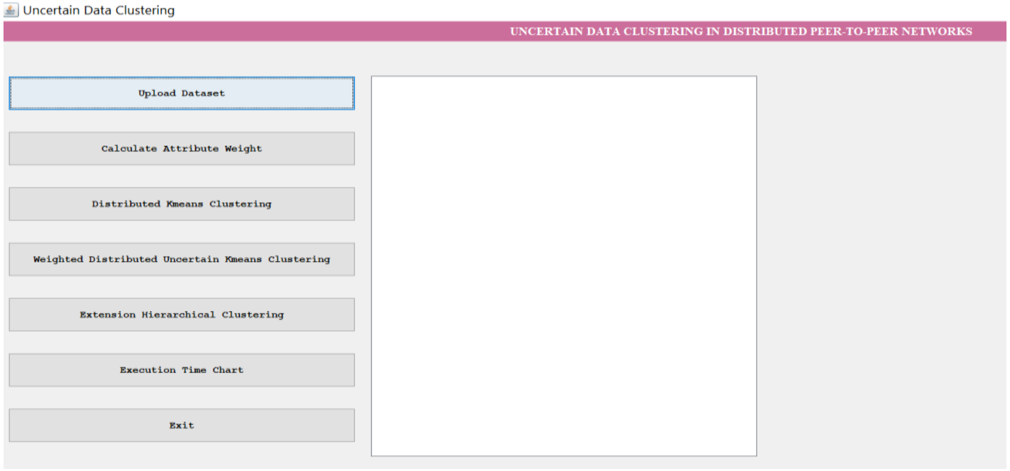
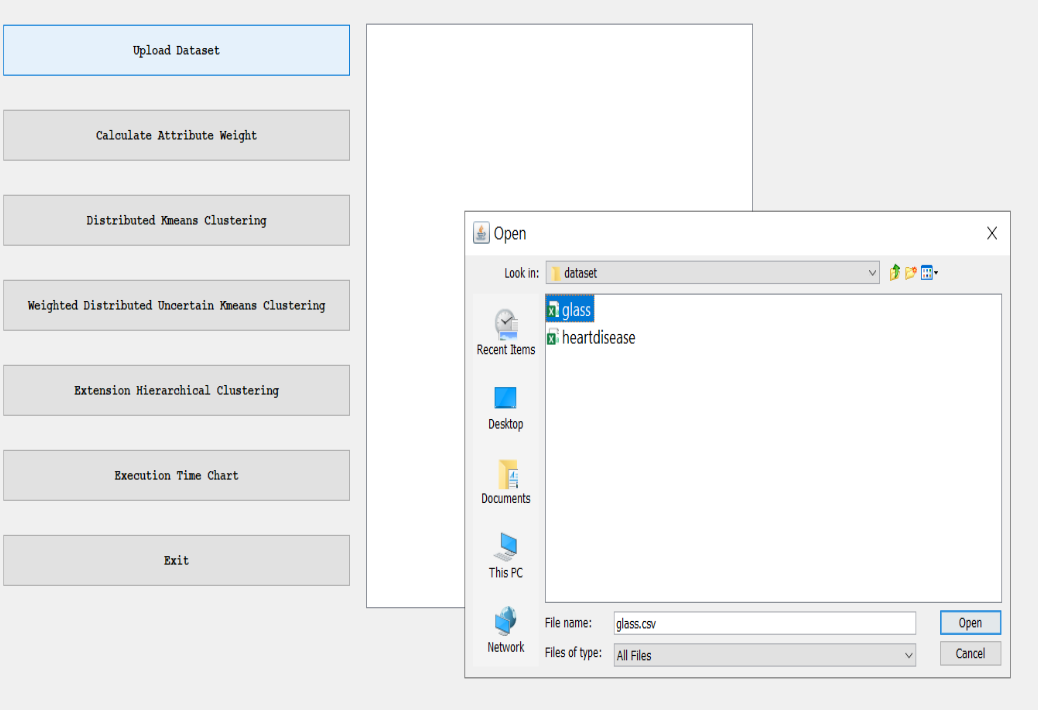

Uncertain Data Clustering
UNCERTAIN DATA CLUSTERING IN DISTRIBUTED PEER-TO-PEER NETWORKS
Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit

Fig. 9 GUI

Step 1) upload the dataset.

The dataset considered contains small number of data points, and a simple network architecture is followed, because larger data set takes more execution time for clustering.

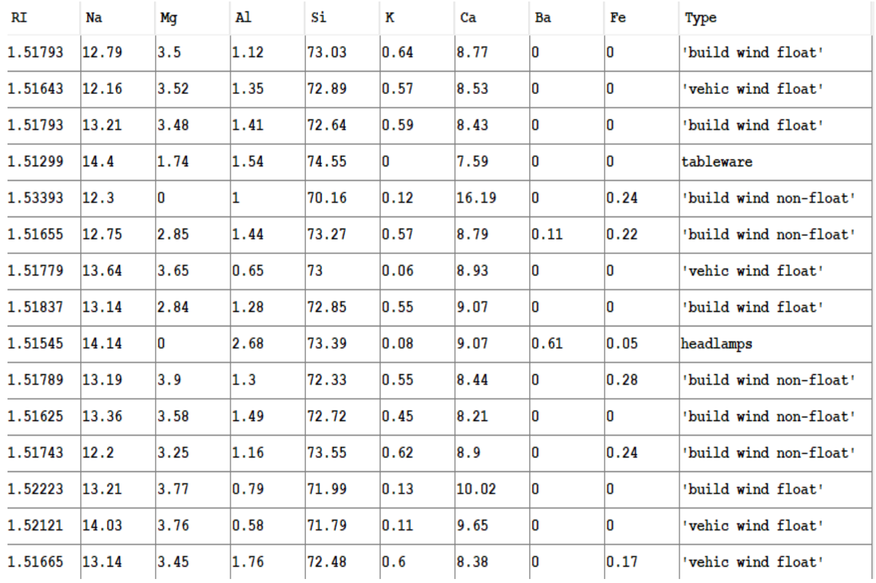

Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit
Open
Look in:
dataset
Recent Items
Desktop
Documents
This PC
Network
glass
heartdisease
Open
Cancel

File name: glass.csv

Files of type: All Files

Fig. 10 Upload dataset

In Figure 10 we select and load the desired dataset file in CSV format.

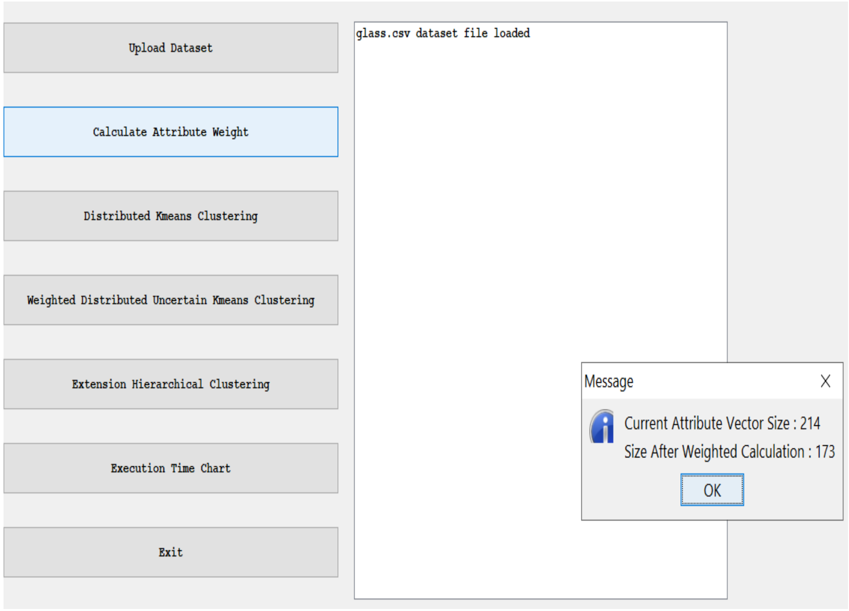


RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1.51793	12.79	3.5	1.12	73.03	0.64	8.77	0	0	'build wind float'
1.51643	12.16	3.52	1.35	72.89	0.57	8.53	0	0	'vehic wind float'
1.51793	13.21	3.48	1.41	72.64	0.59	8.43	0	0	'build wind float'
1.51299	14.4	1.74	1.54	74.55	0	7.59	0	0	tableware
1.53393	12.3	0	1	70.16	0.12	16.19	0	0.24	'build wind non-float'
1.51655	12.75	2.85	1.44	73.27	0.57	8.79	0.11	0.22	'build wind non-float'
1.51779	13.64	3.65	0.65	73	0.06	8.93	0	0	'vehic wind float'
1.51837	13.14	2.84	1.28	72.85	0.55	9.07	0	0	'build wind float'
1.51545	14.14	0	2.68	73.39	0.08	9.07	0.61	0.05	headlamps
1.51789	13.19	3.9	1.3	72.33	0.55	8.44	0	0.28	'build wind non-float'
1.51625	13.36	3.58	1.49	72.72	0.45	8.21	0	0	'build wind non-float'
1.51743	12.2	3.25	1.16	73.55	0.62	8.9	0	0.24	'build wind non-float'
1.52223	13.21	3.77	0.79	71.99	0.13	10.02	0	0	'build wind float'
1.52121	14.03	3.76	0.58	71.79	0.11	9.65	0	0	'vehic wind float'
1.51665	13.14	3.45	1.76	72.48	0.6	8.38	0	0.17	'vehic wind float'

Fig. 11 Calculation of weight screen

Figure 11 displays the uploaded dataset.

