

Artificial Intelligence Project Report Sec E

Submission Date: 09/May/2020

Submitted to: Dr. Fahad Sherwani

Group members:

K173922 Abdullah Qadri

K173607 Hassan Ahmed

K173940 Muhammad Owais

Topic

Agglomerative Hierarchical Clustering

**Abstract:**

Clustering is a task of assigning a set of objects into groups called clusters. In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Motivation for clustering in general, covering hierarchical clustering and applications, includes the following: analysis of data; interactive user interfaces; storage and retrieval; and pattern recognition.

**Introduction:**

Agglomerative hierarchical clustering has been the dominant approach to constructing

Embedded classification schemes. Agglomerative hierarchical clustering algorithms can be characterized as *greedy*, in the algorithmic sense. A sequence of irreversible algorithm steps is used to construct the desired data structure. Assume that a pair of clusters, including possibly singletons, is merged or agglomerated at each step of the algorithm. Then the following are equivalent views of the same output structure constructed on n objects: a set of n−1 partitions, starting with the fine partition consisting of n classes and ending with the trivial partition consisting of just one class, the entire object set; a binary tree (one or two child nodes at each non-terminal node) commonly referred to as a dendrogram; a partially ordered set (poset)

which is a subset of the power set of the n objects; and an ultra-metric topology on the n objects. A wide range of agglomerative hierarchical clustering algorithms have been proposed at one time or another. Such hierarchical algorithms may be conveniently broken down into two groups of methods. The first group is that of linkage methods – the single, complete, weighted and unweighted average linkage methods. These are methods for which a graph representation can be used. Sneath and Sokal (1973) may be consulted for many other graph representations of the stages in the construction of hierarchical clustering’s. The second group of hierarchical clustering methods are methods which allow the cluster centers to be specified (as an average or a weighted average of the member vectors of the cluster). These methods include the centroid, median and minimum variance methods.

**Related work:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Research Paper Name** | **Author/s** | **Source** |
| **1** | Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results | * Ian Davidson * S. S. Ravi | * SpringerLink |
| 2 | Efficient agglomerative hierarchical clustering | [Athman Bouguettaya](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150" \l "!)  [a](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150" \l "!)[Qi Yu](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150" \l "!)[b](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150" \l "!)  [XuminLiub](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150#!)  [Xiangmin Zhouc](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150#!)  [AndySong](https://www.sciencedirect.com/science/article/abs/pii/S0957417414006150#!) | ScienceDirect |
| 3 | Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms | [Alberto Fernández](https://link.springer.com/article/10.1007/s00357-008-9004-x#auth-1)  [Sergio Gómez](https://link.springer.com/article/10.1007/s00357-008-9004-x#auth-2) | SpringerLink |
| **4** | Algorithms for Model-Based Gaussian Hierarchical Clustering | [Chris Fraley](https://epubs.siam.org/author/Fraley%2C+Chris) | (SIAM) Society for industrial and applied mathematics |
| **5** | Agglomerative Hierarchical Clustering Algorithm | K.Sasirekha  P.Baby | International Journal of Scientific and Research Publications |

Of all the research papers mentioned above, all the clustering is being done through different methods, the publications dates of all the papers differ from each other therefore the result or data might also differ from each other.

In general, hierarchic. Bisect k which is a clustering algorithm outperforms partitioned clustering algorithm in terms of cluster quality. But the computational cost (space and time complexity) of agglomerative and divisive hierarchical clustering is very high compared to that of partitioned clustering. Using the partitioned clustering algorithms for generating the hierarchical tree is quite popular in recent years. Experiments have shown that these hybrids partitioned clustering algorithm performs better than the traditional clustering algorithms. i.e Bisect-K, a hybrid partitioned clustering algorithm that uses K-means partitioned clustering algorithm (2-way) to split the clusters in each step to construct the hierarchical tree in a top-down approach

**Dataset and Feature:**

**Stored data approach:**

Step 1: Examine all interposing dissimilarities, and form cluster from two closest

points.

Step 2: Replace two points clustered by representative point (center of gravity)

or by cluster fragment.

Step 3: Return to step 1, treating clusters as well as remaining objects, until

all objects are in one cluster.

**NN-chain algorithm**

Step 1: Select a point arbitrarily.

Step 2: Grow the NN-chain from this point until a pair of RNNs is obtained.

Step 3: Agglomerate these points (replacing with a cluster point, or updating

the dissimilarity matrix).

Step 4 :From the point which preceded the RNNs (or from any other arbitrary

point if the first two points chosen in steps 1 and 2 constituted a pair of

RNNs), return to step 2 until only one point remains.

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion). Divisive Hierarchical Clustering

 A top-down clustering method and is less commonly used. It works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain. GeneLinker™ does not support divisive hierarchical clustering.

**Disadvantages:**

 No provision can be made for a relocation of objects that may have been 'incorrectly' grouped at an early stage. The result should be examined closely to ensure it makes sense.

 Use of different distance metrics for measuring distances between clusters may generate different results. Performing multiple experiments and comparing the results is recommended to support the veracity of the original results.

Before clustering comes the phase of data measurement, or measurement of the

observables. Let us look at some important considerations to be taken into

account. These considerations relate to the metric or other spatial embedding,

comprising the first phase of the data analysis *stricto sensu*.

To group data we need a way to measure the elements and their distances

relative to each other in order to decide which elements belong to a group. This

can be a similarity, although on many occasions a dissimilarity measurement,

or a “stronger” distance, is used.

A distance between any pair of vectors or points i, j, k satisfies the properties

of: symmetry, d(i, j) = d(j, k); positive definiteness, d(i, j) > 0 and d(i, j) = 0

iff i = j; and the triangular inequality, d(i, j) ≤ d(i, k)+d(k, j). If the triangular

inequality is not taken into account, we have a dissimilarity. Finally a similarity

is given by s(i, j) = maxi,j{d(i, j)} − d(i, j).

When working in a vector space a traditional way to measure distances is a

Minkowski distance, which is a family of metrics defined as follows:

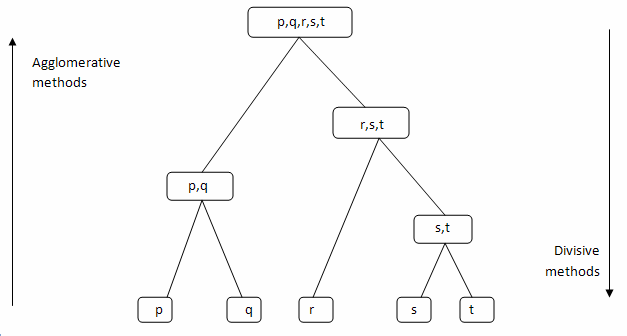
Lp(xa, xb) = (nXi=1|xi,a − xi,b|p)1/p; ∀ p ≥ 1, p ∈ Z+, (1)

where Z+ is the set of positive integers.

