Computer Science 3.3.3(8) 2019-20

Annexure 1

Sustainable Computing: Informatics and Systems 23 (2019) 144–157

Contents lists available at ScienceDirect

Sustainable Computing: Informatics and Systems

journal homepage: www.elsevier.com/locate/suscom

Sustainable automatic data clustering using hybrid PSO algorithm with mutation

Manju Sharma[a], Jitender Kumar Chhabra[b]

[a] Computer Science Department, Govt College for Women, Karnal, India
[b] Computer Engineering Department, National Institute of Technology, Kurukshetra, India

ARTICLE INFO

Article history:
Received 30 January 2019
Received in revised form 15 July 2019
Accepted 29 July 2019
Available online 5 August 2019

Keywords:
Sustainable computing
Automatic clustering
Big data
Genetic algorithm
Particle swarm optimization
Hybrid optimization
Mutation

ABSTRACT

Widespread use of various mobiles, social networks and IOT devices results into continuous generation of the data, often leading to the formation of the big data. Sustainable grouping of such data into various clusters is an open research problem, which aims to provide solutions which are computationally efficient and maintainable over dynamic data. This paper proposes a new sustainable clustering algorithm HPSOM by hybridization of PSO with mutation operator for clustering of the data generated from different networks. The data generated by such networks is usually dynamic and heterogeneous in nature and the number of clusters is not fixed/ known in advance. Hence the proposed algorithm is further extended as AHPSOM for generating and re-adjusting the clusters automatically over the mobile network devices, and it facilitates the generation of sustainable clusters. Firstly, the performance of basic HPSOM is evaluated on six real life data sets and is also compared with some known evolutionary clustering techniques in terms of SWCD and Convergence speed. Then the performance of AHPSOM is evaluated using some synthetic and real life datasets using some validity metrics (cluster numbers, intra cluster distance, inter cluster distance, ARI and F-measure) and is also compared with some prevalent state-of-art automatic clustering techniques. The results show that the proposed algorithm is very efficient in terms of creating well separated, compact and sustainable clusters.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

Widespread use of sensors, mobiles, and networking devices has been instrumental in generating a huge amount of data, which needs to be managed properly for efficient and sustainable processing. The data generation from various social networks, sensors, mobiles and IOT devices etc is a continuous phenomenon and over a period of time, such data takes shape of the big data. Organizing such data into similar and dissimilar groups (clusters) is an open research problem, and researchers are aiming to develop efficient, scalable and sustainable techniques [1]. Data Clustering is an unsupervised learning method that organizes the data sets items into groups so that there must be more cohesiveness within the cluster and less coupling among different clusters [2]. If the data size is too big, arranging the data through traditional deterministic techniques into clusters takes huge processing and exponential time, and is practically not much useful, as such computing may normally be carried out in mobile computing environment, where efficiency as well as sustainability of this computing is must. Nowadays clustering is being widely used in numerous fields like wireless sensor networks [3], social networks, mobile networks, image processing, pattern recognition, biology, software re-modularization etc [4–6,1–8]. Many of these applications are run using the devices which have limited energy, less computing power and need to complete the processing in minimum possible time [9]. Moreover, data generation over sensors, social networks and mobile networks is not a onetime process, but it gets generated regularly as well as dynamically [10]. Hence researchers need to find solutions for clustering, so that big data clustering gets completed in reasonable time and is sustainable as well over the dynamic data. Sustainability of the clustering will need re-clustering at regular intervals (based on dynamicity of the data) and such re-clustering must be fast and efficient, even if clusters are near optimal (instead of the best optimal), as the dynamicity of the data will keep on updating the optimal clusters. Keeping in view of the heterogeneity as well as dynamicity of the data, the number of clusters cannot be pre-determined. Further number of clusters may change after some time due to change in data. Hence the clustering technique will be sustainable only if it is carried out without any prior knowledge of the number of clusters. The algorithm must reach towards optimal clusters by automatically determining number of clusters based on characteristics of the data. Presence of multiple features in the

E-mail addresses: manjusharma.kul@gmail.com (M. Sharma), jtenderchhabra@gmail.com (J.K. Chhabra).

https://doi.org/10.1016/j.suscom.2019.07.009
2210-5379/© 2019 Elsevier Inc. All rights reserved.