Step 2) now calculate attribute weight for the loaded dataset as shown in the Figure 12.

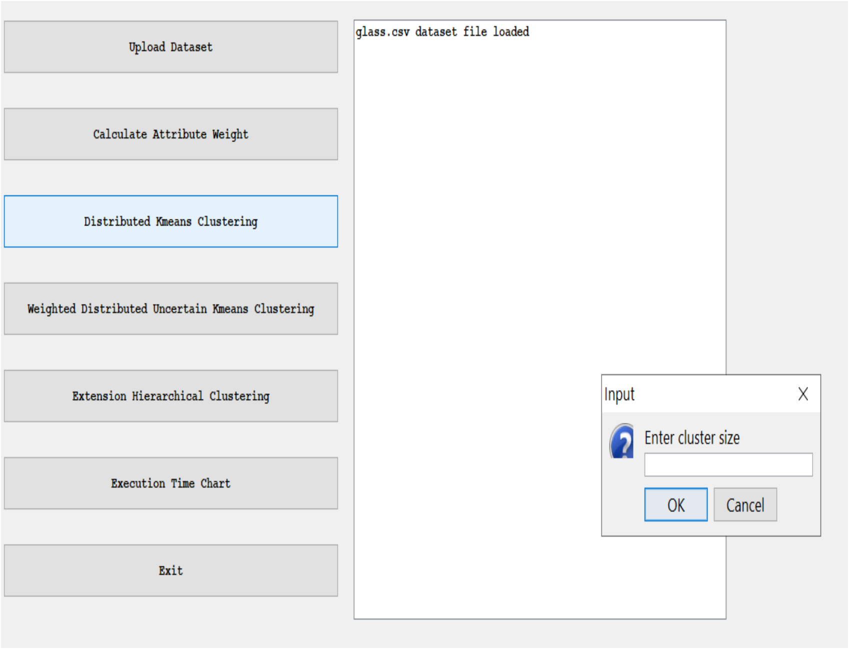

Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit
Message
X
OK

glass.csv dataset file loaded

Current Attribute Vector Size : 214
Size After Weighted Calculation : 173

Fig. 12 Attribute weight of loaded dataset

Step 3) In Figure 6. DK-means (Distributed Kmeans) clustering algorithm is applied on the loaded dataset by providing K value i.e. number of clusters.

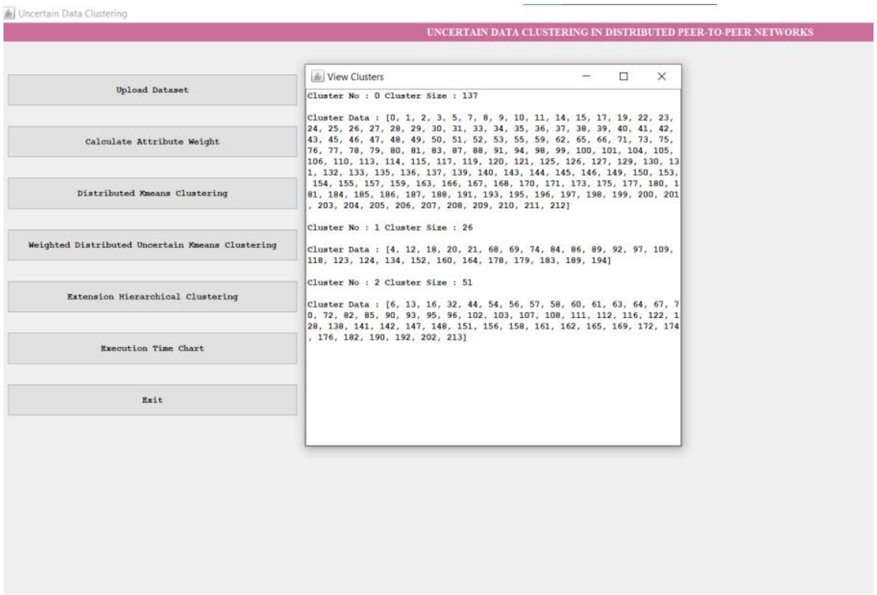

Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit
Input
OK
Cancel

```
glass.csv dataset file loaded
```

Enter cluster size

Fig. 13 Selecting cluster size

After inputting K value in Step 3 we obtain clustering result as shown below in Figure 13

