Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems ••• (••••) •••–•••
FUZZY sets and systems
www.elsevier.com/locate/fss

Speeding up the large-scale consensus fuzzy clustering for handling Big Data

Minyar Sassi Hidri [a,b,*], Mohamed Ali Zoghlami [b], Rahma Ben Ayed [b]

[a] *Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia*
[b] *University of Tunis El Manar, National Engineering School of Tunis, Tunisia*

Received 2 June 2016; received in revised form 13 September 2017; accepted 3 November 2017

Abstract

Massive data can create a real competitive advantage for the companies; it is used to better respond to customers, to follow the behavior of consumers, to anticipate the evolutions, etc. However, it has its own deficiencies. This data volume not only requires big storage spaces but also makes analysis, processing and retrieval operations very difficult and hugely time-consuming. One way to overcome these problems is to cluster this data into a compact format that is still an informative version of the entire data. A lot of clustering algorithms have been proposed. However, their scaling is poor in terms of computation time whenever the size of the data gets larger. In this paper, we make full use of consensus clustering to handle Big Data clustering. We use sampling combined with a split-and-merge strategy to fragment data into small subsets, then basic partitions are locally generated from them using RHadoop's parallel processing MapReduce model and later a consensus tendency is followed to obtain the final result. A scalability analysis is conducted to demonstrate the performance of the proposed clustering models by increasing both the number of computing nodes used and the sample size while satisfying the volume and the velocity dimensions.

Keywords: Big Data analytics; Consensus tendency; Fuzzy clustering; Partial data clustering; Sampling; MapReduce; RHadoop

1. Introduction

Big Data has boomed during the last decade which brought companies to radically change the way they collect, store and analyze their data to discover useful knowledge in order to increase their performance. The conundrum of these organizations is what to do with this huge data and how to glean key insights from them. Moreover, data can create a real competitive advantage to the companies; it is used to better respond to customers, to follow the behavior of consumers, to anticipate the evolution, etc. But it has its own deficiencies as well. This mass of data requires big storage space and makes analysis, processing and retrieval operations very difficult and hugely time-consuming.

* Corresponding author at: University of Tunis El Manar, National Engineering School of Tunis, 1002, Tunisia.
E-mail addresses: mmsassi@iau.edu.sa, minyar.sassi@enit.utm.tn (M. Sassi Hidri), medali.zoghlami@enit.rnu.tn (M.A. Zoghlami), rahma.benayed@enit.utm.tn (R. Ben Ayed).

https://doi.org/10.1016/j.fss.2017.11.003
0165-0114/

It is clear that the existing analyzes have not kept pace with the growth of this change and must evolve towards greater intelligence. As one of unsupervised classification methods, clustering has been used to overcome analysis problems and it is seen as the key of Big Data analytics because it transforms the data in a compact format that is still an information version of the whole data.

The data clustering process consists in partitioning a set of data objects into several subsets called clusters. The partition is made so that the objects within the same cluster are similar while the objects in different clusters are dissimilar.

A lot of clustering algorithms have been proposed. They can be performed in two different modes: hard and fuzzy [22,55,69]. Hard clustering methods assume that the different clusters are disjoint and non-overlapping. In this case, any data object should belong to one and only one cluster, however in practice clusters may overlap, and each data object may belong to several clusters with different degrees of membership. This scenario can be modeled using fuzzy set theory [89], in which cluster's elements are associated with a numeric membership degree in [0, 1] [12,41,46,63,71]. This requirement has led to the development of fuzzy clustering methods.

One of the widely used fuzzy clustering methods is the fuzzy c-means (FCM) algorithm that was proposed by Dunn [21] and improved by Bezdek [8]. FCM is a fuzzy partitional clustering approach, and can be seen as an improvement and a generalization of k-means [60] clustering algorithm.

Several clustering algorithms based on weighted approaches have been developed. In particular, the Possibilistic C-Means (PCM) [54], the Evidential C-Means (ECM) [61] and the Fuzzy-Possibilistic C-Means (FPCM) [65]. The PCM clustering algorithm is a method inspired by the FCM algorithm which has been shown to be advantageous to reduce the effect of noise and outliers in the data. The ECM clustering algorithm is an extension of the FCM one in order to work in the belief functions framework with credal partitions of data. The FPCM approach has been proposed to avoid the undesirable tendency to have identical clusters that can be produced by PCM algorithms in the case of poor initializations.

Distributed systems continuously produce a large volume of data, which imposes a prohibitive communication burden if all the data are transferred to a central node for processing. In the case of massive data, the scaling of the classical clustering algorithms (hard or fuzzy) is poor in terms of computation time as the size of the data gets larger. Traditional single-machine clustering algorithms cannot handle this huge amount of data because of their high complexity and computational cost. To empower clustering algorithms to work with huge datasets, multiple machine clustering techniques [72] have attracted more attention because they are more flexible in terms of scalability and speed. The need to access more resources requires the design of distributed algorithms that can be run on multiple machines or nodes.

Distributed clustering is one of the techniques that reduce the communication load of a mass of data across multiple sites. The goal of such techniques is to find a structure that describes the distributed data without the need to centralize either data or processing. The idea is to compute a local summary at each site in the form of synthetic representations, which are much less voluminous than original data and transmit them to a server for a centralized calculation.

The general approach of distributed data clustering techniques is based on three different steps [47]:

1. Clustering data at local nodes to estimate local cluster models and then transmitting them to a central node.
2. Building a global model which is an aggregate function of local models.
3. Updating all local models.

Most distributed clustering methods apply these steps in order to find a global partition, and then improve the quality of local ones. However, some distributed clustering methods are intended to optimize only the global model [34,40]. Other are intended to optimize exclusively the local models [39,76]. In the first case, the global model is obtained from the aggregation of models estimated on all nodes, then each node tries to improve the quality of the local partition based on the global one. In the second, nodes collaborate together by exchanging data between themselves in order to optimize their local models. These methods do not enforce a global cluster partition for all nodes and each one obtains its own model at the end of the clustering process, which reflects the characteristics of its local data.

Big Data analytics added more challenges to this subject which urges more research to be conducted for clustering algorithms improvement. Motivated by this field, the major contributions of this paper are as follows:

Please cite this article in press as: M. Sassi Hidri et al., Speeding up the large-scale consensus fuzzy clustering for handling Big Data, Fuzzy Sets Syst. (2017), https://doi.org/10.1016/j.fss.2017.11.003