K.J. Somaiya College of Engineering, Mumbai-77.

Report on Technical Paper

An Analysis of Efficient Clustering Methods for Estimates Similarity Measures

Research Paper : https://ieeexplore.ieee.org/abstract/document/8014710

Guided by:
Prof. Ujwala Bhangale

Submitted by:
Rahul Dandona 16010421015
Arya Nair 16010421063
Riya Thapar 16010421118

This paper is based on Clustering Analysis to form a group of data based on similarity and dissimilarity into different clusters using distance as the main parameter.

In data analysis, clustering is a method of unsupervised learning used to group similar patterns together.The data points from two distinct clusters will have the greatest distance, whereas the points within a cluster will have the shortest distance.The two classes or categories of objects are distinguished by this.Market research, pattern recognition, data analysis, and image processing all benefit from cluster analysis.

There are three categories in clustering algorithm:

1. **Partial Clustering Algorithm:** Dividing the entire dataset into smaller data sets representing cluster
2. **Density Based Clustering Algorithm:** Creating arbitrary shaped clusters. The main aim is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

3. **Hierarchical Clustering Algorithm:** It's a clustering algorithm which groups similar objects together. Each cluster is distinct from each other and objects within each cluster are similar to each other.

The objective of this technical paper is the similarity analysis. Clustering is used for similarity analysis since it is used to find the distance between data objects.

The most popular and important distance Functions listed are:

1. **Euclidean Distance:**
   a. It is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

2. **Manhattan Distance:**
   a. It is the distance in blocks between any two points in a city.
   b. It is defined as
      i. $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$
3. **Minkowski Distance:**
   a. Generalization of Euclidean and Manhattan
   b. It is defined as(L_p norm):

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h},$$

4. **Chebyshev Distance:**
   a. It is another variant of Minkowski Distance.
   b. If in the Minkowski Distance formula, h is replaced by ∞, then the distance obtained is Supremum Distance also known as Chebyshev Distance (Denoted by L_{max} or L_∞).

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

5. **Cosine Similarity:**
   a. It is the measure of similarity that can be used to compare documents or rank the documents with respect to a given vector of query words.

b. It is denoted as

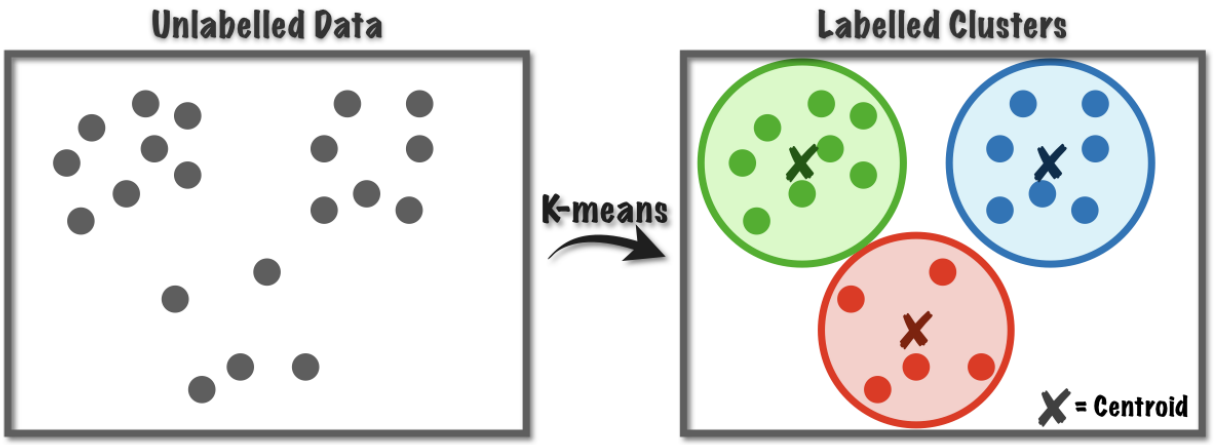
$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||},$$

6. Jaccard Distance:

a. Jaccard Distance is used for comparing the similarity between the datasets.

$$J\ A, B\ = \frac{|A \cap B|}{|A \cup B|}$$

In Similarity Analysis, distance plays an important role. The most widely used algorithm for clustering is the K-Means Algorithm . It is also based on the Euclidean Distance metric. K-Means Algorithm groups the unlabelled dataset into 'k' different clusters.



Dataset can have either Internal Similarity or External Similarity. Existing System makes use of the famous Euclidean distance method to find similarity because it is more effective than other methods.

If there are two objects (di,dj) then the distance is measured using the below formula.

$$\nabla d = \sqrt{\sum_{i=1}^{n} (d_i - d_j)^2}\ \forall_n = 1,2,\ldots.n$$

Along with the object (di & dj) and the distance, a third point, dh, comes into play in Multi viewpoint based Similarity.In order to make an accurate assessment of similarity, it is used to measure the similarity between various points of view.Using the formula below, this novel method determines distance.

$$sin_{d_i d_j \in s_r}(d_i, d_j) = \frac{1}{n-n_r} \sum_{d_h \in S \setminus S_r} \sin(d_i - d_h, d_j - d_h)$$

The technical paper proposes a system that uses an effective data association technique known as the content similarity coefficient to combine all of the benefits of the systems that were looked at with pattern-based similarity clustering. The system's primary objective is to use optimized functions to perform document clustering. The similarity between each document vector and its cluster's centroid is considered in this method. It is mainly used in algorithms for hierarchical clustering to make clustering more effective at finding similarities in text documents.

This report helps us to understand various clustering techniques and its key fundamentals. It makes us realize that distance plays a key role in similarity measurement. Very little research exists on Dissimilarity measures which is negative data. Most of the distance metrics and pattern similarity concepts for clustering aren’t robust enough to handle clustering negative data.

Clustering is a descriptive technique. It helps us address a wide range of challenges faced in the industry such as anomaly detection and recommender systems. By finding similar characteristics and transforming data to a usable and scalable size, clustering algorithms can identify traits and patterns to get insights faster.

References:

https://ieeexplore.ieee.org/abstract/document/8014710

https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e