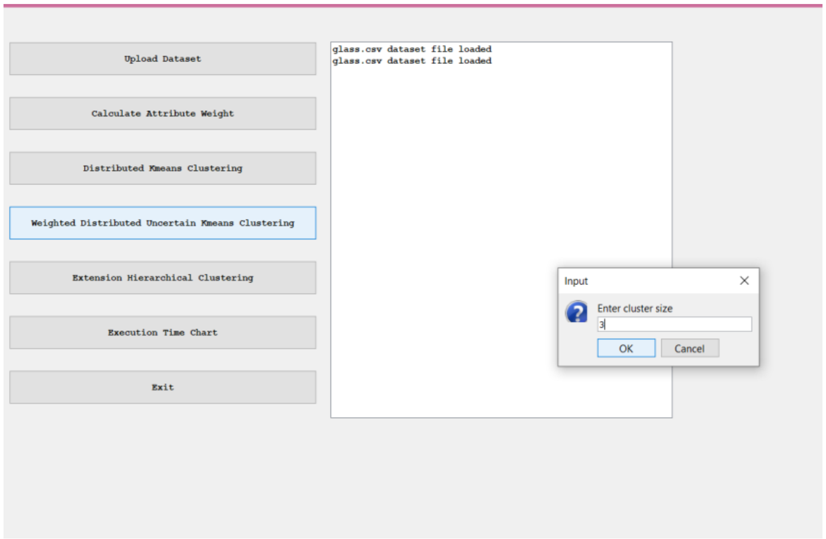

Uncertain Data Clustering
UNCERTAIN DATA CLUSTERING IN DISTRIBUTED PEER-TO-PEER NETWORKS
Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit
View Clusters

```
Cluster No : 0 Cluster Size : 137

Cluster Data : [0, 1, 2, 3, 5, 7, 8, 9, 10, 11, 14, 15, 17, 19, 22, 23,
24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 55, 59, 62, 65, 66, 71, 73, 75,
76, 77, 78, 79, 80, 81, 83, 87, 88, 91, 94, 98, 99, 100, 101, 104, 105,
106, 110, 113, 114, 115, 117, 119, 120, 121, 125, 126, 127, 129, 130, 13
1, 132, 133, 135, 136, 137, 139, 140, 143, 144, 145, 146, 149, 150, 153,
 154, 155, 157, 159, 163, 166, 167, 168, 170, 171, 173, 175, 177, 180, 1
81, 184, 185, 186, 187, 188, 191, 193, 195, 196, 197, 198, 199, 200, 201
, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212]

Cluster No : 1 Cluster Size : 26

Cluster Data : [4, 12, 18, 20, 21, 68, 69, 74, 84, 86, 89, 92, 97, 109,
118, 123, 124, 134, 152, 160, 164, 178, 179, 183, 189, 194]

Cluster No : 2 Cluster Size : 51

Cluster Data : [6, 13, 16, 32, 44, 54, 56, 57, 58, 60, 61, 63, 64, 67, 7
0, 72, 82, 85, 90, 93, 95, 96, 102, 103, 107, 108, 111, 112, 116, 122, 1
28, 138, 141, 142, 147, 148, 151, 156, 158, 161, 162, 165, 169, 172, 174
, 176, 182, 190, 192, 202, 213]
```

Fig. 14 DKmeans Output

In above Figure. 13 in each cluster, we can see data is arranged in 3 clusters, with cluster number and cluster.

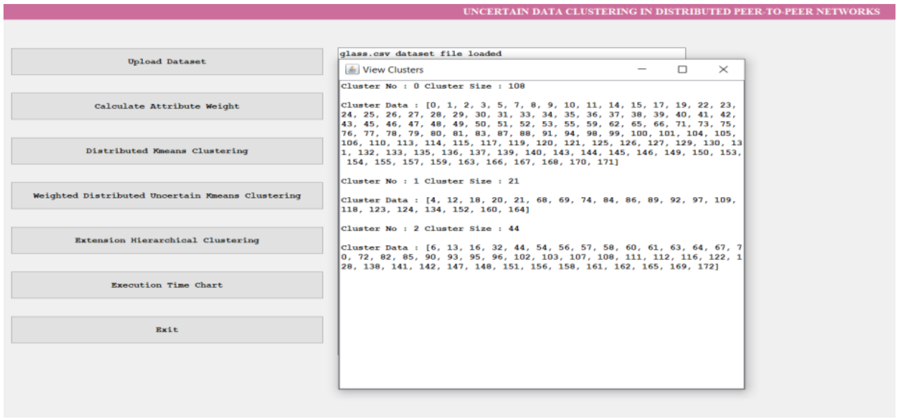
Step 4) now WDUK-means (Weighted Distributed Uncertain Kmeans) is applied. Number of clusters is inputted, in Figure. 8, then data clustering is done using attribute-weight regularization.


Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit
Input
OK
Cancel

```
glass.csv dataset file loaded
glass.csv dataset file loaded
```

Enter cluster size: 3

Fig. 15 Input cluster size to WDUK-means


UNCERTAIN DATA CLUSTERING IN DISTRIBUTED PEER-TO-PEER NETWORKS
Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit
View Clusters

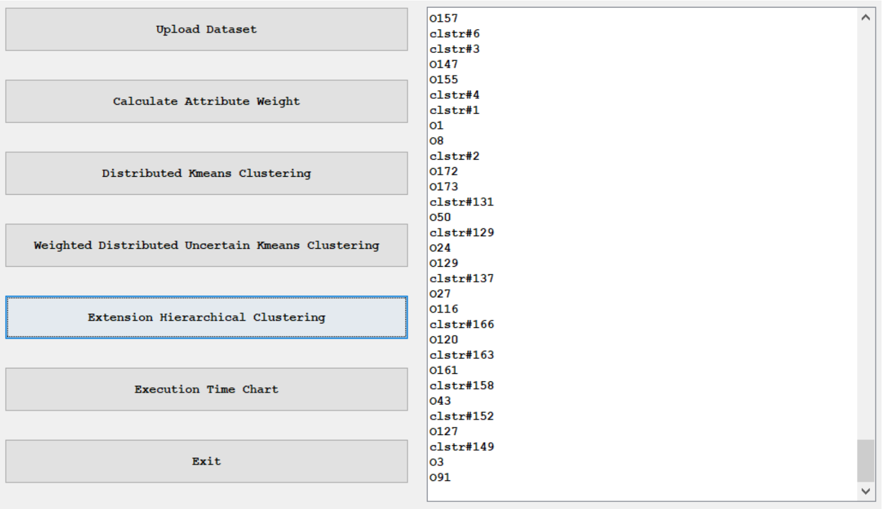
```
glass.csv dataset file loaded
```

```
Cluster No : 0 Cluster Size : 108

Cluster Data : [0, 1, 2, 3, 5, 7, 8, 9, 10, 11, 14, 15, 17, 19, 22, 23,
24, 25, 26, 27, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,
43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 55, 59, 62, 65, 66, 71, 73, 75,
76, 77, 78, 79, 80, 81, 83, 87, 88, 91, 94, 98, 99, 100, 101, 104, 105,
106, 110, 113, 114, 115, 117, 119, 120, 121, 125, 126, 127, 129, 130, 13
1, 132, 133, 135, 136, 137, 139, 140, 143, 144, 145, 146, 149, 150, 153,
 154, 155, 157, 159, 163, 166, 167, 168, 170, 171]

Cluster No : 1 Cluster Size : 21

Cluster Data : [4, 12, 18, 20, 21, 68, 69, 74, 84, 86, 89, 92, 97, 109,
118, 123, 124, 134, 152, 160, 164]

Cluster No : 2 Cluster Size : 44

Cluster Data : [6, 13, 16, 32, 44, 54, 56, 57, 58, 60, 61, 63, 64, 67, 7
0, 72, 82, 85, 90, 93, 95, 96, 102, 103, 107, 108, 111, 112, 116, 122, 1
28, 138, 141, 142, 147, 148, 151, 156, 158, 161, 162, 165, 169, 172]
```

Fig. 16 Weighted Distributed Uncertain K means Output

In above Figure. 16 WDUK-means arrange all data in 3 clusters.

Step 4) now apply Extension Hierarchical Clustering for Modified HWC (Modified Hierarchical Weighted Clustering). In Modified HWC we are not required to enter any prior K value and below is the result


Upload Dataset
Calculate Attribute Weight
Distributed Kmeans Clustering
Weighted Distributed Uncertain Kmeans Clustering
Extension Hierarchical Clustering
Execution Time Chart
Exit

```
o157
clstr#6
clstr#3
o147
o155
clstr#4
clstr#1
o1
o8
clstr#2
o172
o173
clstr#131
o50
clstr#129
o24
o129
clstr#137
o27
o116
clstr#166
o120
clstr#163
o161
clstr#158
o43
clstr#152
o127
clstr#149
o3
o91
```

Fig. 17 Modified Hierarchical Weighted Clustering

In above Figure. 17 shows cluster arrangement. In the result panel, 'clstr#' indicates the cluster number, and

'o' indicates the or data object arranged in that cluster.

For example, clstr#134 is the cluster number and o5 and o103 are the data object arranged in that cluster, which indicates object 5 and 103 contains similar data.