The Manhattan, Euclidean and Chebyshev distances (the latter is also called

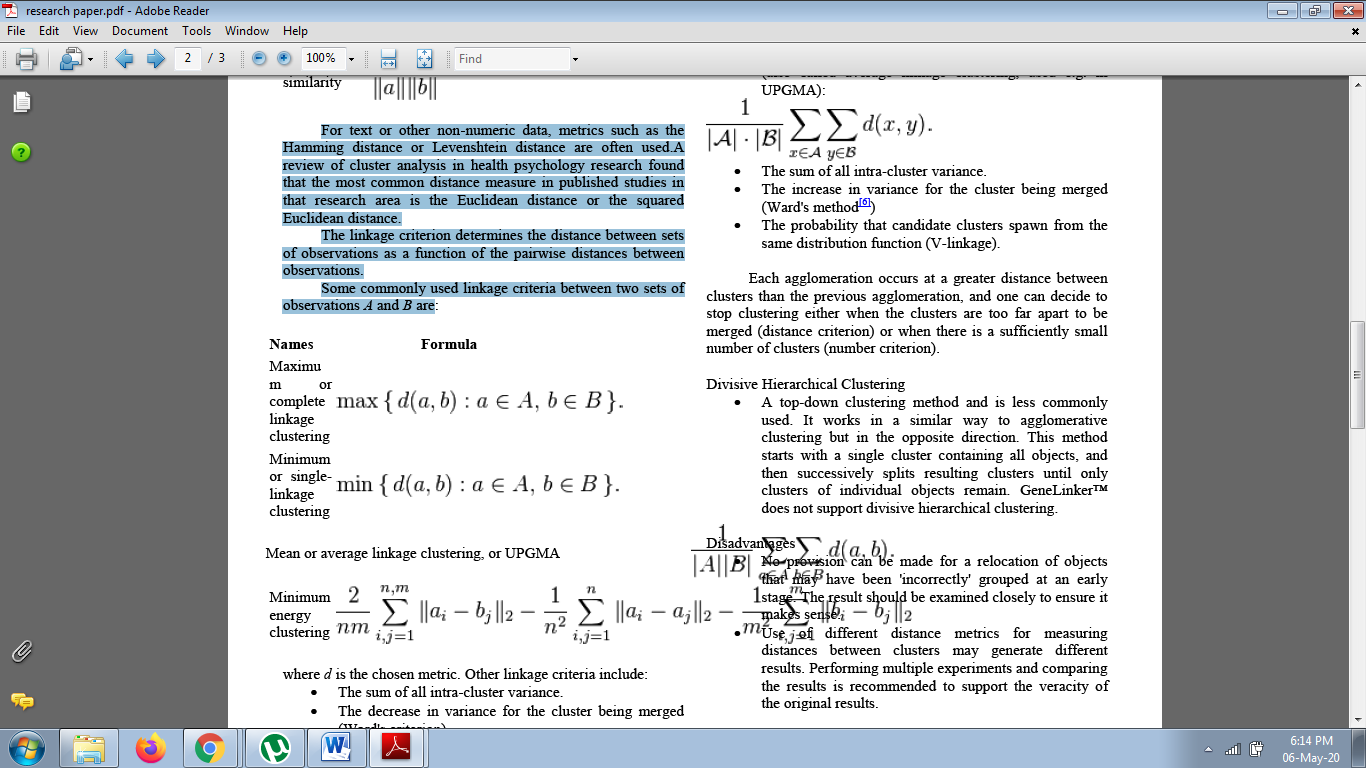
maximum distance) are special cases of the Minkowski distance when p = 1, p =

2 and p → ∞.

