Adapting a Multi-SOM Clustering Algorithm to Large Banking Data

Imèn Khanchouch[1(✉)] and Mohamed Limam[2]

[1] ISG, University of Tunis, Tunis, Tunisia
imen.khanchouch@yahoo.fr
[2] University of Dhofar, Salalah, Oman
limam@du.edu.om

Abstract. It the recent years, Big Data (BD) has attracted researchers in many domains as a new concept providing opportunities to improve research applications including business, science, engineering. Big Data Analytics is becoming a practice that many researchers adopt to construct valuable information from BD. This paper presents the BD technologies and how BD is useful in Cluster Analysis. Then, a clustering approach named multi-SOM is studied. In doing so, a banking dataset is analyzed integrating R statistical tool with BD technologies that include Hadoop Distributed File System, HBase and Map Reduce. Hence, we aim to decrease the time execution of multi-SOM clustering method in determining the number of clusters using R and Hadoop. Results show the performance of integrating R and Hadoop to handle big data using multi-SOM clustering algorithm and to overcome the weaknesses of R.

Keywords: Big data · Big data analytics · Clustering · multiSOM
RHadoop

1 Background

Big data (BD) was initially defined by Laney (2001) with a 3V model: volume (amount of data), velocity (speed of data) and variety (different sources of data types) and this model was used by IBM, Gartner and Microsoft.

BD can handle very large and complex datasets that can be structured or unstructured. Some of the popular organizations that hold Big Data are: Facebook which has 40 Petabytes (PB) of data and Yahoo! has 60 PB of data. Facebook captures 100 Terabytes (TB)/day and Twitter captures 8 TB/day. The data characteristics are different from a researcher to another such as Shah et al. (2015) say 3Vs (Volume, Velocity and Variety) of data, Liao et al. (2014) reported 4Vs (Volume, Velocity, Variety, and Variability) of data and Demchenko et al. (2013) say 5Vs (Volume, Velocity, Variety, Value and Veracity) and finally Gandomi and Haider (2015) say 6Vs (Volume, Velocity, Variety, Veracity, Variability, and Value) of data.

- *Velocity* refers to the low latency, real-time speed at which the analytics need to be applied.
- *Volume* refers to the size of the dataset. It may be in KB, MB, GB, TB, or PB based on the type of the application that generates or receives the data.

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 745, pp. 171–181, 2018.
https://doi.org/10.1007/978-3-319-77703-0_17

- *Variety* refers to the various types of the data that can exist, for example, text, audio, video, and photos.
- *Veracity* refers to increasingly complex data structure, anonymities, imprecision or inconsistency in large data-sets.
- *Variability* refers to the data which is constantly changing.
- *Value* refers to extracting knowledge/value from vast amounts of data without loss for users.

BD is applied in many domains such as Knowledge Management (KM), Cluster Analysis, management, marketing, etc. For example, Khan and Vorley (2017) aim to show how big data text analytics is efficient as an enabler of Knowledge Management. He applied big data text analytics tools such as MapReduce, Zookeeper, HBase, etc. on 196 articles published in two journals (the Journal of Knowledge Management and Knowledge Management Research & Practice) during two years and the results show the 50 most frequently used shared words across these articles. Chan (2014) integrates the concepts of Customer Relationship Management and Knowledge Management process with BD tools to obtain an architecture for BD customer knowledge management. Nowadays, BD is increasingly applied in cluster analysis, Sajana et al. (2016) give an overview on clustering methods for BD Mining. They present the different categories of clustering algorithms with the properties of BD characteristics such as dimensionality, shape of cluster, size, complexity and noise. However, here in this paper we will apply multi-SOM clustering method for BD because it has never been used for BD. Franke et al. (2016) introduce some strategies for BD analysis such as visualization, reducing dimensionality, optimization, regularization and sequential learning. Then, they provided examples of some applications used in BD such as public health, health policy, education, image Recognition and labelling, digital humanities and materials science. Ur Rehman et al. (2016) presented a review of reduction methods for BD and BD complexity. Sivarajah et al. (2017) presented the different analytical methods which are descriptive analytics, predictive analytics, prescriptive analytics and pre-emptive analytics methods. Then, the BD challenges are detailed. Chen et al. (2014) aim to give an overview about BD, its challenges, techniques and tools. BD techniques mentioned are data mining, statistics, optimization methods, machine learning, visualization approaches and social network analysis. However, big data tools are Hadoop, Map/Reduce, Dryad, Mahout, skytree server, pentaho, jaspersoft, Karmasphere, Tableau and Talend Open Studio. Then, they explained the different principles for designing BD systems. Yang et al. (2016) introduced the different future opportunities and innovations for BD and its challenges with cloud computing. Also, García et al. (2016) explained the challenges in data preprocessing for BD.

