Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 770–779

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Empirical Analysis of Data Clustering Algorithms

Pranav Nerurkar[a], Archana Shirke[b], Madhav Chandane[c], Sunil Bhirud[d]

[a]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*
[b]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*
[c]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*
[d]*Dept. of Computer Engineering & IT, VJTI, Mumbai - 400019, India*

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 data-sets and empirically these have been tested on artificial benchmarks to highlight their strengths and weaknesses.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

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, RF-IDs , 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

* Corresponding author. Tel.: +91-961-999-7797.
E-mail address: pranavn91@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.099