International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

A Review of Big Data Clustering Methods and Research Issues

Nweso Emmanuel Nwogbaga

[1]Department of Networking and Communication, Faculty of Computer Science, University Putra Malaysia
Department of Computer Science, Faculty of Science, Ebonyi State University, Nigeria

Abstract: ***Data mining is a method for knowledge discovery from a dataset. The world today is moving toward data-driven in all ramifications, ranging from education, health care, security, customers' management, smart city, etc. Unsupervised learning like clustering is the most big-data mining technique used for grouping large dataset when there is no prior information about the classes in the dataset. The use of the internet of things (wearable, sensors, RFID) and social networks has drastically increased data in the cyber-physical world resulting in what is called Big Data. With the increase in big data as a result of cloud computing, it has proliferated research on knowledge discovery on these avalanche of big data. Clustering is used to extract valuable hidden information from massive complex data. Clustering as unsupervised learning has an advantage over supervised learning when it comes to knowledge discovery in a huge dataset without a prior knowledge of the groups. In this review, we discussed big data mining techniques and narrowed it to clustering method. We also discussed different clustering approaches, and similarities measures used in clustering algorithms. Finally, we discussed the strength and weaknesses of clustering approaches and the research issues in clustering big data for information discovery.***

Keywords: Big Data, Big Data Mining, Clustering, IoT Big Data Clustering, Distance/Similarity Measures, Unsupervised Learning

1. Introduction

The world is moving toward data-driven decision making (Provost & Fawcett, 2013). This implies basing our decisions/actions on the available data around us, for almost everything we do. There are several ways data are being generated from our environment today, ranging from sensors, cameras, Internet traffic generated stream data and other devices. These devices and platforms generate a lot of text, image, audio and video data. This different modality of data results in what is known as Big Data. The benefit of cloud computing as discussed in (Nwogbaga, 2016) also proliferated data generation, because the users careless about processing resources. Internet of things (IoT) is another source of data generation. The use of IoT in vehicular network as presented in (Eze, Sijing, Liu, Nwogbaga, & Eze, 2016; Eze, Zhang, Liu, Nwogbaga, & Eze, 2016), generates huge amount data worldwide daily. Big data imposes challenges of identifying the underlying pattern, groups or hidden information about the data set. Analytics of these big data requires efficient data mining technique posing a lot of challenges in processing and analytics (C. c. Aggarwal, 2015). The big data mining processes through these devices involves different stages (Che, Safran, & Peng, 2013). Characteristics of big data are:- high volume, different types (variety), data quality – ranging from incomplete data, noise data, etc. (veracity) and are collected at high speed (velocity) (J. Chen et al., 2013; Kuang et al., 2014).

1.1 Big Data

The term "Big Data" was first used in 1998 according to (Kitchin & McArdle, 2016; Thakur & Mann, 2014) by John Mashey in a Silicon Graphics (SGI) slide deck. Big Data is the dataset that is beyond the ability of current data processing technology (J. Chen et al., 2013; Riahi & Riahi, 2018). Big data plays a critical role in all areas of human endevour. For instance, governments are now mining the contents of social media networks, blogs, and other online transactions to identify the necessity for government facilities or to recognize the organizational groups and their activities and to predict relevant future events such as threats or promises. Service provider in the other hand track their customers' purchases made through online, in-store, and customer' behaviour through streams of online clicks for improving their marketing and predicting the growth of their profits and increase customers satisfaction (Che et al., 2013). The gap between the big data management and the capabilities of the current DBMSs can offer has reached the historically high peak. The major three characteristics (Volume, Variety and Velocity) of big data, each mean one distinct deficiencies of the present DBMSs. Large volume requires great scalability and massive parallelism that are beyond the capability of present DBMSs; high variety of data types of big data are mostly not compatible with architecture of current database systems. The velocity of big data especially stream data processing needs appropriate real-time efficiency which is far beyond the current DBMSs (Madden, 2012).

1.2 Characteristics of big data

Different authors characterize big data from different perspectives, (Hadi, Lawey, El-gorashi, & Elmirghani, 2018; Russom, 2011) characterized big data with 3Vs as Volume, Variety, and Velocity. Presently big data is a volatile term that has a different definition from different perspectives (De Mauro, Greco, & Grimaldi, 2016; Ylijoki & Porras, 2016). Others like (Mao, Hu, & Kumar, 2018) characterized big data with 4Vs as Volume, Variety, Velocity, and veracity. (Sami & Sael, 2016) Characterized big data with 9Vs as Volume, Variety, Velocity, veracity, validity, volatility, variability, visualization, and value as in table 1.

Table 1: 9Vs characteristics of big data

Characteristics	Meaning
Volume	Refers to the high volume of big data. It is predicted that the world's big data will exceed 40ZB by 2020 (Y. Zhang, Ren, Liu, Sakao, & Huisingh, 2017)
Variety	This is the data type. It can be structured, semi-structured and unstructured
Velocity	This is about the rate at which big data are been collected to be processed to get meaningful information from them
Veracity	This refers to the nature of the big data. Some are noisy, incomplete and abnormalities in the big data.
Validity	Data validity certifies that the data is valid for the purpose for which it is been used. It ensures that it is correct and accurate for intended decision making.
Volatility	Big data volatility refers to the duration the data can still be valid for the intended decision making.
Variability	This is about the variation in the velocity of the big data which may be as a result of daily, seasonal and event-triggered peak data that is unstructured. This can be challenging to manage.
Visualization	Big data visualization refers to the ways the data can be explored for human understanding, like charts or graphs.
Value	Refers to the worth of the big data. This also refers to the cost and management of the big data.

1.3 Categorization of big data characteristics

The characteristics of big data can be categorized into five as follow;- data characteristics based on data collection, data processing, data integrity, data visualization and the worth of the data (Sami & Sael, 2016) as illustrated in Figure 1.

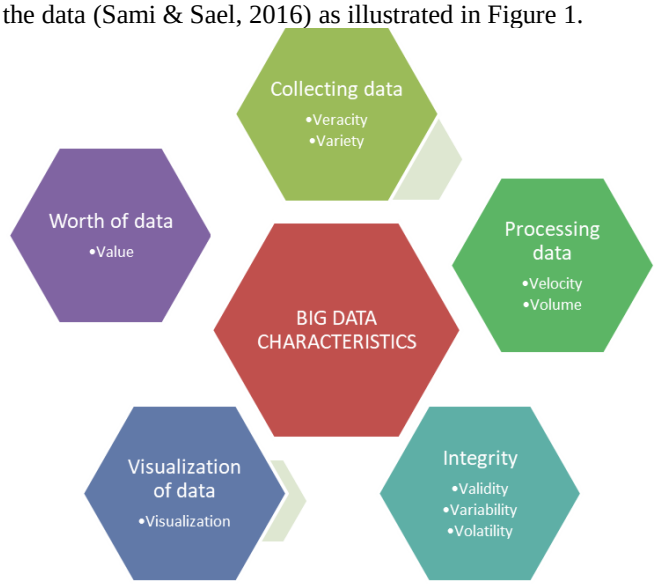

Figure 1: Categorization of 9vs characteristics of big data

1.4 Steps in Data Mining for Big Data

This refers to the steps involved in information extraction from data (object) and using the extracted information to make an informed decision. The aim of information extraction is to discover patterns from unstructured or semi-structured data. The discovered patterns can be for other users like administrators or can an input for other computer systems like search engines and database management systems to provide a better services. The information to be extracted depends on the application and type of data involved (C. C. and C. Z. Aggarwal, 2012). In this section, we present the steps for information extraction from data.

