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.

© 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

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

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.

Clustering is thus a meta learning approach for getting insights into data and in diverse domains such as Market Research, E-Commerce, Social Network Analysis and Aggregation of Search Results among-st others. Multiple algorithms exist for organizing data into clusters however there is no universal solution to all problems. No consensus exists on the ”best” algorithm as each is designed with certain assumptions and has its own biases. These algorithms can be grouped into methodologies such as Partitioning based, hierarchical , density based, grid based, message passing based, neural network based, probabilistic and generative model based. However in terms of complexity it is a NP-hard grouping problem and so existing algorithms rely on approximation techniques or heuristics to reduce the search space in order to find the optimal solution. There is no universally agreed objective criteria for correctness or clustering validity and each of these algorithms has its own drawbacks and successes in solving the challenging problem of unsupervised clustering [3] [4] .

Motivated by these reasons, in this paper a review of the state of the art clustering algorithms is made to highlight their main strengths and weaknesses. Section II covers the theoretical aspects of these algorithms, Section III contains the Experiments performed using these algorithms and Section IV has the conclusion on the results.

2. Types of Clustering Algorithms

Various clustering algorithms are found in literature [1][3][4] and are broadly categorized into categories on the basis of an algorithm designer’s perspective with emphasis on the underlying clustering criteria:

2.1. Partition based clustering algorithms [5]

The general principle in these algorithms is that a cluster should contain atleast one object and that each object must belong to exactly one group i.e. hard clustering. The Number of clusters k is pre-specified by the user making this a semi supervised algorithm although many strategies have been suggested to estimate the ideal number of clusters like the empirical method where $k = \sqrt{n}$ where $n = \mid N \mid$ points and Elbow method where the k is fixed as the turning point on the graph of k v/s Avg. distance to centroid. The objective function to be minimized in these k-partitioning algorithms is SSE i.e. Sum of Squared Distance.

$$SSE = \sum_{k=1}^{k} \sum_{x_i \epsilon c_k} \|x_i - c_k\|^2 \tag{1}$$

where c_k = centroid of the cluster,

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات