**Empirical Analysis of Data Clustering Algorithms**

Pranav Nerurkar1, Sunil Bhirud2

Dept. ofComputer Engineering & IT, VJTI, Mumbai, India.

pranavn91@gmail.com, sgbhirud@vjti.org.in

**Abstract.** Clustering is performed to get insights into the data whose volume makes it problematic for analysis by humans. Due to this, clustering algorithms have emerged as meta learning tools for performing exploratory data analysis. A Cluster is deﬁned as a set of objects which have a higher degree of similarity to each other compared to objects not in the same set. However there is ambiguity regarding a suitable similarity metric for clustering. Multiple measures have been proposed related to quantifying similarity such as euclidean distance, density in data space etc. making clustering a multi objective optimization problem. In this paper, different clustering approaches are studied from the theoretical perspective to understand their relevance in context of massive datasets and empirically these have been tested on artiﬁcial benchmarks to highlight their strengths and weaknesses.

**Keywords:** Clustering algorithms, Community structure, Unsupervised learning.

1. **INTRODUCTION**

As the Digital transformation of the society gathers pace, there is an increase in proliferation of technologies that simplify the process of recording data efﬁciently. Low cost sensors, RFIDs, Internet enabled Point of Sales terminals are an example of such data capturing devices that have invaded our lives. The easy availability of such devices and the resultant simpliﬁcation of operations due to them has generated repositories of data that previously didn’t exist. Today, there exist many areas where voluminous amount of data gets generated every second and is processed and stored such ﬁelds are social networks, sensor networks, cloud storages etc. This has boosted the ﬁelds of machine learning, pattern recognition, statistical data analysis and in general data science.

Even though such a volume provides huge opportunities to academia and industry it also represents problems for efﬁcient analysis and retrieval [1]. To mitigate the exponential time and space needed for such operations data is compacted into meaningful summaries i.e. Exploratory Data Analysis [E.D.A] which shall eliminate the need for storing data in unsupervised learning literature such summaries are equivalent to ”clusters”. E.D.A. helps in visualization and promotes better understanding of the data. It utilizes methods that are at the intersection of machine learning, pattern recognition and information retrieval. Cluster analysis is the main task performed in it.

A Cluster in a data is deﬁned objectively using dissimilarity measures such as edit distance, density in a Euclidean or non-Euclidean data space, distance calculated using Minkowski measures, proximity measures or probability distributions. All measures concur that a threshold value should be set for grouping of objects in a cluster and objects which exceed such a threshold are dissimilar and should be separated from the cluster. Clustering gives a better representation of the data since all objects within a cluster have less variability in their attributes and they can be summarized efﬁciently. Clustering has found applications in other ﬁelds like estimating the missing values in data or identifying outliers in data.