Modified Hierarchical Weighted clustering created 172 clusters for the whole dataset and provides full freeness to arrange data itself in matching cluster and if data is not matching it will create a new cluster.

CHAPTER 6

RESULTS AND DISCUSSION ON RESULTS

Figure 18 shows the final Data flow diagram.

Figure. 19 shows the execution time chart improvement compared with previous methods.

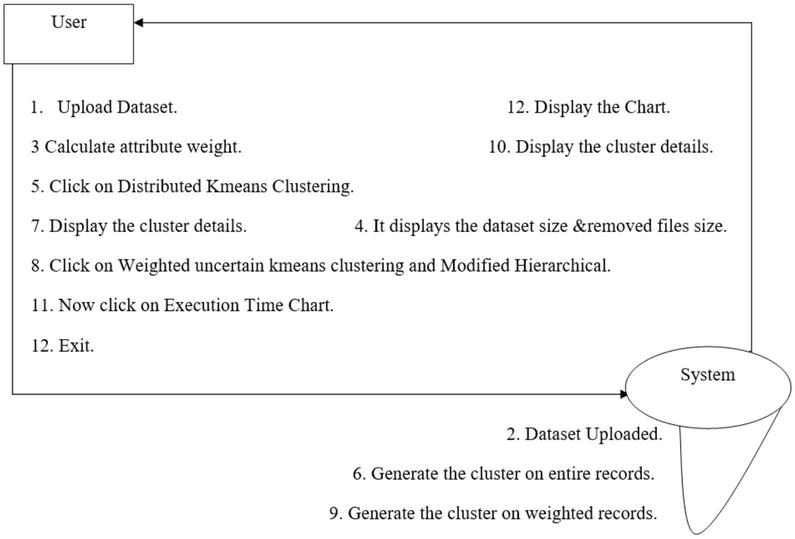

User
1. Upload Dataset.
3 Calculate attribute weight.
5. Click on Distributed Kmeans Clustering.
7. Display the cluster details.
8. Click on Weighted uncertain kmeans clustering and Modified Hierarchical.
11. Now click on Execution Time Chart.
12. Exit.
12. Display the Chart.
10. Display the cluster details.
4. It displays the dataset size &removed files size.
System
2. Dataset Uploaded.
6. Generate the cluster on entire records.
9. Generate the cluster on weighted records.

Fig. 18 Data Flow Diagram

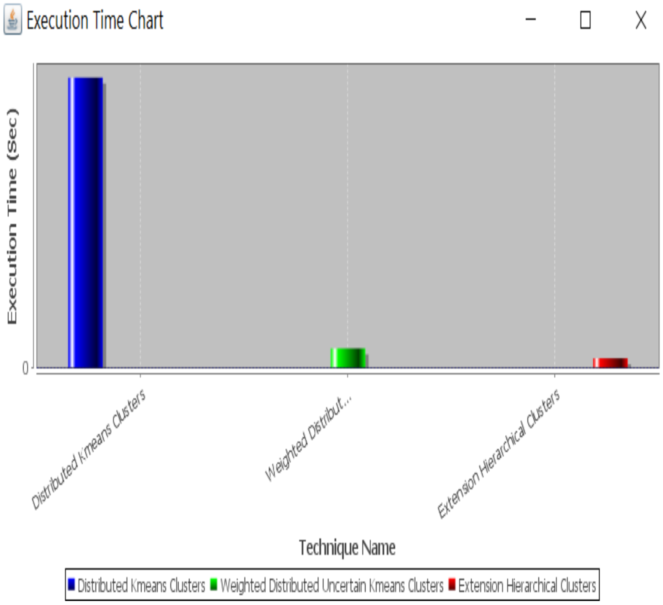

Execution Time Chart
Execution Time (Sec)
0
Distributed Kmeans Clusters
Weighted Distribut...
Extension Hierarchical Clusters
Technique Name
🟦 Distributed Kmeans Clusters 🟩 Weighted Distributed Uncertain Kmeans Clusters 🟥 Extension Hierarchical Clusters

Fig. 19 Execution Time result UI

Figure. 19 displays the execution time of the three algorithms for one data set.

Execution time comparison for all the three algorithms (DK-means, WDK-means, modified HWC) applied on 7 different datasets is shown in Figure 20-22.

