SECP2753 – DATA MINING

SEMESTER 2 2021/2022

GROUP 2

TOPIC 4 - CLUSTERING

Members:

1. **GROUP LEADER**: NAYLI NABIHAH JASNI (A20EC0105)
2. MIKHEL ADAM BIN MUHAMMAD EZRIN (A20EC0237)
3. MUHAMMAD DINIE HAZIM BIN AZALI (A20EC0084)
4. MADINA SURAYA BINTI ZHARIN (A20EC0203)

TABLE OF CONTENT

1.0 Definition

1.1 Cluster

A data cluster is a subset of a larger dataset in which each data point is closer to the cluster center than to other cluster centers in the dataset, as determined via cluster analysis, which involves iteratively decreasing squared distances (Glenn, 2022).

1.2 Cluster Analysis

Cluster analysis is often one of the first steps in the analysis of data, as such, it is an effort at unsupervised learning usually in the context of very little a priori knowledge (Davies et al.,1979).

Cluster analysis is the process of creating data clusters by minimizing the distance between data points and a reference (Glenn, 2022).

1.3 Clustering

Clustering is one of the common methods of unsupervised learning in which data is segmented based on the similarity. Clustering is assigning a set of instances to subgroups class clusters so that each cluster is very similar with other clusters (Teng et al., 2022).

Clustering identifies groups of related records that can be used as a starting point for exploring further relationships. It is the first step in data mining analysis (Bijuraj, 2013).

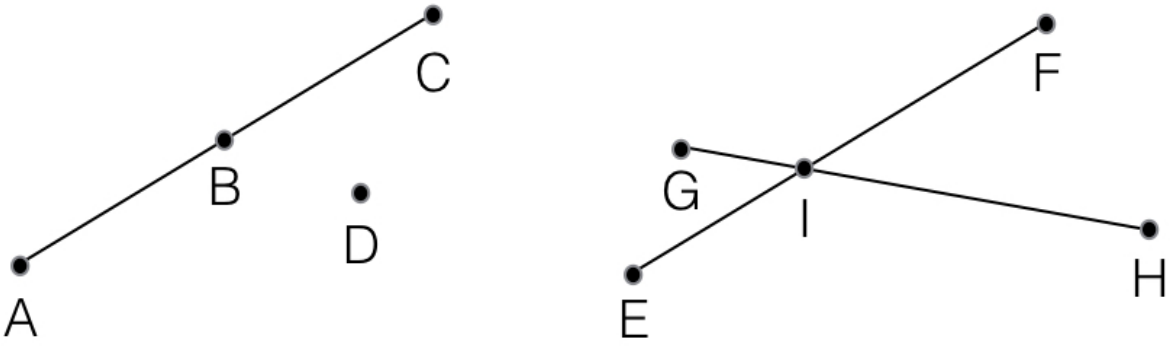
2.0 What is Similarity

2.1 Definition

Similarity, also known as **distance measures,** are core components of distance-based clustering algorithms used to group similar data points into the same clusters, while dissimilar or distant data points are placed into different clusters. The distance between various data points can be defined as a similarity measure. Similarity is an amount that reflects the strength of relationship between two data items. Meanwhile, dissimilarity is the measurement of divergence between two data items.