In this paper, we focus on a neural network approach of clustering. In Sect. 2, we will discuss some BD challenges. The integration of R and Hadoop is introduced in Sect. 3. Then, we give the experimentation results in Sect. 4. In the last section, a conclusion is drawn.

2 Big Data Challenges

2.1 Data Analysis (DA)

Data Analysis or Data Analytics (DA) is a process for obtaining raw data and transforming it into information useful for decision-making. DA is defined by Tukey (1962) as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data". There are several phases in DA process which are: Data requirements, data collection, data processing, data cleaning, exploratory data analysis, modeling and algorithms and data product. The most important methods for DA are the calculation of the mean, the standard deviation and the regression.

2.2 Knowledge Management (KM)

KM is defined by Duhon (1998) as "a discipline that promotes an integrated approach to identifying, capturing, evaluating, retrieving, and sharing all of an enterprise's information assets. These assets may include databases, documents, policies, procedures, and previously un-captured expertise and experience in individual workers". KM depends on the management of the organization's knowledge creation and conversion mechanisms, organizational memory and retrieval facilities, organizational learning and organizational culture. BD could enable better knowledge management.

2.3 Data Mining (DM)

DM is a process of data management that contains a set of algorithms to extract useful information from data, drawing from many areas such as machine learning, pattern recognition and data visualization in different domains. It is a particular DA technique and is closely linked to data visualization. The major tasks of DM are prediction, cluster analysis, classification and association rules. A large set of data mining approaches have been developed in statistics, such as Neural network (NN) which have shown its effectiveness in clustering analysis task. DM refers to the activity of going through big data sets to look for relevant or pertinent information. Hence, BD is the asset and DM is the executer of that is used to provide beneficial results.

2.4 Data Visualization (DV)

DV refers to the techniques used to communicate data or information by encoding it as visual objects such as points, lines or bars contained in graphics in order to communicate information clearly and efficiently via plots. It is closely related to DA and statistical graphics. For Big Data applications, it is sometimes difficult to conduct data visualization because of the large size and high dimension of BD. Therefore, Wu et al. (2012) say that uncertainty can lead to a great challenge to effective uncertainty-aware visualization and the new frameworks for modeling uncertainty are necessary through analytical processes.

3 Cluster Analysis

Cluster Analysis (CA) is one of the most important techniques in data mining. It is the subject of much recent research in different domains such as: bioinformatics, marketing, finance and text mining. The main idea of clustering is to partition a given data set into groups of similar objects where the similarity is commuted based on a distance function. CA improves the efficiency of data mining by combining data with similar characteristics. According to Sheikholeslami et al. (2000) clustering techniques could be classified into four types: partitioned clustering methods, hierarchical clustering, density-based clustering and grid based clustering. NN belongs also to clustering methods. NN are complex systems with high degree of interconnected neurons. Unlike the hierarchical and partitioning clustering methods NN contains many nodes or artificial neurons so it can accept a large number of high dimensional data. Many neuronal clustering methods exist such as (SOM) and multi-SOM.

3.1 Multi-SOM Definition

The Multi-SOM algorithm is an extension of Self Organizing Map (SOM) algorithm introduced by Kohonen (1981). Multi-SOM was firstly introduced by Lamirel (2001) for scientific and technical information analysis specifically for patenting transgenic plant to improve the resistance of plants to pathogen agents. He proposed an extension of SOM to multi-SOM to introduce the notion of viewpoints into the information analysis with its multiple maps visualization and dynamicity. A viewpoint is defined as a partition of the analyst reasoning. Each map in multi-SOM represents a viewpoint and the information in each map is represented by nodes (classes) and logical areas (group of classes). Lamirel (2002) applied multi-SOM on an iconographic database. The latter is the collected representation illustrating a subject which can be an image or a document text. Ghouila et al. (2009) applied multi-SOM algorithm for macrophage gene expression analysis in order to overcome the weaknesses of Self Organizing Map (SOM) method. The idea consists on obtaining compact and well separated clusters using an evaluation criterion namely Dynamic Validity Index (DVI) proposed by Shen et al. (2005). They showed that multi-SOM is an efficient approach for determining the optimal number of clusters. Khanchouch et al. (2014) applied multi-SOM for real data sets to improve the algorithm of Ghouila et al. (2009). The proposed algorithm as shown below aims to find optimal segments using an evaluation criterion which is DB index. It is widely used and based on compactness and separation of clusters. We have already developed the multi-SOM algorithm using R language and conducted an evaluation using different clustering evaluation indices such as silhouette index and Dunn index in Khanchouch et al. (2015). Then, we have used the multisom r package developed by Chair and Charrad (2015) which provides an implementation of the Multi-SOM method within the R programming environment. This package is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=multisom. The authors proposes to combine Multi-SOM to thirty validation indices. At each layer, all indices are computed. The package provides numerous clustering validity indices to estimate the number of clusters in the dataset and determine the best clustering scheme from different results.