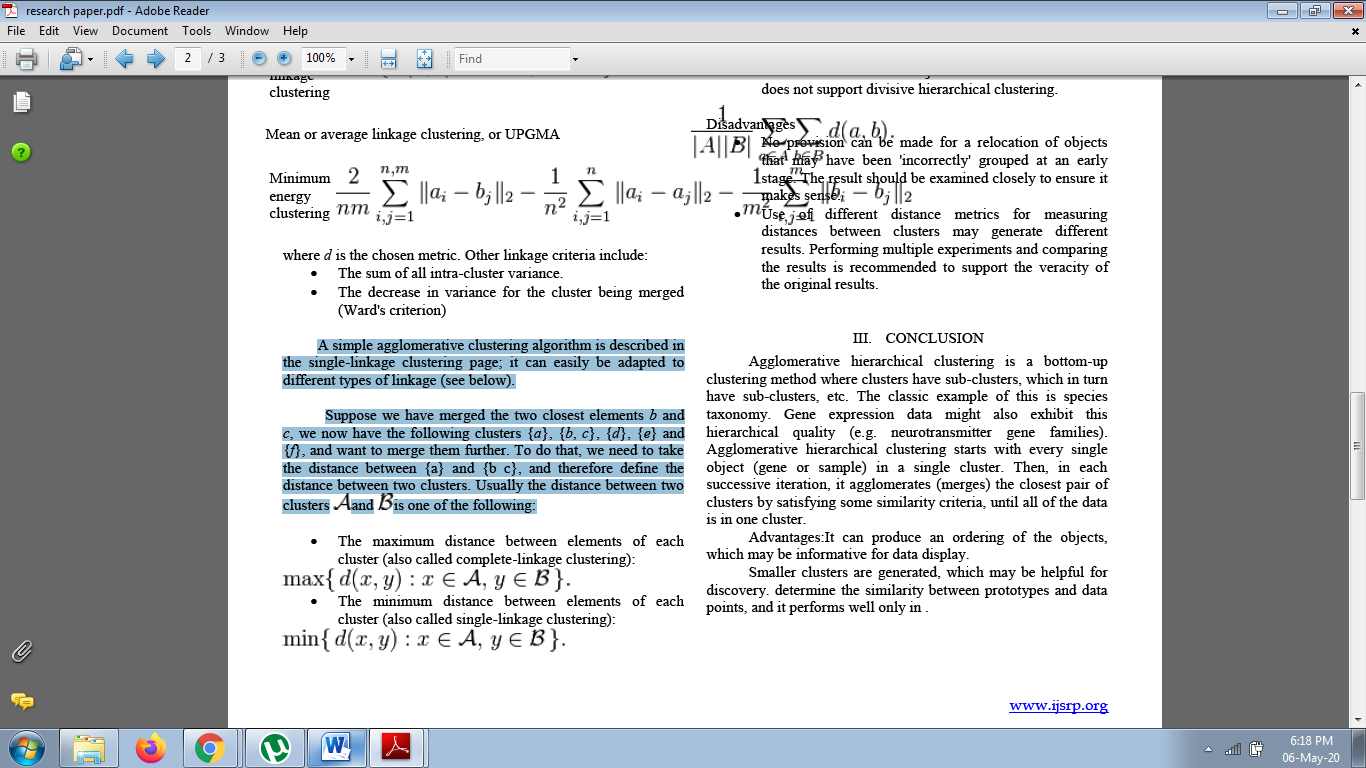


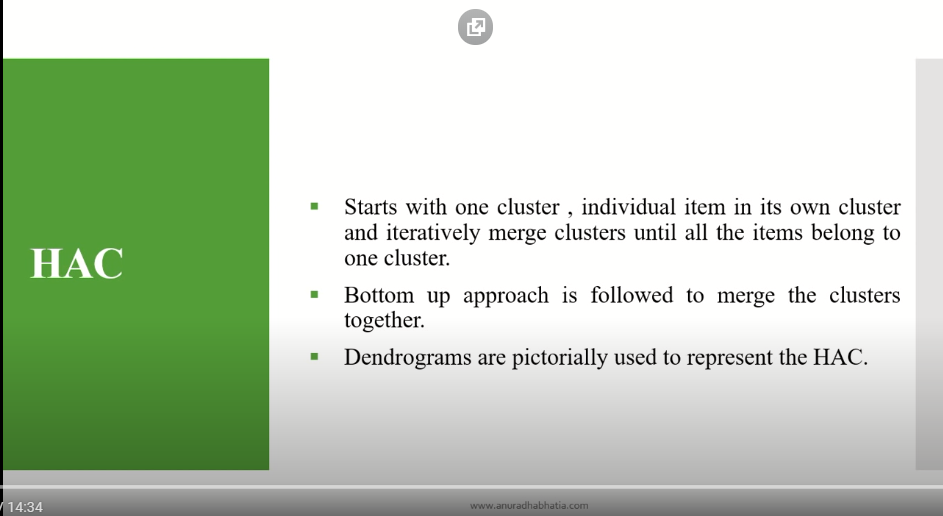
**Methods:**

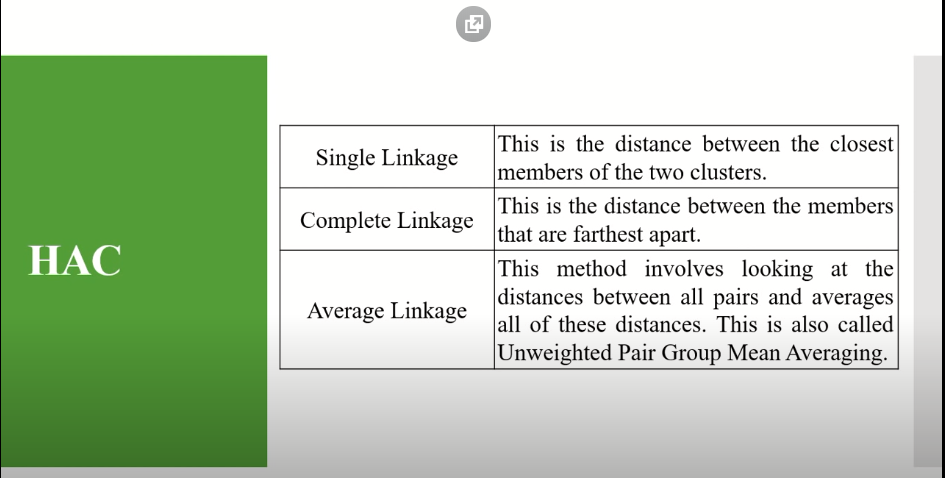
For text or other non-numeric data, metrics such as the Hamming distance or Levenshtein distance are often used.A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance. The linkage criterion determines the distance between sets of observations as a function of the pairwise distances between observations. Some commonly used linkage criteria between two sets of observations *A* and *B* are:

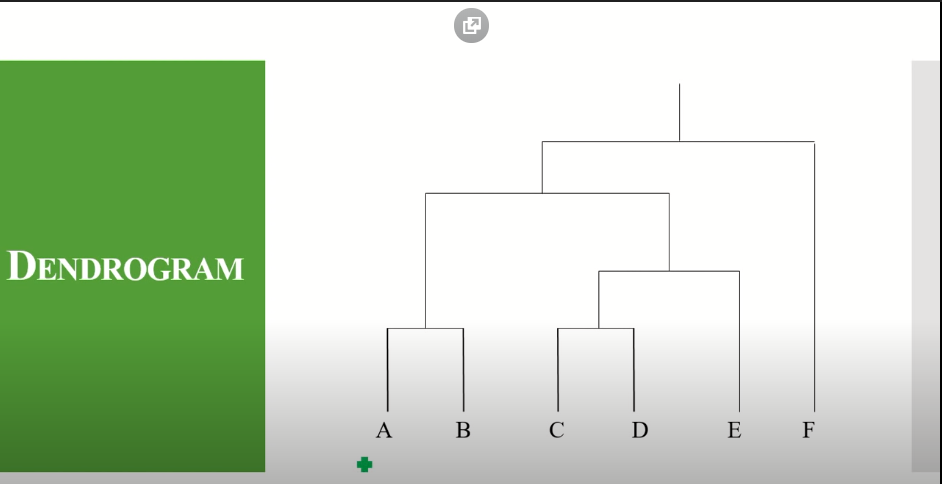


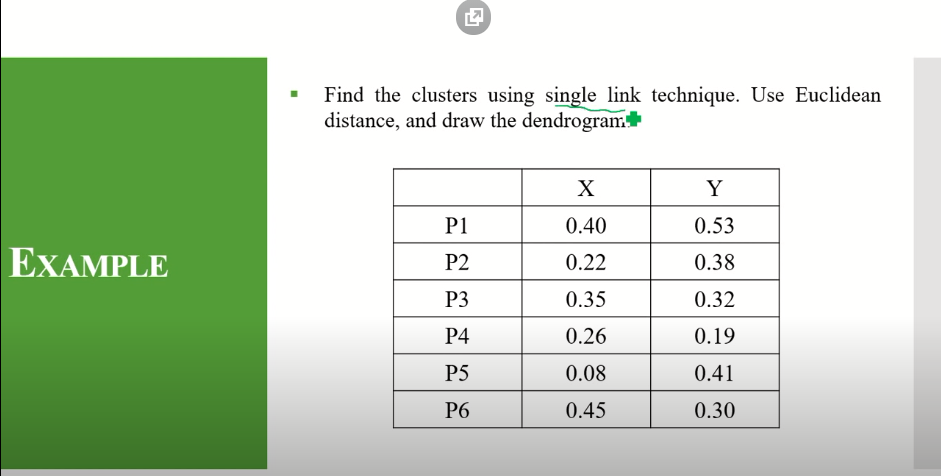
A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage (see below). Suppose we have merged the two closest elements *b* and *c*, we now have the following clusters {*a*}, {*b*, *c*}, {*d*}, {*e*} and {*f*}, and want to merge them further. To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters and is one of the following