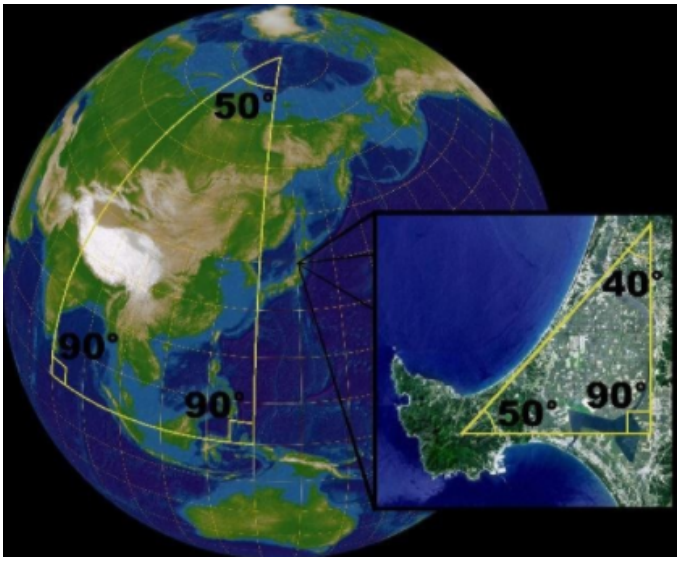
2.2 Euclidean

Euclidean geometry is the study of geometrical shapes (plane and solid) and figures based on different axioms and theorems. It is basically introduced for flat surfaces or plane surfaces. For geometrical problems, Euclidean distance is considered the standard metric. It's simply the ordinary distance between two points. Euclidean distance is also extensively used in clustering problems. It is also the default distance measure used by the K-means algorithm. The Euclidean distance determines the root of square differences between the coordinates of a pair of objects. The most basic terms of geometry are a point, a line, and a plane. A point has no dimension (length or width), but it does have a location. A line is straight and extends infinitely in the opposite directions. A plane is a flat surface that extends indefinitely.



2.3 Non-Euclidean

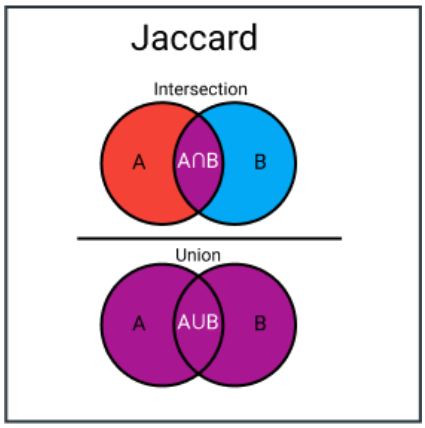
Non-Euclidean spaces are basically any geometry that is not the same as Euclidean geometry. An example of Non-Euclidean geometry can be seen by drawing lines on a sphere, that is straight lines that are parallel at the equator can meet at the poles. This “triangle” has an angle sum of 90 + 90 + 50 = 230 degrees.



For non-Euclidean spaces, there are several distance measures used. Some of them include Jaccard distance, Hamming distance, and Manhattan distance.

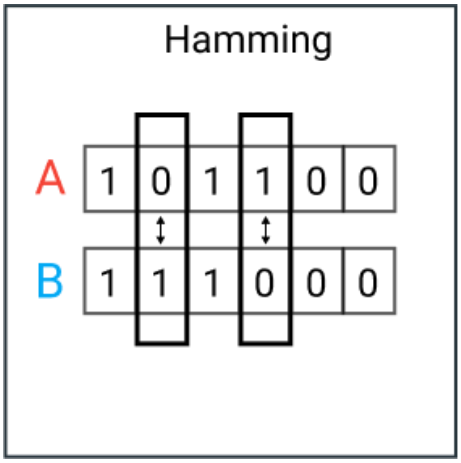
2.3.1 Jaccard Distance (Jaccard Index)

The Jaccard Index (also known as Intersection over Union) is a metric for calculating sample set similarity and diversity. It is equal to the intersection size divided by the sample set union size. In practice, it is the total number of similar entities between sets divided by the total number of entities. For example, if two sets have 1 entity in common and there are 5 different entities in total, then the Jaccard index would be 1/5 = 0.2.



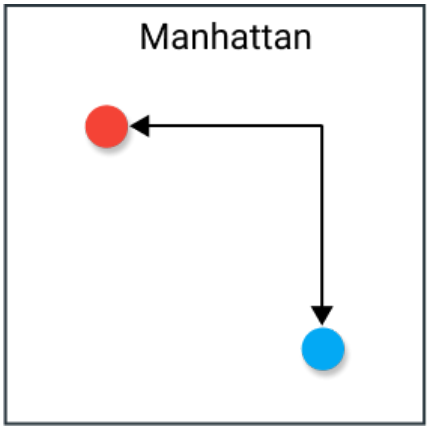
2.3.2 Hamming Distance

The Hamming distance is the difference in values between two vectors. It's usually used to compare two binary strings of the same length. It can also be used to compare the similarity of strings by calculating the number of characters that differ.



2.3.3 Manhattan Distance

The Manhattan distance, also known as the Taxicab distance or City Block distance, is a calculation that determines the distance between two real-valued vectors. Consider vectors that describe objects on a chessboard-like uniform grid. The Manhattan distance is the distance between two vectors if they could only move in the same direction. The distance is calculated without any diagonal movement.