3.2 Multi-SOM Algorithm

The main idea of the multi-SOM approach is that input data are firstly trained by SOM algorithm. Then, the other levels of data are clustered iteratively based on the first SOM grid. Thus, the size of the maps decreases gradually since only a single neuron is obtained in the last layer. Each grid groups similar elements into groups from the previous layer. The objects in a partition could be homogenous or heterogeneous and not necessary similar. However objects in one cluster are similar and homogenous where a criterion of similarity is inevitably used.

The different steps of multi-SOM algorithm are described as follows:

Algorithm. Multi-SOM (Khanchouch et al. 2014)

Begin

• *Step1*: *Clustering data by SOM*

s = 1;

Batch SOM (W1, H1, I1, max_it);

Compute DB index;

s = s+1;

• *Step2: Clustering of the SOM and cluster delimitation*

Hs = Hs − 1;

Ws = Ws − 1;

Repeat

Batch SOM (W1, H1, I1, max_it);

Compute DB index on each SOM grid;

s=s+1;

Until obtaining the minimum value of DB index;

Return (Data partitions, optimal cluster number);

End

4 Integrating R with Hadoop

In order to process large datasets, the processing power of R in data analytics can be combined with the power of Hadoop framework for Big data to get the work done. Therefore, the integration of such data-driven tools and technologies can build a powerful scalable system that has features of both of them. While R is very useful for statisticians and data analysts, it handles data analysis functions such as exploration, loading, visualization, classification, clustering… while Hadoop will realize the storage of parallel data as well as computation power against distributed data. So with the integration of R and Hadoop we can forward data analytics to BD analytics to solve the problem of handling large amount of data and at the same time to decrease the time execution of our multi-SOM algorithm.

4.1 Hadoop Framework

Apache Hadoop is an open source Java framework for BD processing. The strength of Hadoop is to store and process very large amounts of data in the Terabytes (TB) and even Petabytes (PB) range. Hadoop has two main features: Hadoop Distributed File System (HDFS) that provides the data storage and Map Reduce that provides the distributed processing which are defined in Table 1. Hadoop includes an ecosystem of other components built over the HDFS and Map Reduce layer to enable various types of operations on the platform as shown in Table 1. These features provide high degree of scalability and flexibility and also fault tolerance.

Table 1. Available tools for big data analytics

Big data analytics tools	Description
Apache Hadoop Distributed File System (HDFS)	Open source Java based software framework responsible for storing data on the cluster, written in Java
MapReduce	The system used to process data in the Hadoop cluster, consists of two phases: Map, and then Reduce
Zookeeper	Facilitate a centralized infrastructure, provide synchronization across a cluster of computers
HBase (The Hadoop Database)	A column family-store database layered on top of HDFS – Based on Google's Big Table – Provides interactive access to data Can store massive amounts of data – Multiple Terabytes, up to Petabytes of data
Flume	A distributed, reliable, available service for efficiently moving large amounts of data as it is produced – Ideally suited to gathering logs from multiple systems and inserting them into HDFS as they are generated
Hive	A data warehouse system for Hadoop that facilitates easy data summarization, ad hoc queries, and the analysis of large datasets stored in HDFS

4.2 Introducing R

R is an open source software used by data scientist statisticians and others who need to make statistical analysis of data such as regression, clustering, classification, and text analysis. It was developed by Ihaka and Gentleman (1996) at the University of Auckland in New Zealand. It has various functions for machine learning and statistical tasks such as: data extraction, data cleaning, data transformation and data visualization… With its growing list of packages, R can now connect with other data stores, such as MySQL, SQLite and Hadoop for data storage activities. It is the most popular language for data analysis and mining as showed the following Fig. 1.