1.4.1Data acquisition: Object sensing, measuring or raw measurement of data

1.4.1.1 *Attributes (Features):* a feature is an element of a pattern characterizing an object. Pattern attributes types and value domain depend on measurements of object signals, devices used, pre-processing and coding of measurement into patterns. There are two major attributes types

- Quantitative attributes (numerical)
- Qualitative attributes (non-numerical)
- Quantitative can be divided into continues (real – value) and discrete (discrete – value) depending on the value domain.

1.4.2Data Pre-processing: In knowledge discovery (KD), a pattern is very important element. The Pattern is the totality of features used to describe an object or abstract concept. An object is recognized by the characteristic description, representing the information about the object. The term pattern is the sum of all these attributes or features used to describe the objects. If an object has (x_1, x_2 ,…, x_n) features or attributes, then the pattern of that object is given by Pattern = {x_1, x_2, …, x_n}.

1.4.3 Feature extraction: Feature extraction which is also referred to as dimensionality reduction refers to extracting the features assumed to be informative and non-redundant that will facilitate the subsequent decision making from the given dataset. The feature extraction technique to use is problem-specific (Susto, Schirru, Pampuri, & McLoone, 2016).

1.4.4Pattern form: Object's pattern can be formed using measuring devices and/or software. The devices and/or software (measurement system) senses, measures and collects the specific signals from an object. Depending on the goal for which the data are being collected and the type of object involved. The type and nature of the signal necessary for it will be measured. The pattern can be in one row or one column (vector or one dimension), matrix form (two dimensions) or multiple dimensions (n-dimensional where n is greater than 2 other wise refer to higher order tensor).

- **Pattern Vectors** (pattern in a vector space): Pattern is represented as a pattern vector. Given the elements X_i as pattern attributes such that a set $\{x_1, x_2, \ldots, x_n\}$ it implies that X = $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

The number of attributes n in the pattern vector represents the pattern space dimensionality. For instance, a pattern representing a cat can be represented as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} size \\ brightness \\ speed \end{bmatrix}$$

which implies

$$X = \begin{bmatrix} size \\ brightness \\ speed \end{bmatrix}.$$

An instance of a cat (a specific object now) can be characterized by pattern with qualitative and quantitative attribute values as $X = \begin{bmatrix} 15.8 \\ 0.85 \\ high \end{bmatrix}$ this means that the size of the cat is 15.8, and 0.85 is the brightness then the speed is high. The decision about which signals to measure, what kind and how accurate the measuring devices to use or how to form and encode an object pattern, always reflect information content and ability of object recognition. Measuring and object pattern forming are goal-oriented, such as classification (recognition) and clustering (grouping) etc.

- **Types of Objects:** there are three major groups of objects data types which include static object type, temporal object type, and spatio-temporal object type.

Static Object type: - Contains static signals of the object. This means the signal is sensed at a particular time. An example is a patient's health status, blood pressure, weight, temperature, pulse, images, and videos, etc. These types of measurements can be used to form a static pattern. For instance, an image (object) may be characterized by the image rectangular gray – values grid of pixels. In that case, the grayscale will be the basis for the image object pattern. In static objects, it is only the object attributes that will be measured. Temporal object types are data type measured at a specific time for in instance land-use patterns of Hong Kong say in 1995. In temporal object, the object attributes and time will be measured. Spatio-temporal object type is when the object is measured at particular time, and space (Erwig, Güting, Schneider, & Vazirgiannis, 199AD). In spatio-temporal objects, the attributes, time and space of the objects will be measured.

1.5 Big Data Mining Methods

There different methods or approaches used in discovering information in IoT big dataset depending on the available information about the dataset and the goal for the big data mining. There are four major methods used for big data mining which are classification, clustering, association pattern mining and outlier detection as illustrated in Figure 2.

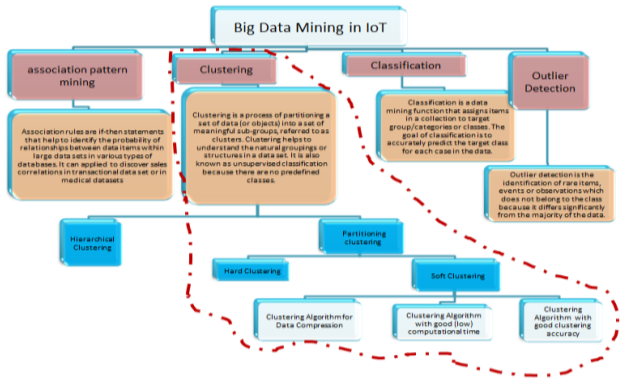

Figure 2: Big Data Mining Methods/Approaches

1.5.1 Classification

Data mining seeks to get information from a dataset. According to (C. c. Aggarwal, 2015) some data mining problem are linked to a value which is from a particular feature or attribute in the data. This feature or attribute that leads to the goal of the data mining problems is referred to as class label. Data mining problems that require class labels are referred to as supervised learning. Data used to learn the relationship between other features and the class labels are called training data. After training the model, it will be used to identify the presence of the label in the dataset. Data mining problems as above is referred to as classification. Classification groups data into predefined categories based on known characteristics of the categories (Phyu, 2009). The Data mining process uses sophisticated big data analysis techniques to discover hidden patterns and relationships in a dataset. The techniques include but are not limited to mathematical algorithms, statistical models and machine learning methods. Classification as a data mining technique can be performed using one or combination of the following techniques, decision tree induction, Bayesian Network and or k-nearest neighbor classifier (Phyu, 2009). Classification in data mining can be classified into the following group, neural networks, rule-based, memory-based learning, Bayesian network, decision tree induction and support vector machines (Kesavaraj & Sukumaran, 2013). There so many examples of classification algorithms which has been applied in so many area including in medical, information retrieval etc. For instance C4.5 classification algorithm (Quinlan, 1996; Rajesh, 2012; Ruggieri, 2002) and KNN classification algorithms (Deng, Zhu, Cheng, Zong, & Zhang, 2016; Mateos-García, García-Gutiérrez, & Riquelme-Santos, 2019).