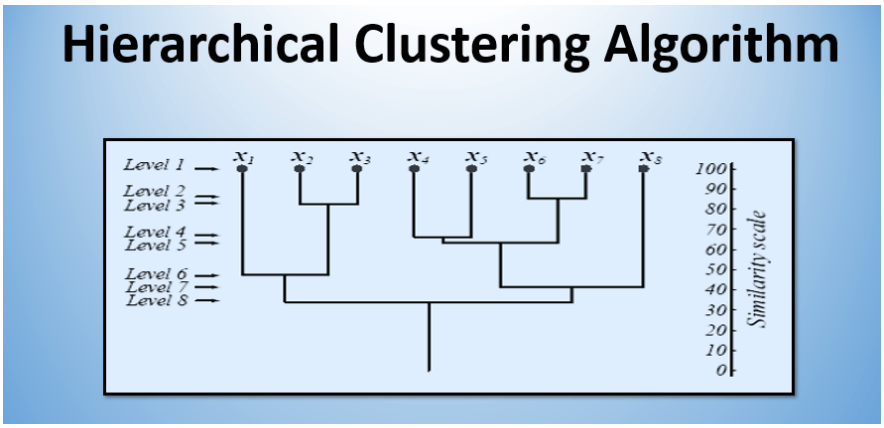

Manhattan

3.0 Applications of Clustering

Applications	Algorithms	Area
KNIME	- K-Means - K-Medoids - Hierarchical Clustering - Fuzzy c-Means - SOTA (self organizing tree algorithm)	- Voting patterns - Market segmentation
Orange	- K-Means - SOM (self organizing maps) - Hierarchical Clustering - MDS (multidimensional scaling)	- Anomaly detection - Medical imaging
RapidMiner Community Edition	- Hierarchical Clustering - Support Vector Clustering - Top-Down Clustering - K-Means - K-Medoids	- Social network analysis - Market segmentation
Tanagra	- K-Means - SOM - LVQ (Learning Vector Quantizers) - Hierarchical Clustering	- Biological models of neural systems
Weka	- DBSCAN - COBWEB - K-Means - EM (Expectation maximization)	- Economic - Recommendation engines

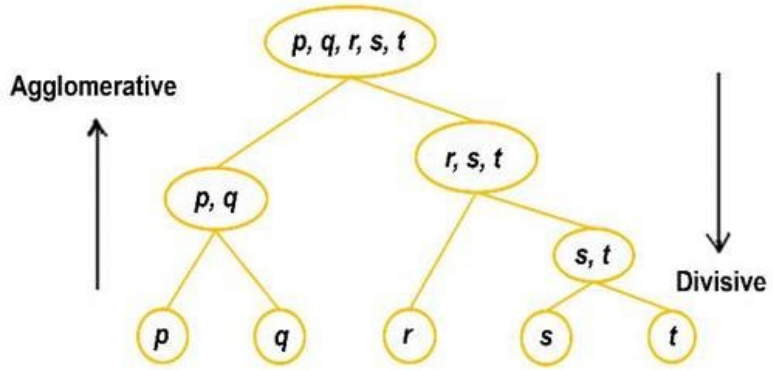
4.0 Types of Clustering

4.1 Hierarchical Algorithm



Hierarchical algorithm is an unsupervised Machine Learning technique. Its aim is to determine natural groupings based on the data's features. Hierarchical algorithm aims to find nested groups of data by constructing a hierarchy. It's similar to the plant or animal kingdom's biological taxonomy. The hierarchical tree known as a dendrogram is commonly used to describe hierarchical groups.

There are 2 types of Hierarchical algorithm:



a. Divisive

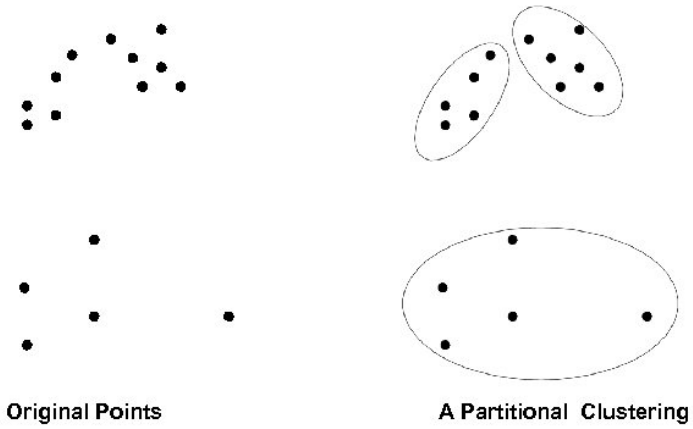
   This is a top-down technique, in which the full data set is first considered as a single group, and then subgroups are created repeatedly. When the number of clusters in a hierarchical clustering algorithm is known, the division process ends when that number is

reached. Otherwise, the procedure ends when the data can no longer be split, implying that the current iteration's subgroup is identical to the prior iteration's (one can also consider that the division stops when each data point is a cluster).

b. Agglomerative

It's a bottom-up technique that relies on cluster merging. The data is first divided into m singleton clusters (where m is the number of samples/data points). Iteratively, two clusters are merged into one, lowering the number of clusters in each iteration. When all clusters have been merged into one or the desired number of clusters has been reached, the process of merging clusters comes to an end.

4.2 Partitional Clustering



This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. The number of clusters that must be formed for the clustering procedures is specified by the data analysts.

When a database(D) contains multiple(N) objects, the partitioning method creates user-specified(K) data partitions, each of which represents a cluster and a specific region. Many algorithms fall within the partitioning method category, including K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications), and others.

5.0 Clustering Algorithms

Algorithm	Function	Advantages	Disadvantages	Field Area	Application
K-Means	Used to cluster numerical data attributes.	- Widely applied by variety of packages - Fast convergence when apply clustering to small datasets - Easy to implement	- Computationally expensive for large datasets (k will be large) - Occasionally, it will be hard to choose the initial value for the number of clusters (k) - Strong sensitivity to outliers	Marketing	Tableau R
Agglomerative	A type of hierarchical clustering used to group objects based on their similarity	- Easy to implement and understand - Able to produce an ordering of objects - No need to pre-specify the number of clusters	- It gives the best result only in some cases - Cannot undo what was done previously - The various distance metrics for measuring distances between clusters may produce different results.	Bioinformatics	XLSTAT Excel
DBSCAN	Used in density-based clustering	- Does not need to specify the number of clusters beforehand	- Sometimes, determining an appropriate distance of neighborhood (eps) is hard and requires domain	Bio-Medical	XLSTAT

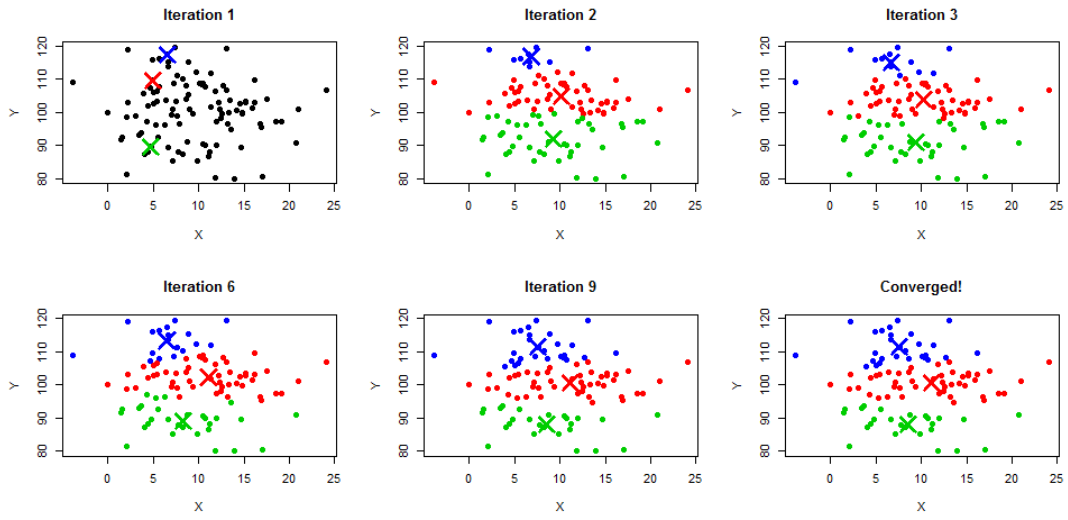
		- Performs well with arbitrary shapes clusters - Robust to outliers and can detect the outliers	knowledge - If clusters are very differ in terms of in-cluster densities, DBSCAN is not well suited to define the clusters		
Self-Organizing-Map (SOM)	Used to help to understand high dimensional data by reducing the dimensions of data to a map	- Data mapping is easily interpreted - Capable of organizing large and complex data sets	- Hard to determine what input weights to use - Mapping can resulting to divided clusters - Nearby points need to behave similarly	Economy	XLSTAT
Fuzzy C-Means (FCM)	Used to cluster multidimensional data by assigning each point a membership in each cluster center	- Gives out the best result for overlapped data compared to k-means - Low time complexity	- Sensitive to the initial values of k and p - Sensitive to outliers	Marketing	MATLAB