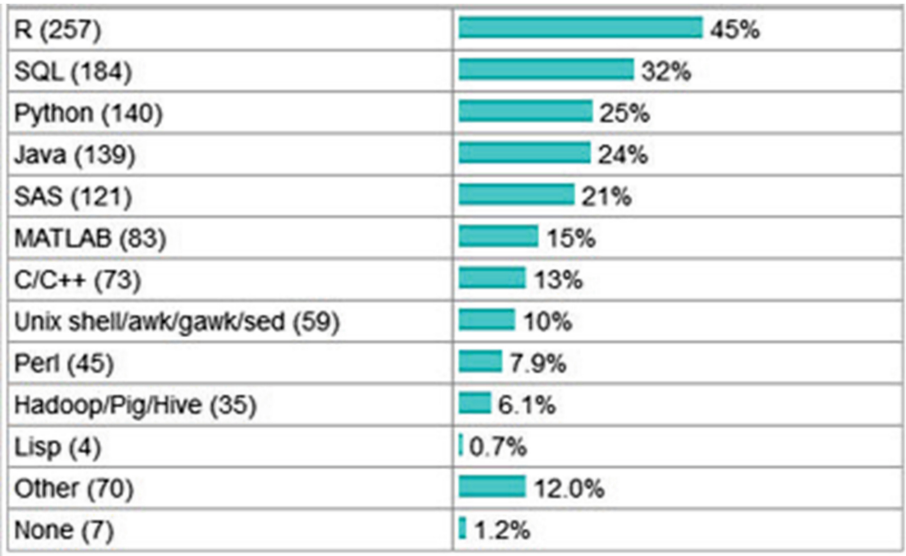

Fig. 1. Programming languages used for data mining and data analysis between 2012 and 2013

4.3 RHadoop

RHadoop is an open source project developed by Revolution Analytics. RHadoop system as shown in Fig. 2 is a result of a combination of R and Hadoop. It allows running a MapReduce jobs within R just like RHIPE (R and Hadoop Integrated Programming Environment) which is an R library which allows running a MapReduce job within R. It requires the installation of R on each data node and allows the user to carry out data analysis of big data directly in R.

We need several R packages to be installed to help it connecting R with Hadoop such as: rJava, itertools, rmr package. RHadoop is a collection of three main R packages which are rhdfs, rmr, and rhbase for providing large data operations with an R environment as shown in Fig. 3.

- *rhdfs* is an R interface for providing the connectivity to the HDFS from the R console. R users can browse, read, write, and modify files stored in HDFS from within R.

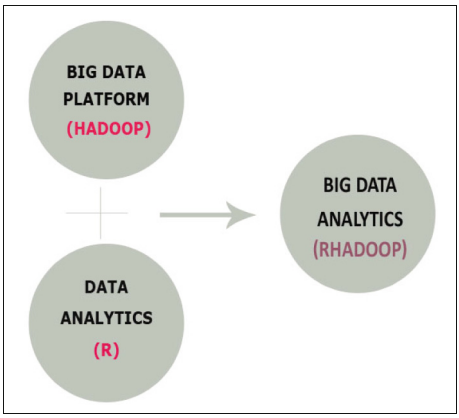

Fig. 2. Big data analytics

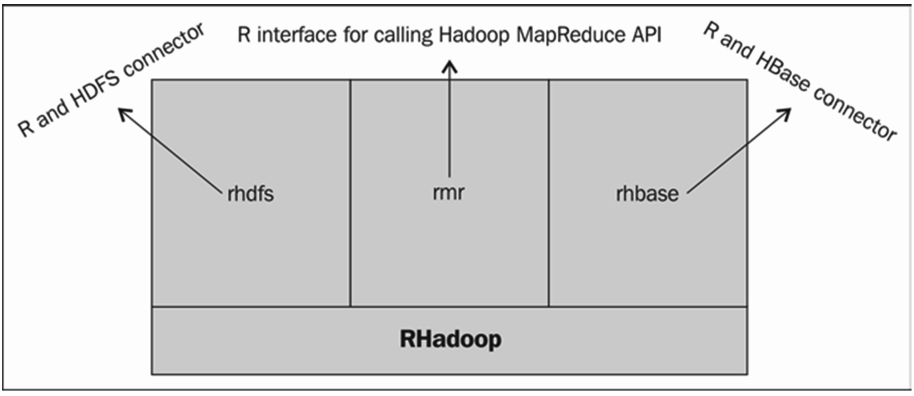

Fig. 3. RHadoop architecture

- *rmr* is an R interface for providing Hadoop MapReduce execution operations inside the R environment.
- *rhbase* is an R interface for operating the Hadoop HBase data source stored at the distributed network. The rhbase package is designed with several methods for initialization and read/write and table manipulation operations.

