Expert Systems With Applications 91 (2018) 198–210

Contents lists available at ScienceDirect

Expert Systems With Applications

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

FWCMR: A scalable and robust fuzzy weighted clustering based on MapReduce with application to microarray gene expression

Behrooz Hosseini[a], Kourosh Kiani[a,*]

[a] Electrical and Computer Engineering Department, Semnan University, Semnan, Iran

ARTICLE INFO

Article history:
Received 28 May 2017
Revised 7 August 2017
Accepted 31 August 2017
Available online 8 September 2017

Keywords:
MapReduce
Distributed density based clustering
Gene expression microarray
Fuzzy weighted clustering
Decision making
Big data

ABSTRACT

Data clustering is a very useful data mining technique to find groups of similar objects present in the dataset. Scalability to handle immense volumes, robustness to intrinsic outlier data and validity of clustering results are the main challenges of any data clustering approach. In order to address these challenges, a fuzzy weighted clustering approach which is comprehensibly parallel and distributed in every phase, is proposed in this research. Although the proposed method can be used for various data clustering purposes, it has been applied in gene expression clustering to reveal functional relationships of genes in a biological process. Conforming to MapReduce, the proposed method also presents a novel similarity measure which benefits from combining ordered weighted averaging and Spearman correlation coefficient. In the proposed method, density reachable genes were joined to establish subclusters. Afterwards, final cluster results were obtained by merging these subclusters. A voting system detects the best weights and consequently the most valid clusters among all possible results for each distinct dataset. The whole algorithm is implemented on a distributed processing platform and it is scalable to process any size of data stored in cloud infrastructures.

Precision of resulting clusters were evaluated using some of the well-known cluster validity indexes in the literature. Also, the efficiency of the proposed method in scalability and robustness was compared with recently published similar researches. In all the mentioned comparisons, the proposed method outperformed recent works on the same datasets.

1. Introduction

Clustering is an unsupervised approach of data mining which attempts to group similar objects according to similar characteristics. Generally speaking, clustering analysis has three main area to address: similarity measurement(s), clustering algorithm, and clustering validation. There has been a rich literature on clustering analysis over the past decades (Aggarwal;charu & Chandan, 2014; Fahad et al., 2014; Kim, 2009). Genes are fundamental biological information storage units found in every living creature. The process by which information from a gene is used in the synthesis of functional gene products is called gene expression. Elucidating patterns laid in gene expression data proposes unique possibilities for a rich understanding of functional genomics (Allison, Cui, Page, & Sabripour, 2006; Sateesh Babu & Suresh, 2013). Quality of gene expression clustering plays a substantial role in the refinement of valuable biological data to find natural groups existing in the gene set. Microarray (Castellanos-Garzón, García, Novais, & Díaz, 2013) across collections of related samples, is one of biotechnological means used to facilitate simultaneous monitoring of the expression levels of thousands of genes during important biological processes. With recent advances in gene microarray technology, high-density DNA arrays can be monitored all in one place (Castellanos-Garzón et al., 2013). However, the growing need for storing, retrieving and processing large-scale genome data presents considerable challenges to biologists (Wong, 2016). Various clustering algorithms have been addressed to more precisely group similar genes according to similar expression patterns (Kerr, Ruskin, Crane, & Doolan, 2008; Yoo et al., 2012). However, there is no single “best” clustering algorithm which is the “winner” in every case (Xu & Wunsch, 2005). Microarray datasets are very complex and have an intrinsic outlier or missing values (Moorthy, Saberi Mohamad, & Deris, 2014). Robustness keeps clustering results valid in the presence of outlier data. With the presence of such undesirable data within any dataset including gene expression, robustness of clustering algorithm is very important. Recently microarray datasets have become great both in volume and complexity. Hence, scalable algorithms are inevitably needed to handle them. To address these challenges, this research proposes a distributed

* Corresponding author.
E-mail addresses: hosseini@semnan.ac.ir (B. Hosseini), kourosh.kiani@semnan.ac.ir, kkiani2004@yahoo.com (K. Kiani).

http://dx.doi.org/10.1016/j.eswa.2017.08.051
0957-4174/

density based clustering algorithm that tries to group the genes with a novel fuzzy weighted similarity metric. In the rest of the paper, the proposed method is named FWCMR which is an acronym for Fuzzy Weighted Clustering approach based on MapReduce.

In distributing the total data through nodes of Hadoop (the most popular big data processing framework) (White, 2012) in a way that does not affect their possible correlations, FWCMR finds the major density centers where the subclusters originate from. Consequently, the proposed clustering algorithm tries to merge the middle step subclusters into an integrated and unified final cluster result. Unlike partially distributed algorithms such as parallel k-means addressed by Zhao, Ma, & He (2009), FWCMR is a completely distributed algorithm that uses MapReduce (MR) (Shim, 2012) in every step. The proposed method is also robust to intrinsic outliers within datasets in comparison with recently published similar approaches. Finally, the quality of resulted clusters is measured by widespread clustering validity measurements in the literature. Experiments demonstrate the effectiveness and efficiency of FWCMR in comparison with similar researches. The proposed approach described in this research is motivated by the need of method for data clustering and specially gene expression clustering that is reliable, scalable and robust to outliers at the same time. The contributions of this paper are: (1) benefiting from a parallel and distributed algorithm which correctly discovers valid gene clusters; (2) a scalable, flexible and yet robust similarity measure for gene expression clustering on the basis of Ordered Weighted Averaging (OWA) operator; (3) compatibility with recent cloud computing frameworks and complete integrity with the MapReduce paradigm and specially, Hadoop; (4) addressing a new density based clustering which introduces gene density connectivity to prune the exploring space for similarity assessment.