1.5.2 Clustering

In Data mining, Clustering is a renowned technique used for knowledge discovery in different scientific areas (Carmona et al., 2012; Ci, Guizani, & Sharif, 2007; A. S. Shirkhorshidi, Aghabozorgi, & Ying Wah, 2015). Clustering algorithms are popular data mining tool used to detect inherent groups of objects that have similar characteristics (Ericson & Pallickara, 2013; Lee, Kim, Kwon, Han, & Kim, 2008). It helps to preprocess and compress big data set for processing (Hüsch, Schyska, & Bremen, 2018). Clustering has played a major role in machine learning (Anaya & Boticario, 2011; Edwards, New, & Parker, 2012; Fan, Chen, & Lee, 2008). For image analysis and image processing (Das & Konar, 2009; Oztimur Karadag & Yarman Vural, 2014; Portela, Cavalcanti, & Ren, 2014; Siang Tan & Mat Isa, 2011; Zhao, Fan, & Liu, 2014). In medical, for gene analysis (An & Doerge, 2012; de Souto, Costa, de Araujo, Ludermir, & Schliep, 2008; Ernst, Nau, & Bar-Joseph, 2005; Ma, Tavares, Jorge, & Mascarenhas, 2010), analysis of medical images or medical image segmentation (Cui, Wang, Fan, Feng, & Lei, 2013; Meyer & Chinrungrueng, 2005; Ye, Lazar, & Li, 2011), energy consumption analysis (Iglesias & Kastner, 2013; Shen, Babushkin, Aung, & Woon, 2013; van Wijk & van Selow, 1999), and environmental pollution analysis (Carbajal-Hernández, Sánchez-Fernández, Carrasco-Ochoa, & Martínez-Trinidad, 2012; Ignaccolo,

Ghigo, & Bande, 2013; Moolgavkar et al., 2013), for instruction detection and security (Francisca, 2011; Singh, Satinder, 2007; Wang & Dong, 2012; Zhuang, Ye, Chen, & Li, 2012). Clustering is one of the challenges in big data management. It refers to grouping of data from a given dataset, such that, given a data set $X = \{x_1, x_2, x_3, ..., x_n\}$ and a number of clusters (say C), then clustering means to define a mapping function such that $f : X \rightarrow \{1, ..., C\}$, Where each item for $i \in \{1, ..., n\}$ is assigned to a particular cluster $C_j, j = 1, ..., C$. A cluster C_j constitutes those items mapped to it by the function (f:). The method for determining which cluster an item should be assigned to, is usually by distance and similarity measures like Euclidean distance, tensor distance, Mahalanobi's distance, etc. (Kurasova, Marcinkevicius, Medvedev, Rapecka, & Stefanovic, 2014). Clustering is applied in a recommendation system (Ghazanfar & Prügel-Bennett, 2014; Sarwar, Karypis, Konstan, & Riedl, 2002), prediction system (Bose & Chen, 2009; Laasonen, 2005), smart city (Ghazanfar & Prügel-Bennett, 2014; Pan et al., 2013; Sarwar et al., 2002), detecting outlier (Jiang, Tseng, & Su, 2001), sensing patterns (Gardner, Hines, & Pang, 1996; Roggen, Wirz, Tröster, & Helbing, 2011) to mention but a few. It is the pre-processing involved in making the data available and accessible for analysis. It is used in big data analysis where structured and unstructured large volume data are grouped according to their similarities (Kurasova et al., 2014) Clustering is the partitioning of a dataset into different clusters based on the similarity measures such that object within the same clusters are more similar to each other than object in another clusters and object in different clusters are more different to each other than object in the same cluster (Q. Zhang et al., 2017). Clustering becomes important as the amount of data generated in our environment increases which needs to be managed to improve on the lives of people, improve the customers' relationship, government performance etc.

1.5.2.1 Clustering Procedure

- **Feature extraction and selection**: the features that mostly represent the data will be extracted from the original dataset using a defined technique.
- **Algorithm Design:** the algorithm to be used for the clustering is designed and implemented at this stage based on the type of clustering the user want.
- **Evaluation of result:** the clustering result will be evaluated and validated (validating means to judge the algorithm's result i.e. to compare the result with the original data set).
- **Interpretation of result**: the result of the algorithm will be explained using a practical example.

2. Types of Clustering Method

Clustering technique are mainly of two types (fig 3), hierarchical clustering and Partitioning clustering (Krishnasamy, Kulkarni, & Paramesran, 2014; Ng, n.d.; Raymond T. Ng, 1994). Clusters is generally referred to as a clustering which are mainly nested clustering (hierarchical) and un-nested clustering (partitioning), though clusters can as well be looked at as exclusive versus overlapping versus fuzzy, and complete versus partial (Tan, P. N., Steinbach, M., & Kumar, 2006).

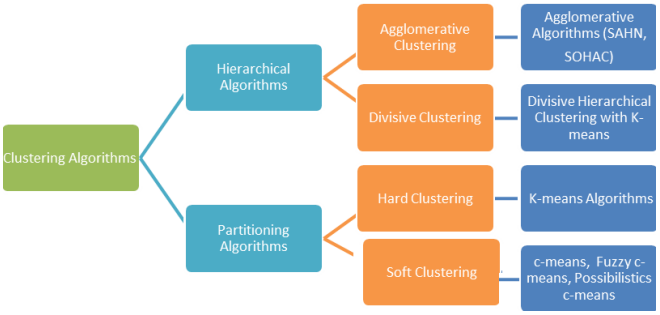

Figure 3: Types of Clustering Algorithms

2.1 Hierarchical Clustering

Hierarchical clustering, groups data in a dendrogram form. It is typically used when the number of clusters is not known (Diaz-valenzuela, Vila, & Martin-bautista, 2016). Hierarchical algorithms combine or divide existing groups by creating a hierarchical structure that reflects the order in which groups are divided or merged (Osama Abu Abbas, 2008). Hierarchical clustering can be of two types. They are;-

1. Divisive
2. Agglomerative

Divisive methods: A divisive hierarchical method is the hierarchical method that begins the clustering by putting the entire objects in one cluster and then split them into two groups at each step, based on the differences and similarities. Divisive Hierarchical Clustering with K-means (Reddy, Vivekananda, & Satish, 2017) splits a cluster into k-smaller clusters using a continuous iteration of k-means clustering until all elements has its own cluster. **Agglomerative method:** Agglomerative method starts from every item as a cluster, then merge two items at a time based on their closest similarity features until all items belong to one bigger cluster (Raymond T. Ng, 1994). Examples of hierarchical clustering include SOHAC algorithm (Buza, Nagy, & Nanopoulos, 2014), Fuzzy Hierarchical Semi-supervised (HSS) algorithm (Diaz-valenzuela et al., 2016), and Sequential, Agglomerative, Hierarchical and non-overlapping (SAHN) algorithms which uses a quantitative specification of dissimilarity between pairs of objects in the set of objects being clustered. This information can be provided in any of the two forms, stored matrix, approach where a non-negative real-value matrix is used to measure the dissimilarity between object x and y. the second approach is stored data, where each object x is described by a k-tuple $X = (x_1, x_2, ... x_k)$ of real numbers x_i being the score pertaining to the i^{th} variable or character, and the rule specifying how to calculate k-tuple, the dissimilarity between the corresponding objects (Day & Edelsbrunner, 1984).

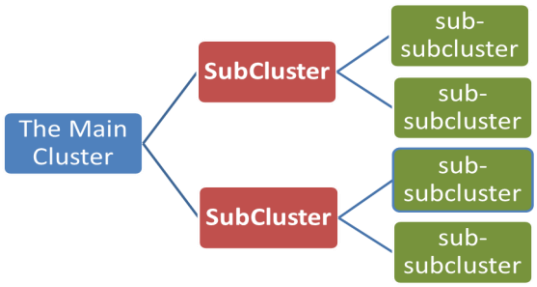

Figure 4: Divisive hierarchical clustering method

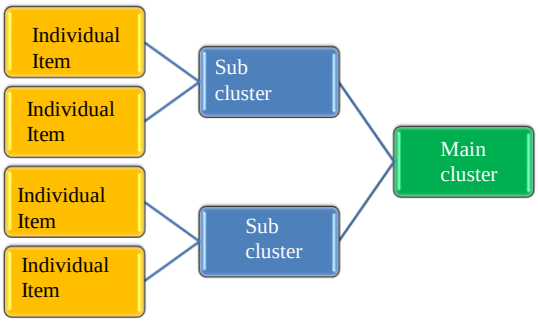

Figure 5: Agglomerative hierarchical clustering method

Strengths and weaknesses of hierarchical clustering:

Hierarchical clustering is good when the data set involved, belong to one whole family that can be split or combined based on their ranked relationship. For instance members of students from the same countries can be grouped according to their states at first level, then according to their schools at the second level, according to their level/class at the third level. Likewise, if you start from the individual students, the relationship can be combined in the reverse order as above. If the data do not have a ranked relationship, then hierarchical clustering will not give a good result. For example, clustering a data set containing members of a school into students and staff, there are some students who are also staff but are doing part-time studies and vice versa.

2.2 Partitioning Clustering Method

Partitioning clustering algorithms clusters dataset into subgroups by their similarity features which can intersect.
There are mainly two categories of partitioning clustering method according (Q. Zhang & Chen, 2014).
(1) Hard clustering and
(2) Soft clustering

2.2.1Hard Clustering Method

Partition clustering uses a membership function. It uses the membership values between zero (0) and one (1) to represent every element of the data set. Hard or soft clustering is determined by the way the membership value is calculated. In calculating the membership value which is the degree of membership of that data item to any of the clusters, hard clustering will always get a value of zero (0) or one (1). This means that it either belong to that cluster (in that case the value is one (1)) or does not belongs to the cluster (the value will be zero (0))(Q. Zhang, Yang, Chen, & Li, 2018)

2.2.2 Soft Clustering Method

In soft clustering, each of the data items may or may not completely belong to one particular cluster, which means that the membership value can take any value between 0 and 1. As the data can partly belong to one group and partly belong to another group or cluster in real sense. In grouping the data, the data will be assigned to the cluster that has the highest membership value.

2.2.3 Strengths of Partition Clustering

Partitioning clustering is always used when the data set is closely related and similar. In image and videos dataset for instance, if you are clustering the dataset into human beings and building, each image may contain building and or human beings. In this case, the group the image is to belong will be determined by the degree of either the building or human being in the image. Specifically, the problems of partition clustering are:-

Ability to identify close relationships among data

Big data is characterized by high volume, variety of data which may somehow be similar in a way. For instance, a video clip may contain human beings, cats, trees and sky. A picture may also contain sky, trees, and human beings. Grouping these into humans and trees for instances is not distinct. It therefore, depends on how the features in question, appear in the picture or video in relation to other features in consideration. Partition clustering is most appropriate for data whose similarities are fuzzy. For instance, (Q. Zhang et al., 2018) used PCM to develop CP-HOPCM algorithm which is efficient for clustering images from Flickr (Chua et al., 2009) and video clips from YouTube.

2.2.4Weaknesses of partition clustering

High computational complexity

Because the data are closely related or similar, a lot of calculations are involved to actually identify the distinct similarities and difference. It usually involves high data points and high computations as in Higher Order Possibilistic c-means Algorithm (HOPCM) (Q. Zhang, Yang, Chen, & Xia, 2015).

2.3 Parallel and Distributed Clustering of Big IoT Data

Big data has increased drastically during last decade which makes companies to proactively change the way they collect, store and analyze their big data to discover hidden and useful knowledge in order to increase their performance and customers relations. Such big data are real competitive advantage to companies and the big data is used to better respond to customers and actively follow the behavior of customers, to predict the evolution to avoid been over taken by other companies. But this big data has its own problems of storage, difficulties in analysis, processing and retrieval operation are incredibly time consuming and very difficult. Big Data analytics methods like clustering (unsupervised learning) can be done on a single system but as the volume of data increases, the clustering in parallel and distributed systems becomes paramount to ensure speedy and coherent analytics from the data (Sassi Hidri, Zoghlami, & Ben Ayed, 2017). As the amount of data increases, the single system clustering algorithms can no longer be suitable for big data analytics because of the data high complexity and computational requirement. The clustering algorithm can be

deployed to work with multiple systems to reduce the computational time. Hence, the need for distributed algorithms that can run on multiple systems to address big data challenges. Generally, big data clustering methods can be grouped as single-machine clustering and multiple-machine clustering. When the data involved are too large, multiple-machine clustering is preferred. Usually single-machine clustering are based on dimensionality reduction technique while multiple-machine clustering is based on parallel clustering (A. S. S. A. T. Y. W. and T. H. Shirkhorshidi, 2014).

3. Clustering Distances/Similarities Measures

Distance or similarity measure is a very important step in clustering. It is used to determine how two elements or data can be assigned to their groups (C. C. Chen & Chu, 2005).

The most common distance/similarity measures are Minkowski family. Others include Intersection family, L1 family, Squared L2 family or χ2 family, Shannon's entropy family(Cha, 2007).

Minkowski Distance: Minkowski distance is defined as

$$d_{\min} = (\sum_{i=1}^{n} \| x_i - y_i \|^m)^{1/m}$$

m > 1. The most used distance measures for clustering are Euclidean and Manhattan distances which are versions of Minkowski depending on the value of m. if m = 2 it is Euclidean $d_{\min} = (\sum_{i=1}^{n} \| x_i - y_i \|^2)^{1/2}$ while when m = 1 it is Manhattan distance $d_{\min} = \sum_{i=1}^{n} \| x_i - y_i \|$ x_i and y_i are the vectors or tensors of n attributes.

4. Comparison of existing clustering algorithms

In an attempt to use the most relevant features of a data set for clustering process to avoid some redundant features which introduces a lot of challenges in clustering activities such as increase in run time, clustering accuracy drop and high memory usage during clustering, so many feature reduction technique has been proposed by different researcher to overcome these challenges, but each of them also comes with its own limitations. Here, we present some feature extraction/reduction techniques used for clustering and their limitations.