5 Experimentations

5.1 Data Set Description

We used a real data set of a Tunisian Bank. It contains 30,298 customers. The descriptive statistics of the four numerical variables are shown in Table 2 and Fig. 4. We started with this data set, then we duplicated it 3 times in order to increase the size of our initial banking data set.

Table 2. Descriptive statistics for banking dataset

Attribute	Min	Max	Mean	Std. dev
Age	20	97	47	12.8415
Salary	300007	14830160	1063127	133003771
Nb of incidents	0	366	99.2549	121.919
Assets	0	1466777863	4701149	203869543

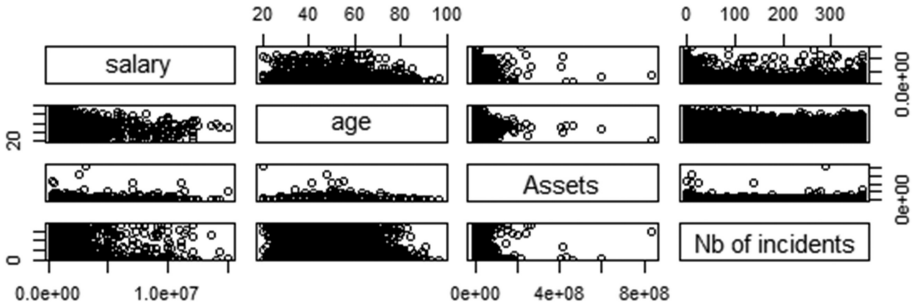

Fig. 4. Variables plot of banking data set

5.2 Time Execution Results

We have used a virtual machine (Oracle VM) with these characteristics:
RAM: 4294MO
Operating system: Red-hat (64 bits)

Table 3. Time execution of different banking data sets in seconds

Data size	212K (10000 lines)	1M (30298 lines)	3M (90891 lines)
R	1200 s	43200 s	Limited computer memory (impossible)
R + Hadoop	3 s	240 s	900 s
R + Hadoop + HBase	1,5 s	60 s	540 s

Firstable, we have applied only 10000 lines of customers from 30000 because of the limited computer memory using only R, it was not possible to enter the whole data set. That took 20 min as shown in Table 3. Due to these two problems; the inability of applying a data set of 1M and the time spent in the execution of our multi-SOM algorithm. We thought about BD and to integrate hadoop with R in order to use our data set of 30000 customers and why not more than this data and additionally to accelerate the time execution. Then, to solve these problems we integrated Hadoop with R using three different data sets as shown in Table 3. The time execution of the first data set of 212K is 3 s using R and Hadoop which is much better compared to the time execution using only R. Then, it takes 1 s and half adding Hbase tool to Hadoop

and R. The second data set of 1M of size takes 12 h using R, however with R and Hadoop it takes just four minutes and with R, Hadoop and HBase it takes one minute.

The last data set of 3M is the most largest one, it was impossible to enter it using only R because of the limited computer memory and the limitation of R to handle huge data. But, after integrating Hadoop it took 15 min and 9 min using R, Hadoop and Hbase. In this work, we have used a clustering algorithm developed in R under the package multisom in one hand and in another hand, we integrated BD technologies in order to solve existing problems of rapidity and limited memory of our computers. BD means big systems, big challenges and big profits, so more research works are necessary to resolve it. However, BD techniques and tools are very limited to solve the real existing Big Data problems. Therefore, Big Data analytics is still in the initial stage of development and it was an efficient solution to our problem.

6 Conclusion

Due to the problem of time execution spent using our banking dataset and R language, we studied the BD concept and integrated R and Hadoop to get faster results with huge banking data sets. We tested multi-SOM clustering R package on BD for the first time. So, we applied multisom R package containing 30 different indices for a real banking data set. Then, we compare the time execution of this algorithm using only R and R with Hadoop. The results show the efficiency of BD in time saving and the limitation of R language to handle huge data sets. We are fortunately witnessing the birth and development of Big Data.

As a future work, it will be interesting to apply this approach to fuzzy clustering and for bi-clustering integrating other tools of Hadoop such as Flume and HBase. Also, comparing clustering methods along with multi-SOM method using Hadoop and R.