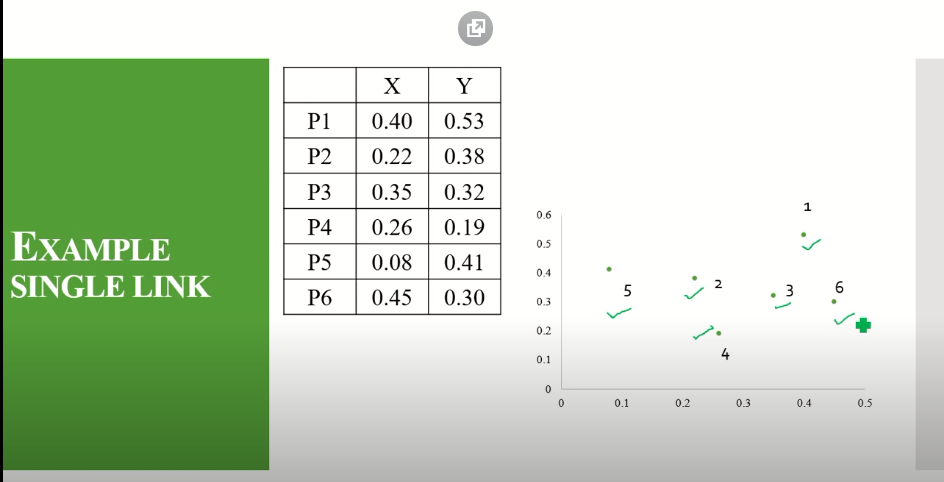


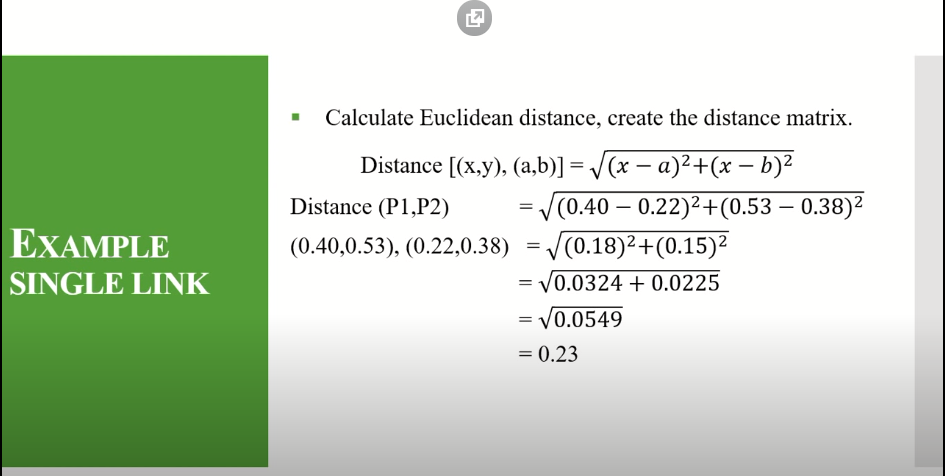
****

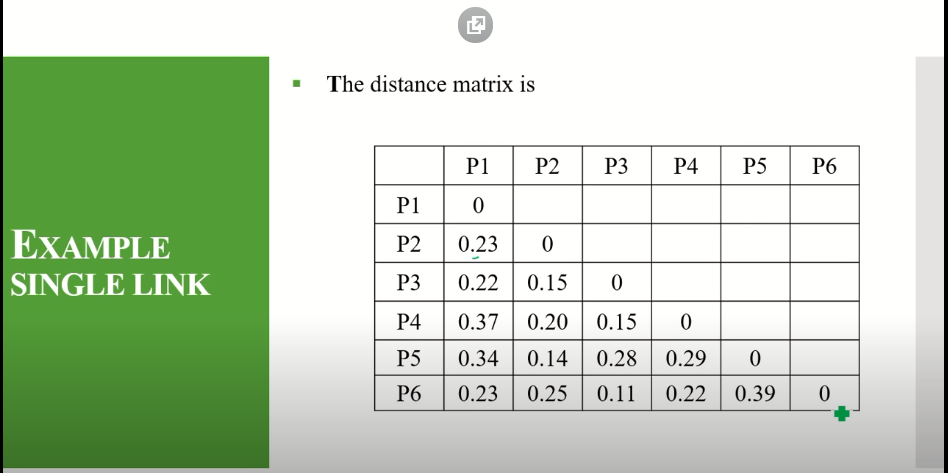
****

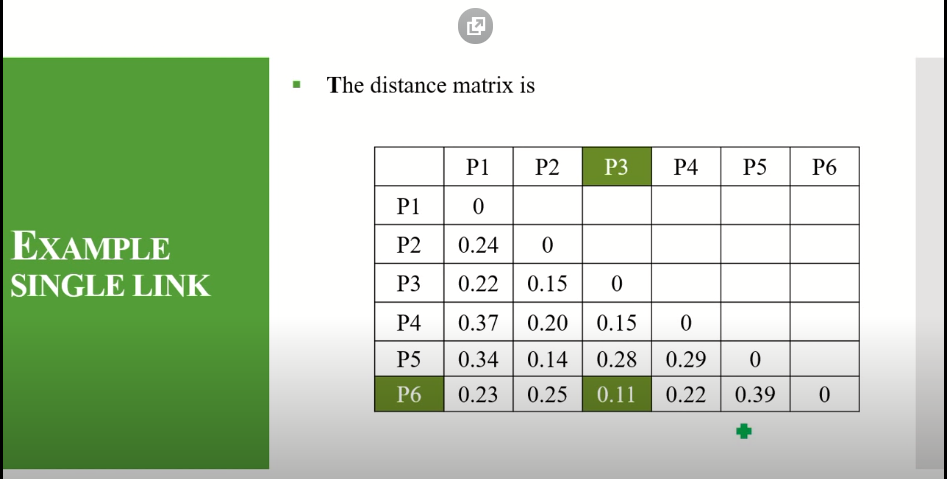
****

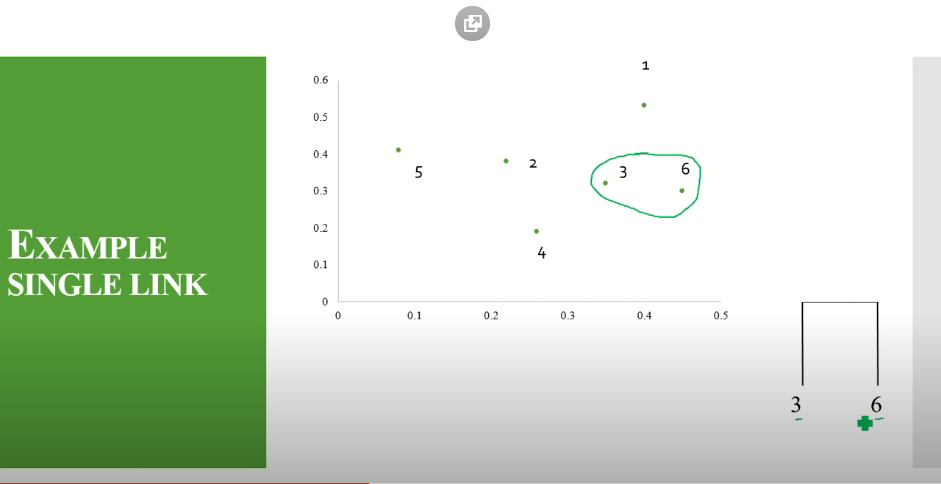
****

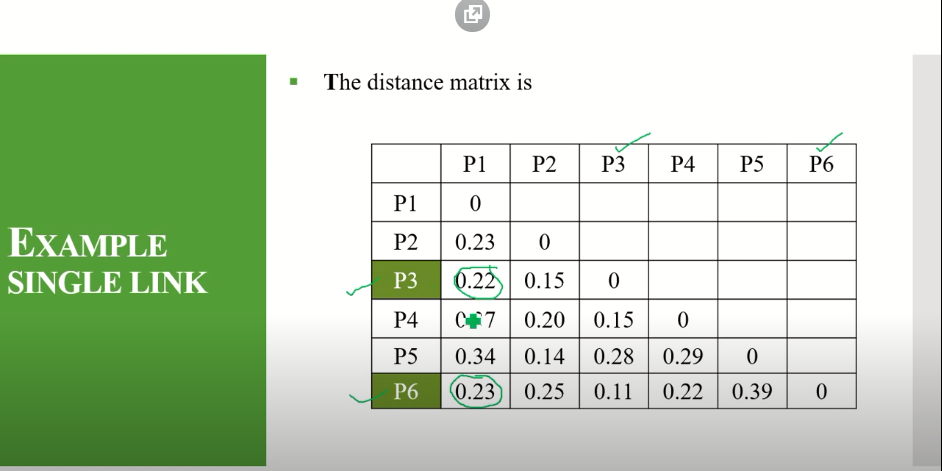
****

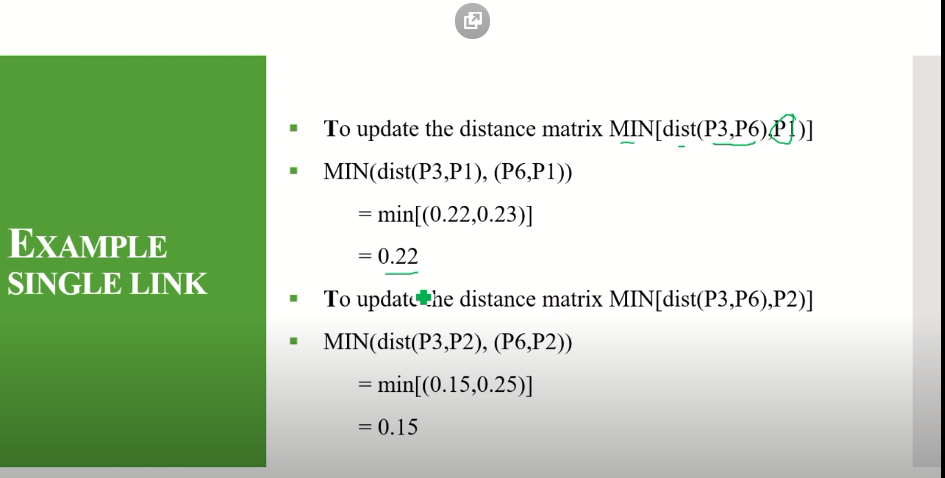
****

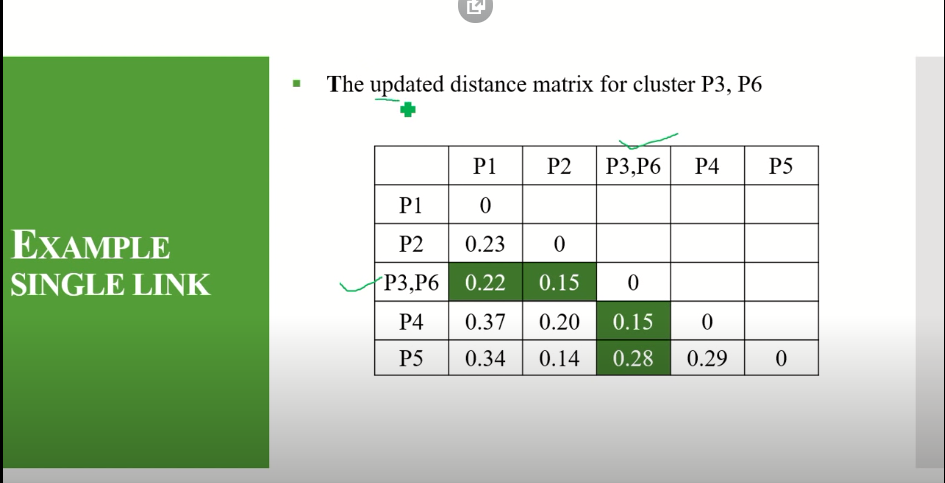
****

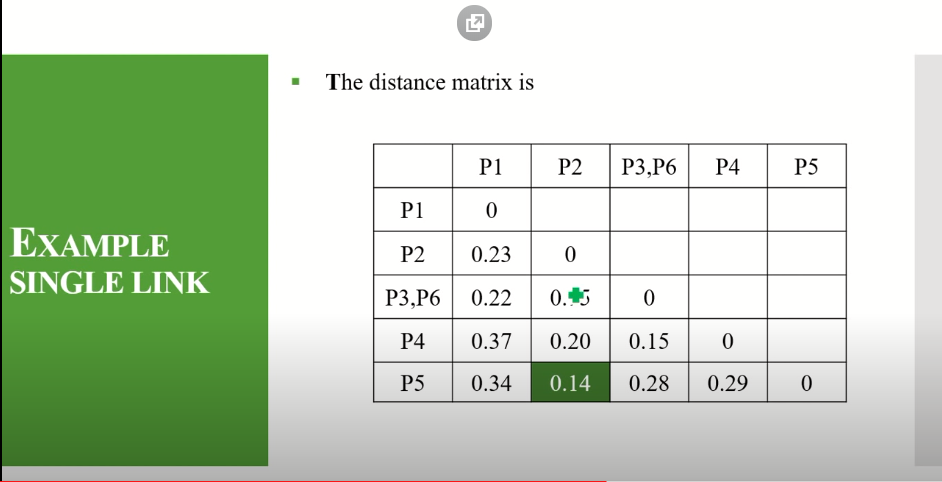
****

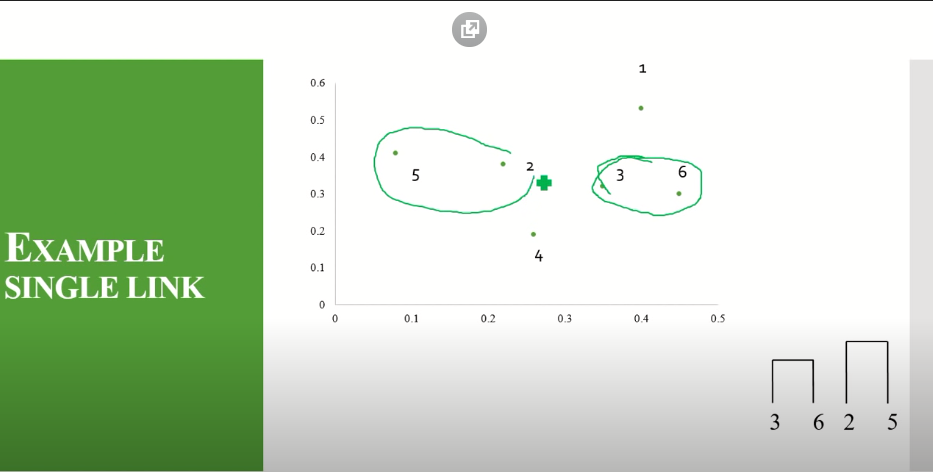
****

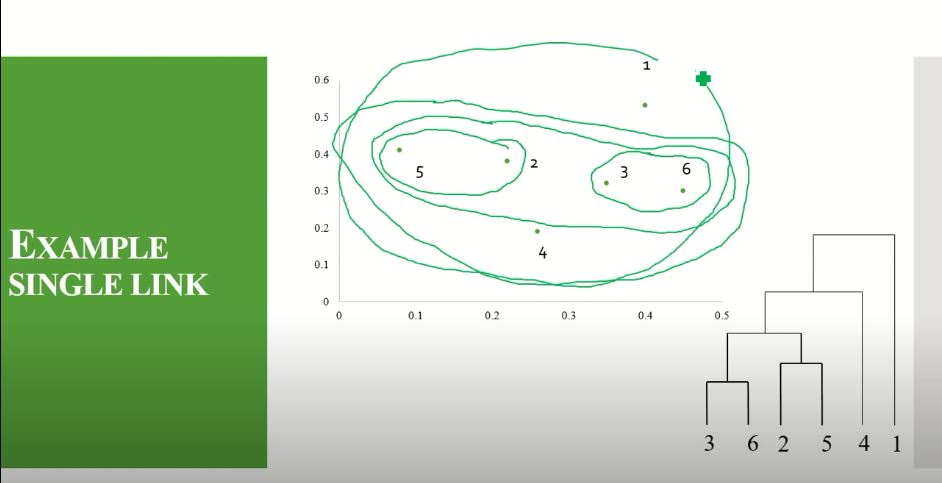
****

****

****

****

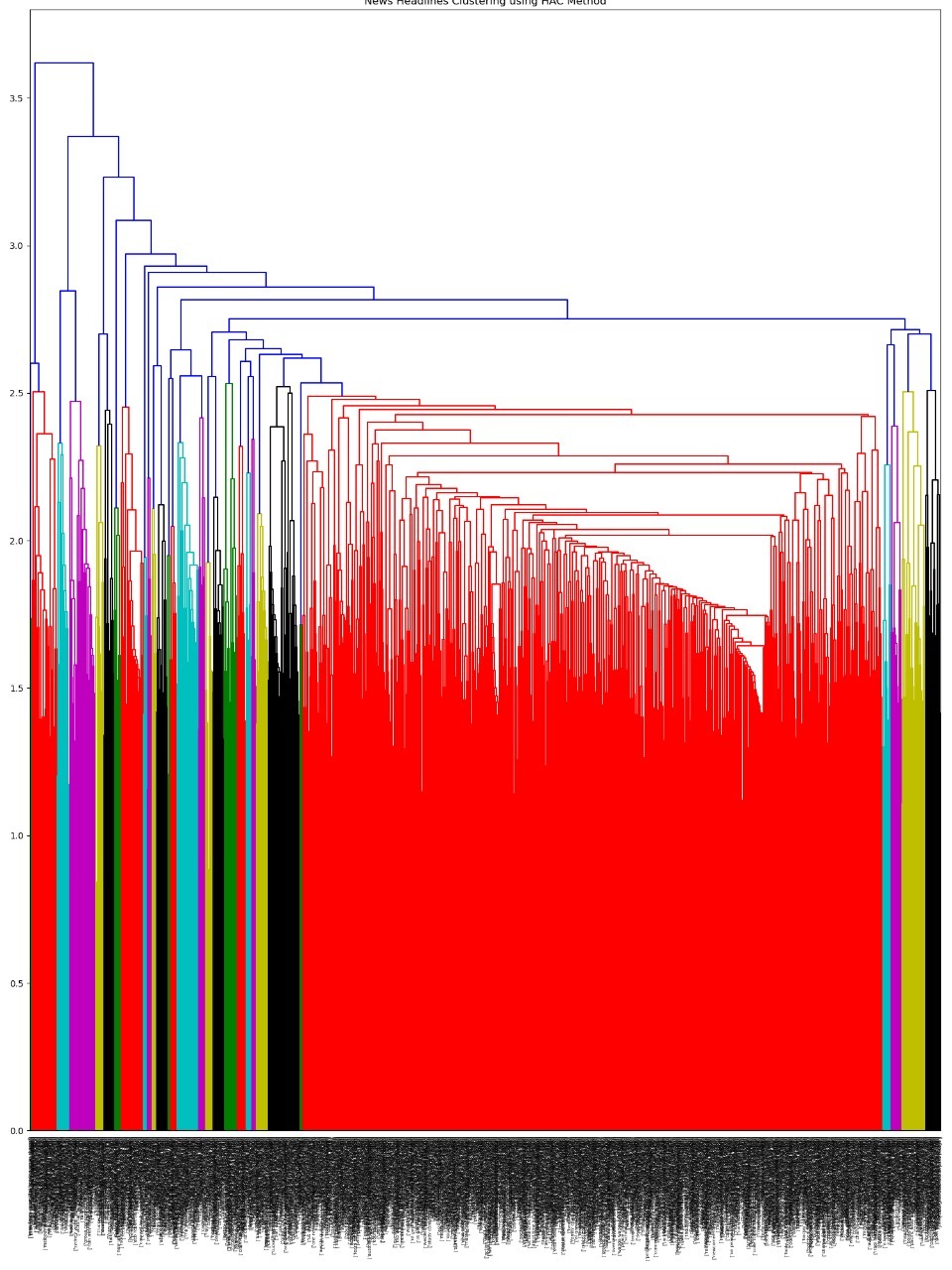
****

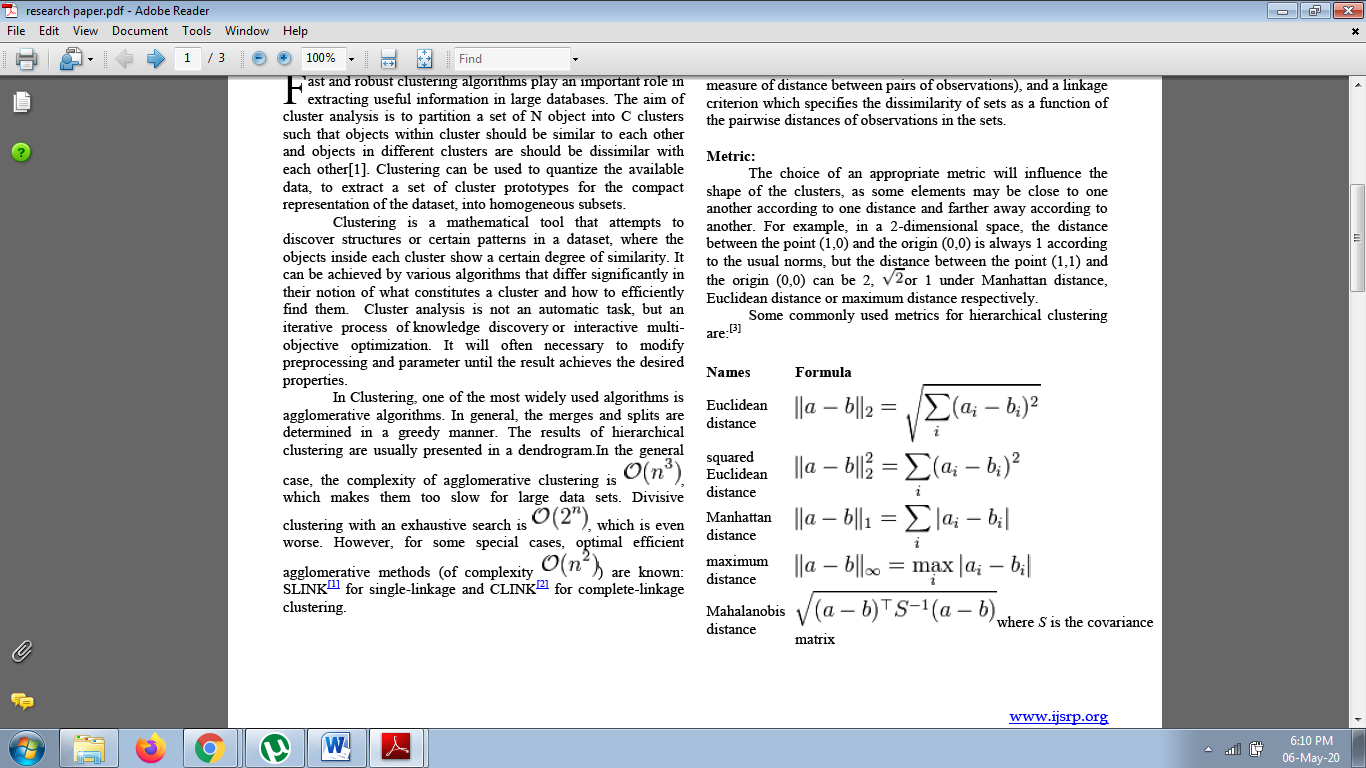
****

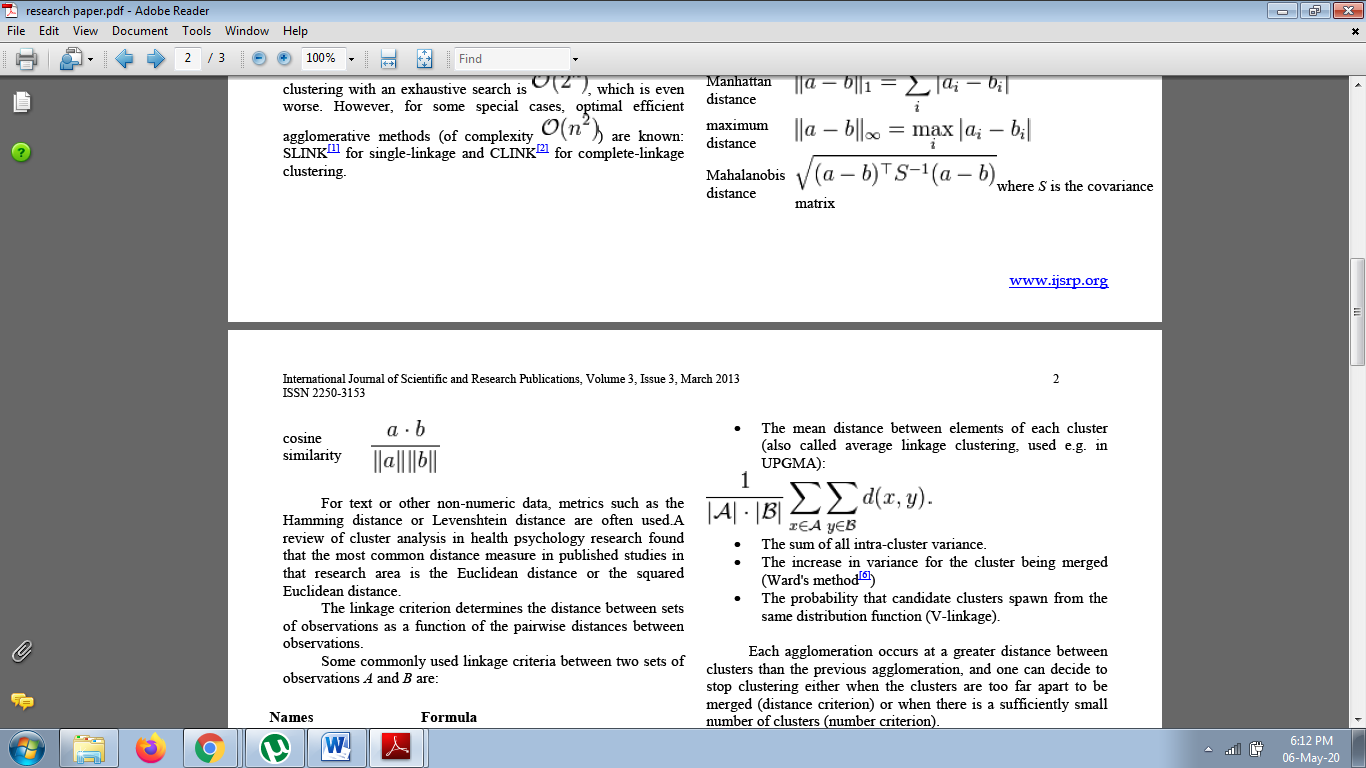
**Experiment/Results/Discussion:**

**Metric:** The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (1,0) and the origin (0,0) is always 1 according to the usual norms, but the distance between the point (1,1) and the origin (0,0) can be 2, or 1 under Manhattan distance, Euclidean distance or maximum distance respectively. Some commonly used metrics for hierarchical clustering are:

**Dendogram:**







**Density and Grid-Based Clustering Techniques**

Many modern clustering techniques focus on large data sets. In Xu and Wunsch

(2008, p. 215) these are classified as follows:

• Random sampling

• Data condensation

• Density-based approaches

• Grid-based approaches

• Divide and conquer

• Incremental learning

we select density and grid based approaches, i.e., methods that either look for data densities or split the data space into cells when looking for groups. In this section we take a look at these

two families of methods. The main idea is to use a grid-like structure to split the information space,

separating the dense grid regions from the less dense ones to form groups. In general, a typical approach within this category will consist of the following steps as presented by Grabusts and Borisov (2002):

1. Creating a grid structure, i.e. partitioning the data space into a finite

number of non-overlapping cells.

2. Calculating the cell density for each cell.

3. Sorting of the cells according to their densities.

4. Identifying cluster centers.

5. Traversal of neighbor cells.

**Conclusions & Future work**

Agglomerative various leveled grouping is a base up bunching strategy where groups have sub-groups, which thus have sub-bunches, and so on. The exemplary case of this is species scientific classification. Quality articulation information may likewise display this various leveled quality (for example synapse quality families). Agglomerative various leveled grouping begins with each and every object (quality or test) in a solitary bunch. At that point, in each progressive cycle, it agglomerates (blends) the nearest pair of bunches by fulfilling some comparability rules, until the entirety of the information is in one bunch.

**Advantages:**

* It can deliver a requesting of the items, which might be educational for information show.
* Littler bunches are created, which might be useful for disclosure. decide the likeness among models and information focuses, and it performs well just in.

**Future work:**

This paper was intended to analyze between two algorithms. Through my broad inquiry I couldn't discover any study that endeavors to compare between all algorithms under examination. As a future work correlation between these algorithms can be endeavored by various factors other than those considered in this paper. Comparing between the aftereffects of calculations utilizing standardized information or non-standardizes information will give diverse results. Of course standardization will influence the execution of the calculation and nature of the outcomes. Another methodology may consider utilizing information bunching calculations in applications, for example, item and character acknowledgment or data recovery which is worried about programmed archives.

**Appendices:**

# 1) Single chaining

# 2) Ward method for cluster generation.

# 3) Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using dendrogram

**Contributions:**

We divided the project work into three parts so each one could have an equal share in this project.

1. Hassan Ahmed: Research Work
2. Abdullah Qadri: Implementation of Algorithm
3. Muhammad Owais: Preprocessing of dataset

**References:**

1. M.S.Yang,” A Survey of hierarchical clustering” Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.
2. vathy-Fogarassy, B. Feil, J. Abonyi”Minimal Spanning Tree based clustering” Proceedings of World academy of Sc., Eng & Technology, vol- 8, Oct-2005, 7-12.
3. Pal N.R, Pal K, Keller J.M. and Bezdek J.C, “A Possibilistic Clustering Algorithm”, IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517– 530, 2005.
4. R. Krishnapuram amd J.M. Keller, “A possibilistic approach to clustering”, IEEE Trans. Fuzzy Systems, Vol. 1, Pp. 98-110, 1993.
5. J. C. Dunn (1973): "A Agglomerative Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57
6. <https://en.wikipedia.org/wiki/Hierarchical_clustering>
7. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>
8. https://scikitlearn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

* **Libraries Included**

In order for the completion of the specific project we also needed to download a few additional libraries which are mentioned below:

* SK learn
* Numpy
* Matplot
* NLKT

# \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*