4.1 CP-HOPCM Algorithm with HOPCM Algorithm

In CPD, given $X_j \in R^{I_1 \times I_2 \times \ldots \times I_N}$ for $J = 1, 2, \ldots, n$ canonic Polyadic decomposition will decompose X_j into $\sum_{r=1}^{R} a_{jr}^{(1)} o\, a_{jr}^{(2)} o \ldots o\, a_{jr}^{(N)}$

$$= [\left[A_j^{(1)}, A_j^{(2)}, \ldots, A_j^{(N)}\right]]$$

Taken $I = \{I_1, I_2, \ldots, I_N\}$ then each sample has the storage complexity of O(NIR) in CPD format instead of O(I^N) in the original tensor format. At the beginning of CPD, $\prod_n^N I^{I_n}$ elements are loaded into the memory, but only $R\sum_n^N I_n$ are stored for each sample in the CPD format. Therefore the number of element in CPD grows linearly in CPD with regard to N instead of exponentially with regard to N (Q. Zhang et al., 2018). However, this method though reduces the number of attribute used for clustering; it courses a high clustering accuracy drop in the clustering algorithms at low R values. If R values should be increased to increase the accuracy, the attribute reduction rate of CPD will drastically drop. If R value is increased above a certain level, CPD will increase the number of original attributes instead of reducing it, thereby making it inefficient for certain values of data dimensionalities. The results of different values of CPD with different I_n is shown on Table 2 and the fig 6 show the attribute reduction for $R^{512x256x3}$ at R = 4, 8, 16, 32 and 64. The CPD value zero (0) means that the produced attributes are equal to the original number of attribute. Any values less than 0 means that at such dimensions of the input data, CPD will increase the number of attribute instead of reducing. The clustering accuracy, computational time and memory usage of CPD when implemented with High Order Possibilistic c-means are presented in Table 3. The results shows that the higher the value of R the lower the attribute reduction and the higher the accuracy but higher computational time and memory usage.

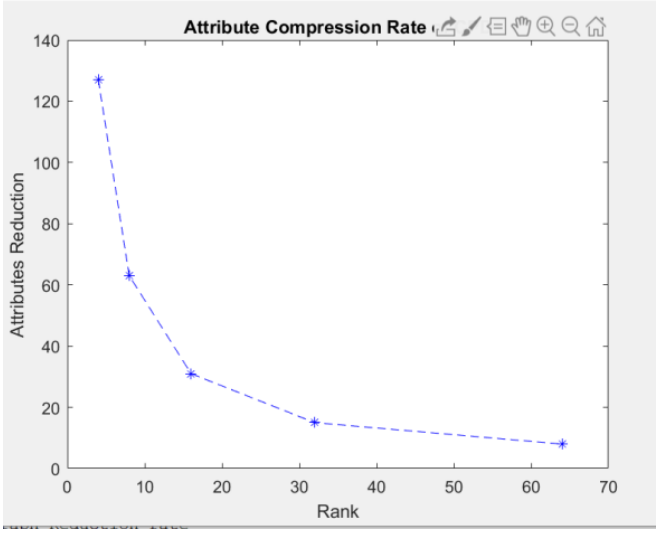

Figure 6: CPD Attribute Reduction Rate

4.2 Tensor Train Decomposition (TT) HOPCM Algorithm with HOPCM Algorithm

Tensor Train Decomposition is similar to CPD because it also decomposes with rank (R) values. Like CPD, TT also reduces clustering accuracy at low value of R though better than CPD as shown in Table 3, when implemented with

HOPCM (Q. Zhang et al., 2018). But higher computational time and memory usage as R increases to improve on the accuracy.

Table 2: Attribute Reduction of Canonical Polyadic Decomposition at Rank (R)

S/N	Image Height	Image Width	Colour space	Total Attributes	Attribute Reduction at Rank				
					4	8	16	32	64
1	128	32	96	393216	384	192	96	48	24
2	128	1024	3	393216	85.11168831	42.55584	21.27792	10.63896	5.319480519
3	512	256	3	393216	127.5019455	63.75097	31.87549	15.93774	7.968871595
4	64	64	3	12288	23.45038168	11.72519	5.862595	2.931298	1.465648855
5	32	64	3	6144	15.51515152	7.757576	3.878788	1.939394	0.96969697
6	32	32	3	3072	11.46268657	5.731343	2.865672	1.432836	0.71641791
7	12	12	3	432	4	2	1	0.5	0.25
8	6	6	3	108	1.8	0.9	0.45	0.225	0.1125

4.3 Efficient Fuzzy c-means algorithm with Standard Fuzzy c-means algorithm

The author presented Efficient Fuzzy c-means approach based on tensor canonical Polyadic decomposition scheme for big data clustering. The result of data compression based on tensor canonical Polyadic decomposition of the dataset is used for clustering using efficient fuzzy c-means and the result compared with standard fuzzy c-means as shown in table 3. Based on table 3, efficient fuzzy c-means approach with different ranks R as compared with standard fuzzy c-means regarding the execution time increases as the rank R increases from 4 to 64. The running time of efficient fuzzy c-means grows gradually. This is because as the rank increases the compression rate of the attributes reduces which the weaknesses of CPD for feature reduction are. At R = 16, the running time of efficient fuzzy c-means is 460 minutes while that of the standard fuzzy c-means is 611 minutes. This implies that if the rank should be increased further to achieve a better clustering accuracy, the execution time will be higher than the execution time for standard fuzzy c-means. The clustering accuracy of efficient fuzzy c-means and standard fuzzy c-means were also compared using E_* and Adjusted

Rand Index (ARI) metrics. Table 3, shows that as R increases from 4 to 64, the clustering accuracy of efficient fuzzy c-means improves gradually regarding E_* and ARI. At a smaller value of R, the accuracy drops. At R = 64 the efficient fuzzy c-means produces E_* = 13.22 and ARI = 0.89 while standard fuzzy c-means produced 12.98 and 0.89 for E_* and ARI respectively. The performance of efficient fuzzy c-means as compared to standard fuzzy c-means was also evaluated using sWSN dataset as shown in table 3. The increase in R values increases the execution time of the efficient fuzzy c-means. For instance as R increases from 8 to 32, the execution time increases from 462 minutes to 643 minutes. As the rank increases to 64 the execution tome of efficient fuzzy c-means and standard fuzzy c-means are about the same with the execution time of 660 minutes and 700 minutes for efficient fuzzy c-means and standard fuzzy c-means respectively. The comparison of efficient fuzzy c-means and standard fuzzy c-means in terms of E_* and ARI shows that the efficient fuzzy c-means achieved a better accuracy on sWSN dataset over standard fuzzy c-means. As R increases, the accuracy of efficient fuzzy c-means increases for instance, when R increases from 4 to 64, the efficient fuzzy c-means accuracy based on E_* improves from 1.26 to 0.61 while ARI value increased from 0.73 to 0.90 as against E_* of 0.59 and ARI of 0.91 for standard fuzzy c-means. Though the efficient fuzzy c-means fuzzy c-means at lower R reduces the computational time, it causes high clustering accuracy drop on the other hand. However, efficient fuzzy c-means reduces execution time at lower rank but it has high clustering accuracy drop. As the rank increases to improve the accuracy, the time also increases. With the increase in the rank above 64, the execution time will increase more than that of the standard fuzzy c-means.

4.4 HOK-Means algorithm with HOPCM algorithm:

This is a three stacked auto-encoder, each with three hidden layers proposed by (Bu, 2017). It is deep learning technique for feature learning. Feature learning is very important for heterogeneous data clustering because if plays major role on the clustering accuracy. Deep learning technique has advantage over other feature extraction technique because of its ability to extract multiple levels of features from objects, by stacking some basic machine learning models. (Bu, 2017) Presented three dropout stacked auto-encoders, each auto-encoder with three hidden layer to learn the features for each modality of the object. The feature vectors from the result of the stacked auto-encoder are used for clustering using High Order K-means (HOK-means) algorithm. To evaluate the performance of the auto-encoder, HOK-means was implemented and evaluated using NUS-WIDE and CUAVE dataset and compared with HOPCM algorithm using two metrics Error of the clustering center (E_*) and Rand Index (RI). According the results as presented in Table 3, the HOK-means produced better clustering accuracy compared with HOPCM. However, five different results produced by HOK-means in the five experiments conducted using 80, 000 images from NU-WIDE dataset indicate that the method is inconsistent. For instance, from the experimental result of KOK-means, the second experiment produced E_* of 2.76 as against 2.90 in the first experiment and about 2.8 in the third experiment which shows inconsistency in the result produced.