Acknowledgement. We are gratefully thankful to Mohamed Rahal for his helpful comments and suggestions.

References

Chan, J.O.: Big data customer knowledge management. Commun. IIMA **14**(3) (2014). Article 5

Chen, C.P., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014)

Demchenko, Y., Grosso, P., De Laat, C., Membrey, P.: Addressing big data issues in scientific data infrastructure. In: International Conference on Collaboration Technologies and Systems (CTS) IEEE, pp. 48–55 (2013)

Duhon, B.: It's all in our heads. Assoc. Inf. Image Manage. Int. **12**(8), 8–13 (1998)

Douglas, L.: 3D data management: controlling data volume, velocity and variety, 6 Feb 2001

Franke, B., Plante, J.-F., Roscher, R., et al.: Statistical inference, learning and models in big data. Int. Stat. Rev. **84**(3), 371–389 (2016)

Gandomi, A., Haider, M.: Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manage. **35**(2), 137–144 (2015)

García, S., Ramírez-Gallego, S., Luengo, J., et al.: Big data preprocessing: methods and prospects. Big Data Anal. **1**, 9 (2016)

Ghouila, A., BenYahia, S., Malouche, D., Jmel, H., Laouini, D., Guerfali, Z., Abdelhak, S.: Application of multi-SOM clustering approach to macrophage gene expression analysis. Infect. Genet. Evol. **9**, 328–329 (2009)

Ihaka, R., Gentleman, R.: R: A language for data analysis and graphics. J. Comput. Graph. Stat. **5**, 299–314 (1996)

Khan, Z., Vorley, T.: Big data text analytics: an enabler of knowledge management. J. Knowl. Manage. **21**, 18–34 (2017)

Khanchouch, I., Charrad, M., Limam, M.: A comparative study of multi-SOM algorithms for determining the optimal number of clusters. Int. J. Future Comput. Commun. **4**(3), 198–202 (2014)

Khanchouch, I., Charrad, M., Limam, M.: An improved multi-SOM algorithm for determining the optimal number of clusters. In: Computer and Information Science, pp. 189–201. Springer (2015)

Kohonen, T.: Automatic formation of topological maps of patterns in a self-organizing system. In: Proceedings of the 2SCIA, Scand, Conference on Image Analysis, pp. 214–220 (1981)

Lamirel, J.C.: Using artificial neural networks for mapping of science and technology: a multi self-organizing maps approach. Scientometrics **51**, 267–292 (2001)

Lamirel, J.C.: Multisom: a multimap extension of the som model. Application to information discovery in an iconographic context, pp. 1790–1795 (2002)

Liao, Z., Yin, Q., Huang, Y., Sheng, L.: Management and application of mobile big data. Int. J. Embed. Syst. **7**(1), 63–70 (2014)

Sajana, T., Sheela Rani, C.M., Narayana, K.V.: A survey on clustering techniques for big data mining. Indian J. Sci. Technol. **9** (2016)

Shah, T., Rabhi, F., Ray, P.: Investigating an ontology-based approach for big data analysis of inter-dependent medical and oral health conditions. Cluster Comput. **18**(1), 351–367 (2015)

Sheikholeslami, G., Chatterjee, S., Zhang, A.: WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. Int. J. Very Large Data Bases (VLDB J.) **8**, 289–304 (2000)

Shen, J., Chang, S.I., Lee, E.S., Deng, Y., Brown, S.J.: Determination of cluster number in clustering microarray data. Appl. Math. Comput. 1172–1185 (2005)

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V.: Critical big data analysis challenges and analytical methods. J. Bus. Res. **70**, 263–286 (2017)

Tukey, J.W.: The Future of Data Analysis. Ann. Math. Stat. **33**, 1–67 (1962). https://doi.org/10.1214/aoms/1177704711, http://projecteuclid.org/euclid.aoms/1177704711

ur Rehman, M.H., Liew, C.S., Abbas, A., et al.: Big data reduction methods: a survey. Data Science and Engineering **l.1**, 265–284 (2016)

Wu, Y., Yuan, G.-X., Ma, K.-L.: Visualizing flow of uncertainty through analytical processes. IEEE Trans. Visual. Comput. Graph. **18**(12), 2526–2535 (2012)

Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F.: Big data and cloud computing: innovation opportunities and challenges. Int. J. Digital Earth **10**, 13–53 (2016)