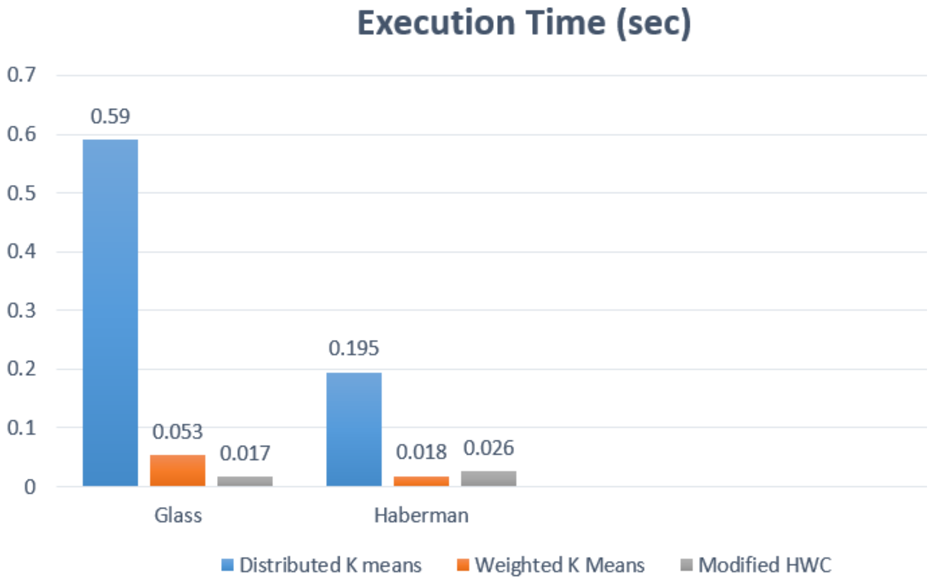


Fig. 20 Execution time for Glass & Haberman dataset

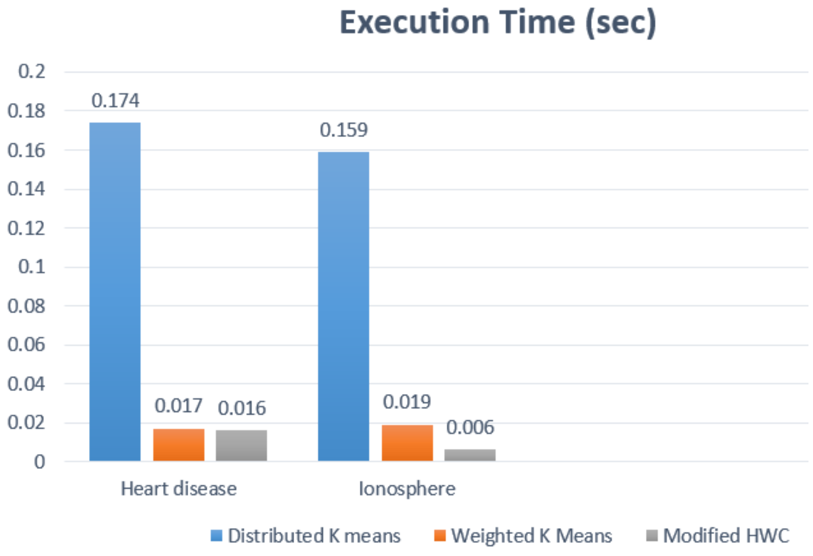


Fig. 21 Execution time for Heart & Ionosphere dataset

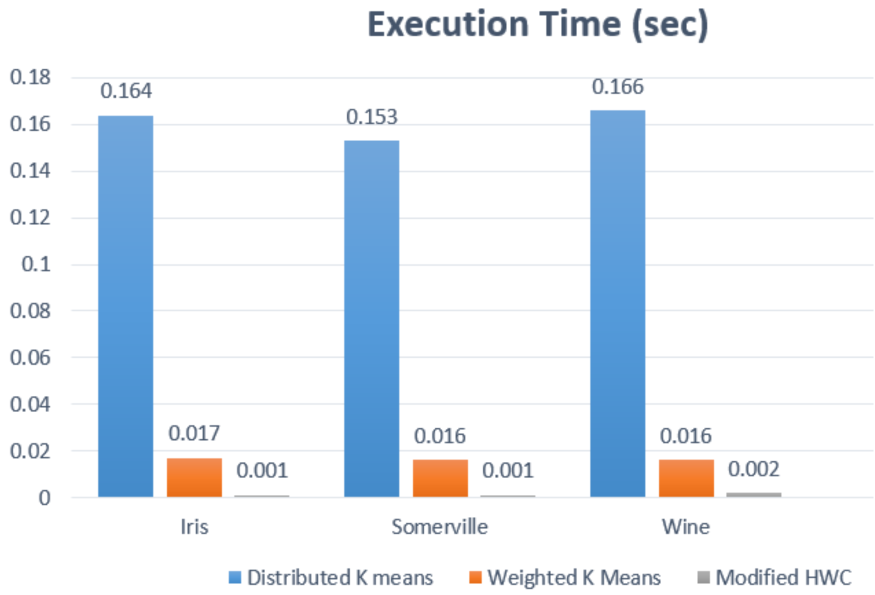


Fig. 22 Execution time for Iris, Somerville & Wine dataset

The results are listed in Table 2. As we can see, in most of the cases, Weighted Distributed K-means and Modified Hierarchical Weighted clustering algorithm outperformed Distributed K-means algorithm. The performance of the proposed algorithm is worth noting. Its Execution Time is significantly lower than other two algorithms. In above graphs Modified Hierarchical Clustering is better in performance as it took less execution time. The big advantage is Modified HWC does not require any prior K value.

