“ A Concise Report on Clustering , Subspace Clustering , Feature Selection / Reduction and Multiobjective optimization based Clustering ’’:- The prerequisites and an exhaustive approach.

## Clustering and Subspace Clustering

After enormous maneuvring and going through countless books and papers , I have indeed been enlightened with the overview of this genre of subject and study and have developed quite a lot of interest in this regard.

I would describe , in layman’s terms , Clustering to be a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes and/or dimensions. In fact Clustering is data mining technique of grouping objects or data into clusters in which objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Cluster analysis is an important data mining technique which is used to find data segmentation and pattern information. By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters.

At the very outset , the aim of cluster analysis is that the objects in a group should be similar to one another and different from the objects in other group which directly implies the fact that Clustering is much better when there is greater similarity within a group and greater the difference amongst the groups. Similarities and Dissimilarities are measured on the attribute values which describes the objects. Clustering methods are used to formulate and typecast the data, for data compression and model construction, for detection of outliers etc. Common approach of all clustering methods is to find clusters centre which represent each cluster. Based on the similarity metric and input vector cluster centre helps in determining which cluster is nearest or most similar one.

## *TYPES OF CLUSTERS:*

## 1 .Well-Separated Cluster:

A cluster is a set of points such that any point that is in a cluster is closer (or more similar) to every other point in the cluster than to any point which is not in the cluster.

## 2. Centre-based Cluster:

A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “centre” of a cluster, than to the centre of any other cluster. The centre of a cluster is often a centroid, the average of all the points in the cluster , or a medoid , the “most representative” point of a cluster.

## 3. Contiguous Cluster (Nearest Neighbour or Transitive Clustering):

A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

**4. Density-based Cluster**:

A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density .This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present.

Raw data --->> Clustering algorithms --->> Clusters

## *Ways of Clustering*

**a) Partitioning Clustering:**

Data objects are divided into non overlapping clusters so that each and every object is in exactly in one subset. The reason of dividing the data into several subsets is that checking of the all possible subset systems is computationally not feasible

**b) Hierarchical Clustering:**

Hierarchical clustering is a clustering technique in which the similar dataset is divided by constructing a hierarchy of clusters. This method is based on the connectivity approach. This hierarchy is created using two algorithms which are : Agglomerative and Divisive.

**Agglomerative** - The method starts with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest each other thus merged together to form the next largest cluster. The merging thus continues until a hierarchy of clusters is constructed with just a single cluster comprising all the records at the top of the hierarchy.

**Divisive** – The technique take the opposite approach from agglomerative techniques. They start with all the records in one cluster and then split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

**c) Density Based Clustering:**

Density-based clustering algorithms try to find clusters based on density of data points in a region. Objects in these sparse areas - that are required to separate clusters – are pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity.

**d) Grid Based Clustering:**

Grid-based clustering is used where the data space is divided into finite number of cells which forms the grid structure and performs clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids.

## SUBSPACE CLUSTERING

Based on my understanding , Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. Often in high dimensional data, many dimensions are irrelevant and can mask existing clusters in noisy data. Feature selection removes irrelevant and redundant dimensions by analysing the entire dataset. Subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces . Just as with feature selection, subspace clustering requires a search method and an evaluation criteria.

Subspace clustering must evaluate features on only a subset of the data, representing a cluster. They must use some measure to define this context. We refer to this as a “measure of locality”.

## Feature Selection

What I understood was that basically , Feature selection is the process of selecting relevant features or a candidate subset of features. The evaluation criteria are used for getting an optimal feature subset. In high-dimensional data where the number of samples is much less than the number of features , finding the optimal feature subset is a difficult task . For the data with N number of features, there exists 2 raised to the power N candidate subset of features . Hence Feature Selection becomes of crucial and vital importance.

Feature selection is a pre-processing step, used to improve the mining performance by reducing data dimensionality. Even though there exists a number of feature selection algorithms, still it is an active research area in data mining, machine learning and pattern recognition communities. Many feature selection algorithms confront severe challenges in terms of effectiveness and efficiency, because of recent increase in data dimensionality (data with thousands of features or attributes or variables). The aim of feature selection is to determine a feature subset as small as possible. It is the essential preprocessing step prior to applying data mining tasks. It selects the subset of original features, without any loss of useful information. It removes irrelevant and redundant features for reducing data dimensionality. As a result it improves the mining accuracy, reduces the computation time and enhances result comprehensibility.

I found , The four key steps of a Feature selection process are feature subset generation, subset evaluation, stopping criterion and result validation. The feature subset generation is a heuristic search process which results in the selection of a candidate subset for evaluation. It uses searching strategies like complete, sequential and random search to generate subsets of features. On the basis of selection strategy, feature selection algorithms are broadly classified into three categories namely Filter, Wrapper and Hybrid Method. It can be applied to data with high dimensionality. Wrapper Method requires a predetermined algorithm to determine the best feature subset. Predictive accuracy of the algorithm is used for evaluation. This method guarantees better results, but it is computationally expensive for large dataset. Hybrid Method combines Filter and Wrapper to achieve the advantages of both the methods. It uses an independent measure and a mining algorithm to measure the goodness of newly generated subset.

## Feature Reduction

After an exhaustive study of Feature Selection , studying this was rather easier . Feature reduction revolves around reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause a [classification](https://en.wikipedia.org/wiki/Statistical_classification) algorithm to [over fit](https://en.wikipedia.org/wiki/Overfitting) to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. When Data accumulates in an unprecedented speed , Data preprocessing becomes an important part for effective machine learning and data mining and feature reduction plays a pivotal role and has an effective approach to downsizing the data. Hence , When the input data to an [algorithm](https://en.wikipedia.org/wiki/Algorithm) is too large to be processed and it is suspected to be redundant , then it can be transformed into a reduced set of [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) thereby involving feature reduction.

## Multiobjective optimization based Clustering

This topic was , in fact , the most intriguing yet the most nuanced to learn . Nevertheless , I tried my best to extract as much knowledge as I could .

The goal of multi-objective clustering (MOC) is to decompose a dataset into similar groups maximizing multiple objectives in parallel. In this paper, we provide a methodology, architecture and algorithms that, based on a large set of objectives, derive interesting clusters regarding two or more of those objectives. The proposed architecture relies on clustering algorithms that support plug-in fitness functions and on multi-run clustering in which clustering algorithms are run multiple times maximizing different subsets of objectives that are captured in compound fitness functions.

A main goal of multi-objective clustering is to find individual clusters that are good with respect to multiple objectives; due to the nature of multiobjective clustering only clusters that are good with respect to at least two objectives are reported. In the remainder of this section we focus on a specific architecture and algorithms for .

After my detailed stint with such intriguing and insightful topics , I have come to not only realise their importance but also appreciate it. Hence I have concluded that each of the topics listed and explained here play crucial and pivotal roles in the this branch of study and provide not only the framework but also the sheer base , without which imagining this form of data mining or machine learning scope of study.

Lastly , Sir , I would absolutely like to express my heart-felt and deepest sense of gratitude to you for giving me this wonderful opportunity that has brought my knowledge a little more closer to these brilliant topics.

Regards,

Ayush Anand

Indian Institute of Engineering Science and Technology , Shibpur