Table 3: Comparison of different algorithms with different attribute reduction techniques

Title	Author	Year	Feature Extraction used	Clustering algorithm	DATASET 1				DATASET 2			
					Metrics 1	Metrics 2	Metrics 3	Metrics 4	Metrics 1	Metrics 2	Metrics 3	Metrics 4

					(accuracy clustering centre) Using Clustering Error (E∗)	(accuracy clustering data items) Using Adjusted Rank Index (ARI)	(execution time in minutes)	(memory usage in percentage)	(accuracy clustering centre) Using Clustering Error (E∗)	(accuracy clustering data items) Using Adjusted Rank Index (ARI)	(execution time in minutes)	(memory usage in percentage)
High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT	Zhang, Qingchen Yang, Laurence T. Chen, Zhikui Li, Peng	2018	Canonical Polyadic Decomposition (CPD)	Canonical Polyadic – High Order Possibilistic c-means algorithms compared with HOPCM algorithm	**NUS-WUDE-14 dataset**				**SNAE2 dataset**			
					CP-HOPCM	**2.71**	**0.90**	**399**	**47.7**	**7.957**	**0.851**	**≅530**	**≅36**
					HOPCM	2.72	0.91	275	55.5	7.944	0.856	≅390	≅43
			Tensor Train Decomposition (TT)	Tensor Train – High Order Possibilistic c-means algorithms compared with HOPCM algorithm	**TT - HOPCM**	2.70	0.92	399	50.5	7.936	0.869	≅550	≅37
					HOPCM	2.72	0.91	275	55.5	7.944	0.856	≅390	≅43
An efficient fuzzy c-means approach based on canonical polyadic decomposition for clustering big data in IoT	Bu, Fanyu	2018	Canonical Polyadic Decomposition (CPD)	Efficient Fuzzy c-means compared with standard Fuzzy c-means algorithm	**eGSAE dataset**				**sWSN dataset**			
					Efficient FCM	13.22	0.89	≅580	-	0.61	0.90	660	-
					Standard FCM	12.98	0.89	611	-	0.59	0.91	700	-
A High-Order Clustering Algorithm Based on Dropout Deep Learning for Heterogeneous Data in Cyber-Physical-Social Systems	Bu, Fanyu	2017	Dropout Deep Learning Model (this is a three stacked auto-encoder , each with three hidden layers)	HOK-Means algorithm compared with HOPCM	**NUS-WIDE dataset**				**CUAVE dataset**			
					HOK – means	2.76	0.902	261	-	-	0.912	111	-
					HOPCM	3.01	0.844	295	-	-	0.864	127	-

4.5 Research Issues in IoT Big Data Clustering

Challenges in big data are mostly to identify the natural patterns (hidden information) in the data. Clustering is one of the techniques used in identifying these natural groupings in a dataset. Clustering basically has its own challenges regarding clustering IOT big data as follows as show in Table 3.

1) The clustering accuracy of existing algorithms is affected by so many things such as the number of attributes required; attribute reduction method, method of similarity check and so on. If the accuracy is not good enough, the essence of clustering the dataset will not be achieved as some hidden information in the dataset will not be identified.
2) The computational complexity of the existing algorithms is another problem of clustering that needs serious research attention. If the algorithm is taking too long, it will not be efficient to be used in a situation that requires urgent attention such as medical, instruction detection and so on. So the computational time of the algorithm affects the response rate which makes it inefficient for real-time IoT big data analytics.
3) Compression rate of algorithms makes the existing algorithms to require high memory space during the execution time. This is another big challenge of the clustering algorithms. Big data usually involves a large volume of data with large attributes, using the data just the way they are, usually will cost much memory space during processing and will as well be time-consuming. To come up with a good attribute reduction method that will not affect hidden information of the dataset is still a challenge in big data mining.

5. Conclusion

We are in the era of enormous data called big data, resulting from human daily activities. These data are heterogeneous, structured, semi-structured and unstructured data which are constantly generated at unprecedented scale. The present data mining methods are no longer suitable for the present big data in our cyber-social systems. With the emergence of big data, the limitations of the present data mining techniques are uncovered which resulted to the challenges of IoT big data. Despite the few work done on big data, there are need for more research on big data to overcome the challenges related to accuracy of big data clustering algorithms, computational time of the algorithm (speed), data compression/attribute reduction rate of clustering algorithms, heterogeneity, and scalability. In this paper we discussed big data, big data mining approaches, then delved into clustering technique, types of clustering methods, strengths and weaknesses, compared some clustering techniques and identified their weaknesses. Finally we presented three main research issues based on clustering accuracy, speed of clustering algorithm and data compression/attribute reduction rate problems in clustering.

References

[1] Aggarwal, C. c. (2015). Data Mining. In *Springer*. https://doi.org/10.1007/978-3-319-14142-8
[2] Aggarwal, C. C. and C. Z. (2012). *Mining Text Data*. Springer Science+Business Media, LLC 2012.
[3] An, L., & Doerge, R. W. (2012). Dynamic Clustering of Gene Expression. *ISRN Bioinformatics*, *2012*, 1–12. https://doi.org/10.5402/2012/537217

[4] Anaya, A. R., & Boticario, J. G. (2011). Application of machine learning techniques to analyse student interactions and improve the collaboration process. *Expert Systems with Applications*, *38*(2), 1171–1181. https://doi.org/10.1016/j.eswa.2010.05.010

[5] Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce*, *19*(2), 133–151. https://doi.org/10.1080/10919390902821291

[6] Bu, F. (2017). *A High-Order Clustering Algorithm Based on Dropout Deep Learning for Heterogeneous Data in Cyber-Physical-Social Systems*. *6*, 11687–11693.