However, Modified HWC may take more execution time as it gives full freeness to data. Thus number of clusters will increase and the execution time may also increase. We can see this in the case of Haberman dataset.

Table 2: Result comparison

Dataset	Distributed K means	Weighted K Means	Modified HWC
Glass	0.59	0.053	0.017
Haberman	0.195	0.018	0.026
Heart disease	0.174	0.017	0.016
Ionosphere	0.159	0.019	0.006

Iris	0.164	0.017	0.001
Somerville	0.153	0.016	0.001
Wine	0.166	0.016	0.002

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

This paper focuses on the problem of distributed data clustering for uncertain data and proposes a modified hierarchical algorithm in peer to peer network. The distributed algorithm is performed for improving the execution time. At each peer clustering is done by teaming up with its topological peers. Experiments conducted in this paper are tested several real world datasets. The proposed algorithm shows really good results. Performance of modified hierarchical weighted clustering is worth noting. It performed better than distributed k-means and weighted distributed k-means. The advantage of applying modified hierarchical weighted clustering is that it does not require prior k value, which improves the execution time of clustering process. However, from the results we can see that wdk-means clustering shows the similar performance as the modified hwc, while distributed k-means performed worst. But wdk-means clustering have a limitation of specifying number of clusters prior the start of the algorithm, which modified hwc overcomes.

The results of this work provides a good direction of future work. As we saw, modified hwc may take more time because it grants freedom to the data to adjust itself, which may increase number of clusters, and execution time. It can be a good direction of future work. This work can also be extended to different dataset with soft computing algorithms like particle swarm optimization.

REFERENCES

[1] K. M. Hammouda and M. S. Kamel, "Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 681-698, May 2009, doi: 10.1109/TKDE.2008.189.

[2] S. Datta, C. Giannella and H. Kargupta, "Approximate Distributed K-Means Clustering over a Peer-to-Peer Network," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 10, pp. 1372-1388, Oct. 2009, doi: 10.1109/TKDE.2008.222.

[3] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in Proc. 10th Pacific–Asia Conf.Knowl. Discovery Data Mining, vol. 3918. Apr. 2006, pp. 199–204.

[4] Zhou, Jin & Chen, Long & Chen, C. & Wang, Yingxu & Li, Han-Xiong. (2017). Uncertain Data Clustering in Distributed Peer-to-Peer Networks. IEEE Transactions on Neural Networks and Learning Systems. PP. 1-15. 10.1109/TNNLS.2017.2677093..

[5] Ajit Patil and M.D. Ingle , "Clustering on Uncertain Data using Kullback Leibler Divergence Measurement based on Probability Distribution", International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161 Pune. Vol.5, No.4 (Aug 2015).

[6] Azimi, Rasool & Sajedi, Hedieh & Ghayekhloo, Mohadeseh. (2016). A Distributed Data Clustering Algorithm in P2P Networks. Applied Soft Computing. 51. 10.1016/j.asoc.2016.11.045.

[7] Zhu, Youwen & Li, Xingxin. (2020). Privacy-preserving k-means clustering with local synchronization in peer-to-peer networks. Peer-to-Peer Networking and Applications. 13. 10.1007/s12083-020-00881-x.

[8] Azimi, Rasool & Sajedi, Hedieh. (2014). Distributed Data Clustering in Peer-to-Peer Networks: A Technical Review.

[9] Korda, Nathan & Szorenyi, Balazs & Li, Shuai. (2016). Distributed Clustering of Linear Bandits in Peer to Peer Networks.

[10] Wang, Ting & Wang, Rongrong & Zhou, Jin & Jiang, Hui & Han, Shiyuan & Wang, Lin & Chen, Yuehui. (2020). A Collaborative Kernel Clustering Algorithm for Non-Linear Data in Peer-to-Peer Networks. 849-853. 10.1109/ICCSS52145.2020.9336887.

[11] A. D. Sarma, O. Benjelloun, A. Halevy, S. Nabar, and J. Widom, "Representing uncertain data: Models, properties, and algorithms," VLDB J., vol. 18, no. 5, pp. 989–1019, May 2009.

[12] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, Sep. 2013.