Clustering Techniques for Big Data Analysis

Stelios Zimeras
http://orcid.org/0000-0002-9940-0041
University of the Aegean, Greece

ABSTRACT

Clustering is the process by which data is classified into semantically consistent clusters based on some measure of similarity. Typically, clustering is an unsupervised machine learning problem, meaning that the structure of the data must be detected without any label being available as to which category it belongs to. Various clustering techniques have been developed, which aim to find coherent groups among a large number of data registered in large databases. We could say that the clustering technique is directly related to the optimization technique and thus its applications multiply in finding homogeneous groups of elements. This work deals with clustering algorithms and their application to big data. First, the clustering concept, objectives, and techniques are studied. Then, the main clustering algorithms are analyzed, their positive and negative characteristics, the steps to be followed for their application, their mathematical formulas, and a small application for each one on a small data set.

INTRODUCTION TO CLUSTERING ANALYSIS

Clustering is the process by which data is classified into semantically consistent clusters based on some measure of similarity (Jain & Dubes, 1988). Typically, clustering is an unsupervised machine learning problem, meaning that the structure of the data must be detected without any label being available as to which category it belongs to. The goal in clustering is to create groups, each of which will gather homogeneous elements and each of these groups maintains a center, the most central (Singh & Singh; 2024).

Various clustering techniques have been developed, aimed at finding coherent groups among a large amount of data. The clustering technique is directly related to the optimization technique. The mathematical formulation is given by the following analysis. Let be the data set $X= \{x_i \in R^k, k=1,2,\ldots,i=1,2,\ldots,n\}$ and let be a set of clusters $C= \{c_j, j=1,2,\ldots,m\}$ and m>1. The clustering problem is to find a matching $f: X \rightarrow C$ so that each observation x_i from the data set is associated with a cluster $C_j, 1 \leq j \leq m$, and for each observation the similarity between it and any other observation from the same cluster is greater than the similarity between it and any observation from other clusters (Singh & Singh, 2024; Zhou et. al., 2024;

DOI: 10.4018/407568

Moujahid & Dornaika, 2025). To solve a data clustering problem, the clustering stages are described based on the following steps:

- Representation of the characteristics
- Defining a metric
- Data clustering technique
- Evaluation of the final result.

Distance metrics

A metric or distance on a set X is a non-negative function d: $XxX \rightarrow R$ for which the following three axioms hold:

$$d(x,y) = 0 \leftrightarrow x = y$$

$$d(x,y) = d(y,x)$$

$$d(x,z) \leq d(x,y) + d(y,z)$$

To cluster observations, we need to define a specified distance between points and between sets of points. Let consider dimension space n (in R^n) and two points $x = (x_1,x_2,\ldots,x_n)$ and $y = (y_1,y_2,\ldots,y_n)$; the most common distances between the two points are (Figure 1):

Euclidean: The Euclidean distance between two points is the length of a line segment between two points

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{1/2}$$

Manhattan: The Manhattan distance is a measurement of distance between two points in an N-dimensional vector space

$$d(x,y) = \sum_{i=1}^{n}\left|(x_i - y_i)\right|$$