[7] Buza, K., Nagy, G. I., & Nanopoulos, A. (2014). Storage-optimizing clustering algorithms for high-dimensional tick data. *Expert Systems with Applications*, *41*(9), 4148–4157. https://doi.org/10.1016/j.eswa.2013.12.046

[8] Carbajal-Hernández, J. J., Sánchez-Fernández, L. P., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2012). Assessment and prediction of air quality using fuzzy logic and autoregressive models. *Atmospheric Environment*, *60*, 37–50. https://doi.org/10.1016/j.atmosenv.2012.06.004

[9] Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., Del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. *Expert Systems with Applications*, *39*(12), 11243–11249. https://doi.org/10.1016/j.eswa.2012.03.046

[10] Cha, S. (2007). Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*(4), 300–307. https://doi.org/10.1007/s00167-009-0884-z

[11] Che, D., Safran, M., & Peng, Z. (2013). From big data to big data mining: Challenges, issues, and opportunities. *International Conference on Database Systems for Advanced Applications. Springer, Berlin, Heidelberg.*, *7827 LNCS*, 1–15. https://doi.org/10.1007/978-3-642-40270-8_1

[12] Chen, C. C., & Chu, H. T. (2005). Similarity measurement between images. *Proceedings - International Computer Software and Applications Conference*, *2*, 41–42. https://doi.org/10.1109/COMPSAC.2005.140

[13] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: A data management perspective. *Frontiers of Computer Science*, *7*(2), 157–164. https://doi.org/10.1007/s11704-013-3903-7

[14] Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: A real-world web image database from National University of Singapore. *CIVR 2009 - Proceedings of the ACM International Conference on Image and Video Retrieval*, (4), 368–375. https://doi.org/10.1145/1646396.1646452

[15] Ci, S., Guizani, M., & Sharif, H. (2007). Adaptive clustering in wireless sensor networks by mining sensor energy data. *Computer Communications*, *30*(14–15), 2968–2975. https://doi.org/10.1016/j.comcom.2007.05.027

[16] Cui, W., Wang, Y., Fan, Y., Feng, Y., & Lei, T. (2013). Localized FCM clustering with spatial information for medical image segmentation and bias field estimation. *International Journal of Biomedical Imaging*, *2013*. https://doi.org/10.1155/2013/930301

[17] Das, S., & Konar, A. (2009). Automatic image pixel clustering with an improved differential evolution. *Applied Soft Computing Journal*, *9*(1), 226–236. https://doi.org/10.1016/j.asoc.2007.12.008

[18] Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*(1), 7–24. https://doi.org/10.1007/BF01890115

[19] De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061

[20] de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics*, *9*, 1–14. https://doi.org/10.1186/1471-2105-9-497

[21] Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, 143–148. https://doi.org/10.1016/j.neucom.2015.08.112

[22] Diaz-valenzuela, I., Vila, M. A., & Martin-bautista, M. J. (2016). *Short Paper On the Use of Fuzzy Constraints in Semisupervised Clustering*. *24*(4), 992–999.

[23] Edwards, R. E., New, J., & Parker, L. E. (2012). Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, *49*, 591–603. https://doi.org/10.1016/j.enbuild.2012.03.010

[24] Ericson, K., & Pallickara, S. (2013). On the performance of high dimensional data clustering and classification algorithms. *Future Generation Computer Systems*, *29*(4), 1024–1034. https://doi.org/10.1016/j.future.2012.05.026

[25] Ernst, J., Nau, G. J., & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, *21*(SUPPL. 1), 159–168. https://doi.org/10.1093/bioinformatics/bti1022

[26] Erwig, M., Güting, R. H., Schneider, M., & Vazirgiannis, M. (199AD). Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, *3*(3), 269–296. https://doi.org/10.1023/A:1009805532638

[27] Eze, E. C., Sijing, Z., Liu, E., Nwogbaga, N. E., & Eze, J. C. (2016). Timely and Reliable Packets Delivery over Internet of Vehicles (IoVs) for Road Accidents Prevention : A Cross-Layer Approach. *IET Networks*, *5*(5), 127–135.

[28] Eze, E. C., Zhang, S., Liu, E., Nwogbaga, N. E., & Eze, J. C. (2016). RECMAC: Reliable and efficient cooperative cross-layer MAC scheme for vehicular communication based on random network coding technique. *2016 22nd International Conference on Automation and Computing, ICAC 2016: Tackling the New Challenges in Automation and Computing*,

(September), 342–347. https://doi.org/10.1109/IConAC.2016.7604943
[29] Fan, S., Chen, L., & Lee, W. J. (2008). Machine learning based switching model for electricity load forecasting. *Energy Conversion and Management*, *49*(6), 1331–1344. https://doi.org/10.1016/j.enconman.2008.01.008
[30] Francisca, N. O. (2011). Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology*, *6*(3), 314–325.
[31] Gardner, J. W., Hines, E. L., & Pang, C. (1996). Detection of vapours and odours from a multisensor array using pattern recognition: Self-organising adaptive resonance techniques. *Measurement and Control*, *29*(6), 172–177. https://doi.org/10.1177/002029409602900603
[32] Ghazanfar, M. A., & Prügel-Bennett, A. (2014). Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications*, *41*(7), 3261–3275. https://doi.org/10.1016/j.eswa.2013.11.010
[33] Hadi, M. S., Lawey, A. Q., El-gorashi, T. E. H., & Elmirghani, J. M. H. (2018). Big data analytics for wireless and wired network design : A survey. *Computer Networks*, *132*, 180–199. https://doi.org/10.1016/j.comnet.2018.01.016
[34] Hüsch, M., Schyska, B. U., & Bremen, L. Von. (2018). CorClustST — Correlation-based clustering of big spatio-temporal datasets. *Future Generation Computer Systems*, (2007). https://doi.org/10.1016/j.future.2018.04.002
[35] Iglesias, F., & Kastner, W. (2013). Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. *Energies*, *6*(2), 579–597. https://doi.org/10.3390/en6020579
[36] Ignaccolo, R., Ghigo, S., & Bande, S. (2013). Functional zoning for air quality. *Environmental and Ecological Statistics*, *20*(1), 109–127. https://doi.org/10.1007/s10651-012-0210-7
[37] Jiang, M. F., Tseng, S. S., & Su, C. M. (2001). Two-phasee clustering process for outliers detection. *Pattern Recognition Letters*, *22*(6–7), 691–700. https://doi.org/10.1016/S0167-8655(00)00131-8
[38] Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013*, 1–7. https://doi.org/10.1109/ICCCNT.2013.6726842
[39] Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 205395171663113. https://doi.org/10.1177/2053951716631130
[40] Krishnasamy, G., Kulkarni, A. J., & Paramesran, R. (2014). A hybrid approach for data clustering based on modified cohort intelligence and K-means. *Expert Systems with Applications*, *41*(13), 6009–6016. https://doi.org/10.1016/j.eswa.2014.03.021
[41] Kuang, L., Hao, F., Yang, L. T., Lin, M., Luo, C., & Min, G. (2014). A tensor-based approach for big data representation and dimensionality reduction. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 280–291. https://doi.org/10.1109/TETC.2014.2330516
[42] Kurasova, O., Marcinkevicius, V., Medvedev, V., Rapecka, A., & Stefanovic, P. (2014). Strategies for Big Data Clustering. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2014-Decem*, 740–747. https://doi.org/10.1109/ICTAI.2014.115
[43] Laasonen, K. (2005). Clustering and prediction of mobile user routes from cellular data. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3721 LNAI*, 569–576. https://doi.org/10.1007/11564126_59
[44] Lee, K., Kim, J., Kwon, K. H., Han, Y., & Kim, S. (2008). DDoS attack detection method using cluster analysis. *Expert Systems with Applications*, *34*(3), 1659–1665. https://doi.org/10.1016/j.eswa.2007.01.040
[45] Ma, Z., Tavares, J. M. R. S., Jorge, R. N., & Mascarenhas, T. (2010). A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. *Computer Methods in Biomechanics and Biomedical Engineering*, *13*(2), 235–246. https://doi.org/10.1080/10255840903131878
[46] Madden, S. (2012). From databases to big data. *IEEE Internet Computing*, *16*(3), 4–6. https://doi.org/10.1109/MIC.2012.50
[47] Mao, Q., Hu, F., & Kumar, S. (2018). Simulation methodology and performance analysis of network coding based transport protocol in wireless big data networks. *Simulation Modelling Practice and Theory*, *84*, 38–49. https://doi.org/10.1016/j.simpat.2018.01.005
[48] Mateos-García, D., García-Gutiérrez, J., & Riquelme-Santos, J. C. (2019). On the evolutionary weighting of neighbours and features in the k-nearest neighbour rule. *Neurocomputing*, *326–327*, 54–60. https://doi.org/10.1016/j.neucom.2016.08.159
[49] Meyer, F. G., & Chinrungrueng, J. (2005). Spatiotemporal clustering of fMRI time series in the spectral domain. *Medical Image Analysis*, *9*(1), 51–68. https://doi.org/10.1016/j.media.2004.07.002
[50] Moolgavkar, S. H., Mcclellan, R. O., Dewanji, A., Turim, J., Georg Luebeck, E., & Edwards, M. (2013). Time-series analyses of air pollution and mortality in the United States: A subsampling approach. *Environmental Health Perspectives*, *121*(1), 73–78. https://doi.org/10.1289/ehp.1104507
[51] Ng, R. T. (n.d.). *E cient and E ective Clustering Methods for Spatial Data Mining 1 Introduction*. 1–25.
[52] Nwogbaga, N. E. (2016). Critical Analysis Of Cloud Computing And Its Advantages Over Other Computing Techniques. *Journal of Multidisciplinary Engineering Science and Technology*, *3*(2), 3955–3960.
[53] Osama Abu Abbas. (2008). Comparisons Between Data Clustering Algorithms. *International Arab Journal of Information Technology*, *5*(3), 320–325.
[54] Oztimur Karadag, O., & Yarman Vural, F. T. (2014). Image segmentation by fusion of low level and domain specific information via Markov Random Fields. *Pattern Recognition Letters*, *46*, 75–82. https://doi.org/10.1016/j.patrec.2014.05.010
[55] Pan, G., Qi, G., Zhang, W., Li, S., Wu, Z., & Yang, L. (2013). Trace analysis and mining for smart cities: Issues, methods, and applications. *IEEE*