FWCMR gives considerable results using scientific case studies and easily scales with any size of dataset. All the steps were implemented in MapReduce and no serial bottleneck was observed in the whole algorithm. The roadmap of this paper is organized as follows: In Section 2, preliminaries including the background on similarity measurement, density based clustering steps and the theoretical concepts of the approach are introduced, together with an overview of the previous works on microarray gene expression clustering. In Section 3, the proposed method is presented by introducing a novel fuzzy weighted similarity measurement and a distributed clustering procedure benefiting from MapReduce. Section 4 explains and demonstrates the practical details of the experiments, evaluations and results. Finally, the paper results are reviewed and discussed in Section 5 and concluded in Section 6.

2. Preliminaries and related works

2.1. Clustering definition

A good clustering method for gene expression should have characteristics such as, 1) simply coping with noisy high dimensional data, 2) being unaffected by the order of input, 3) having amenable computation complexity (that is, allow increased data load without breakdown or least requirement of major changes), 4) requiring few input parameters, 5) being independent of any metadata or a priori knowledge on cluster numbers or results structure (Kerr et al., 2008).

Clustering approaches on big data can be grouped into various methods (Jiang, Tang, & Zhang, 2004; Shim, 2012), but generally, they are divided into partition based, hierarchical and density based algorithms. A good deal of clustering algorithms applied so far produce hard clusters, that is, each gene is assigned only to one cluster. Hard clustering is suitable in the case where clusters are well separated. However, in the microarray data clustering, where gene clusters might frequently overlap, soft clustering is more useful (Futschik & Carlisle, 2005).

Cluster analysis is an exploratory procedure. Hence, having no a priori information on gene groups such as their actual number or possible outlier pattern is common. Arbitrary selection of this number may undesirably bias the search. Partition based clustering algorithms like k-means (Azimi, Ghayekhloo, Ghofrani, & Sajedi, 2017), fuzzy c-means (Nayak, Naik, Behera, & Abraham, 2017), and similar algorithms suffer from the dependency of having a priori knowledge on the number of clusters and possible pattern of data distribution. Partition based algorithms are also sensitive to outlier so that if one of the outliers is accidentally selected as the centroid or a simple member of the cluster, it can affect the accuracy of the final results. Results are also dependent on initial cluster centers (Aggarwal;charu & Chandan, 2014). In hierarchical clustering (Gao, Jiang, She, & Fu, 2010), if two nodes are incorrectly joined to a cluster, there is no correction step in the later stages to solve this problem and this wrong assignment will remain in the final results. More than that, hierarchical clustering in each step relies on the results from previous step to decide on the rest of the procedure. This voracious characteristic is not suitable for distributed and parallel computation. However, density based algorithms (Kriegel, Kröger, Sander, & Zimek, 2011) require no special prior knowledge on the data or the number of clusters. They are robust to outlier's side effects and flexible to change in any shape that fits the data distribution.

Density based clustering approaches has shown better performance in average in comparison with other clustering methodologies (Aggarwal;charu & Chandan, 2014). In order to benefit from the mentioned advantages, FWCMR is a kind of density based clustering algorithms. Extremely high dimensional gene expression data make the conventional clustering process time consuming with heavy computational cost. FWCMR benefits from the distributed processing schema inspired by MapReduce in every step which enables the parallel execution of gene expression clustering for any arbitrary size of data.

2.2. Literature review

Gene clustering has been proven to be helpful in understanding gene function, regulation and their involvement in the cellular processes (Daxin Jiang et al., 2004). Clustering is also beneficial in revealing subcell types and better understanding of the transcriptional regulatory network. Clustering gene expression data have been an area of interest in recent decade (Kerr et al., 2008). Therefore, many profitable reviews on gene expression clustering have also been published (Kotlyar, Fuhrman, Ableson, & Somogyi, 2002; Perdew, Vanden Heuvel, & Peters, 2014). One of the earliest approaches to gene expression clustering was addressed by Sharp, Tuohy, & Mosurski (1986) in which two predefined distinct clusters are established. One is for extremely expressed genes and the other one is for ordinary expressing ones. Later, after the introduction of microarray technologies, many clustering approaches were used to clarify this mass of data, which have been reviewed by Sherlock (2000). A good research on microarray clustering was conducted by Salem, Jack, and Nandi (2008) in which self-organizing oscillator networks are customized for microarray clustering. Their method does not require the knowledge of the number of clusters and indicates constructive achievements in comparison with hierarchical, k-means and Self Organizing Map (SOM) clustering using Mahalanobis distance. However, the algorithm suffers from the calculation of a large distance matrix that records the distance of every point. It also imposes heavy computation costs on intensive datasets and it is dependent on the serial oscillation phases, which makes it very difficult to be adopted in a parallel implementation. Authors in P Maji & Paul (2013) addressed a rough