6.0 How does the K-Means clustering algorithm work?

K-means clustering uses "centroids", K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it. Then the process repeats: every point is assigned to its nearest centroid, centroids are moved to the average of points assigned to it. The algorithm is done when no point changes the assigned centroid. K-means clustering distinguishes itself from Hierarchical since it creates K random centroids scattered throughout the data (Robinson, 2022).

The algorithm will look like below:

a. Initialize K random centroids. You could pick K random data points and make those your starting points. Otherwise, you pick K random values for each variable.
b. For every data point, look at which centroid is nearest to it. Using some sort of measurement like Euclidean or Cosine distance.
c. Assign the data point to the nearest centroid.
d. For every centroid, move the centroid to the average of the points assigned to that centroid.
e. Repeat the last three steps until the centroid assignment no longer changes. The algorithm is said to have "converged" once there are no more changes.



1. Iteration 1 shows the random centroid centers.
2. Iteration 2 shows the new location of the centroid centers.
3. Iteration 3 has a handful more blue points as the centroids move.
4. Jumping to iteration 6, we see the red centroid has moved further to the right.

5. Iteration 9 shows the green section is much smaller than in iteration 2, blue has taken over the top, and the red centroid is thinner than in iteration 6.
6. The 9th iteration's results were the same as the 8th iteration, so it has "converged".

7.0 Classification vs. Clustering

Aspects	Classification	Clustering
Type of Machine Learning	Supervised	Unsupervised
How does it work	It uses algorithms to categorize the new data as per the observations of the training set.	It uses statistical concepts in which the data set is divided into subsets with the same features.
Number of class	Known	Unknown
Training Data	Required	Not required
Complexity	More complex than clustering	Less complex
Example of Algorithm	- Naive Bayes Classifier - Decision Tree - Random Forests	- K-Means - Mean-Shift Clustering - Gaussian (EM) Clustering

8.0 Reference

Chen, J., Zhang, H., Pi, D., Kantardzic, M., Yin, Q., & Liu, X. (2021). A Weight Possibilistic Fuzzy C-Means Clustering Algorithm. Scientific Programming, 2021.

Dalby, A. R. (2015, December 11). *A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4686108/#:%7E:text=Similarity%20or%20distance%20measures%20are,are%20placed%20into%20different%20clusters

DBSCAN (density-based spatial clustering of applications with noise). (n.d.). XLSTAT, Your data analysis solution. Retrieved May 17, 2022, from https://www.xlstat.com/en/solutions/features/dbscan-density-based-spatial-clustering-of-applications-with-noise

Edpresso Team. *(2019, October 11). Classification vs. clustering. Educative: Interactive Courses for Software Developers.* Retrieved May 17, 2022, from *https://www.educative.io/edpresso/classification-vs-clustering*

Harmouch, M. (2021, May 2). *17 clustering algorithms used in data science & mining.* Retrieved May 17, 2022, from https://towardsdatascience.com/17-clustering-algorithms-used-in-data-science-mining-49dbfa5bf69a#7e1d

ML | Classification vs clustering. (2019, October 3). GeeksforGeeks. Retrieved May 17, 2022, from https://www.geeksforgeeks.org/ml-classification-vs-clustering/

Partitioning Method (K-Mean) in Data Mining. (2020, February 5). GeeksforGeeks. Retrieved May 17, 2022, from https://www.geeksforgeeks.org/partitioning-method-k-mean-in-data-mining/

Pedamkar, P. (n.d.). *Hierarchical Clustering Algorithm | Types & Steps of Hierarchical Clustering*. eduCBA. Retrieved May 17, 2022, from https://www.educba.com/hierarchical-clustering-algorithm/

Prado, K. S. do. (2019, June 3). *How DBSCAN works and why should we use it?* Medium. Retrieved May 17, 2022, from https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

Robinson, D. (2022). *K-Means Clustering – What it is and How it Works – Learn by Marketing*. Learn by Marketing. Retrieved May 17, 2022, from https://www.learnbymarketing.com/methods/k-means-clustering