Communications Magazine, *51*(6), 120–126. https://doi.org/10.1109/MCOM.2013.6525604

[56] Phyu, T. N. (2009). Text Classification and Classifiers:A Survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85–99. https://doi.org/10.5121/ijaia.2012.3208

[57] Portela, N. M., Cavalcanti, G. D. C., & Ren, T. I. (2014). Semi-supervised clustering for MR brain image segmentation. *Expert Systems with Applications*, *41*(4 PART 1), 1492–1497. https://doi.org/10.1016/j.eswa.2013.08.046

[58] Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

[59] Quinlan, J. R. (1996). Improved Use of Continuous Attributes in C4.5. *Journal of Arti?Cial Intelligence Research*, *4*, 77–90. https://doi.org/10.1613/jair.279

[60] Rajesh, K. and S. A. (2012). Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(2), 72–77. Retrieved from www.ijarcce.com

[61] Raymond T. Ng, & J. H. (1994). *Efficient and Effective Clustering Methods for Spatial Data Mining*. 144–155.

[62] Reddy, M. V., Vivekananda, M., & Satish, R. U. V. N. (2017). Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering. *International Journal of Computer Science Trends and Technology*, *5*(Sep-Oct), 6–11.

[63] Riahi, Y., & Riahi, S. (2018). Big Data and Big Data Analytics: concepts, types and technologies. *International Journal of Research and Engineering*, *5*(9), 524–528. https://doi.org/10.21276/ijre.2018.5.9.5

[64] Roggen, D., Wirz, M., Tröster, G., & Helbing, D. (2011). Recognition of crowd behavior from mobile sensors with pattern analysis and graph clustering methods. *Networks and Heterogeneous Media*, *6*(3), 521–544. https://doi.org/10.3934/nhm.2011.6.521

[65] Ruggieri, S. (2002). Efficient C4. 5. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *14*(2), 438–444. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstract%5Cnhttp://ukpmc.ac.uk/abstract/CIT/325293

[66] Russom, P. (2011). Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, *19*(4), 1–34.

[67] Sami, S., & Sael, N. (2016). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. *International Journal of Advanced Computer Science and Applications*, *7*(3), 254–258. https://doi.org/10.14569/ijacsa.2016.070337

[68] Sarwar, B. M., Karypis, G., Konstan, J., & Riedl, J. (2002). Recommender Systems for Large-scale E-Commerce : Scalable Neighborhood Formation Using Clustering. *Communications*, *50*(12), 158–167. https://doi.org/10.1.1.4.6985

[69] Sassi Hidri, M., Zoghlami, M. A., & Ben Ayed, R. (2017). Speeding up the large-scale consensus fuzzy clustering for handling Big Data. *Fuzzy Sets and Systems*, *1*, 1–25. https://doi.org/10.1016/j.fss.2017.11.003

[70] Shen, W., Babushkin, V., Aung, Z., & Woon, W. L. (2013). An ensemble model for day-ahead electricity demand time series forecasting. *E-Energy 2013 - Proceedings of the 4th ACM International Conference on Future Energy Systems*, 51–62. https://doi.org/10.1145/2487166.2487173

[71] Shirkhorshidi, A. S., Aghabozorgi, S., & Ying Wah, T. (2015). A Comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0144059

[72] Shirkhorshidi, A. S. S. A. T. Y. W. and T. H. (2014). Big Data Clustering: A Review. *International Conference on Computational Science and Its Applications. Springer, Cham*, *8583 LNCS*(PART 5). https://doi.org/10.1007/978-3-319-09156-3

[73] Siang Tan, K., & Mat Isa, N. A. (2011). Color image segmentation using histogram thresholding Fuzzy C-means hybrid approach. *Pattern Recognition*, *44*(1), 1–15. https://doi.org/10.1016/j.patcog.2010.07.013

[74] Singh, Satinder, and G. K. (2007). Unsupervised anomaly detection in network intrusion detection using clusters. *Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007) RIMT-IET, Mandi Gobindgarh. 2007*, *38*, 333–342.

[75] Susto, G. A., Schirru, A., Pampuri, S., & McLoone, S. (2016). Supervised Aggregative Feature Extraction for Big Data Time Series Regression. *IEEE Transactions on Industrial Informatics*, *12*(3), 1243–1252. https://doi.org/10.1109/TII.2015.2496231

[76] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Cluster analysis: Basic Concepts and Algorithms. *Introduction to Data Mining*, (8), 487–568. https://doi.org/10.1109/IPTA.2008.4743793

[77] Thakur, B., & Mann, M. (2014). Data Mining for Big Data : A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(5), 469–473.

[78] van Wijk, J. J., & van Selow, E. R. (1999). Cluster and calendar based visualization of time series data. *Proceedings of the IEEE Symposium on Information Visualization*, 4–9. https://doi.org/10.1109/infvis.1999.801851

[79] Wang, B., & Dong, A. (2012). Online clustering and outlier detection. *Meta-Heuristics Optimization Algorithms in Engineering, Business, Economics, and Finance*, 529–545. https://doi.org/10.4018/978-1-4666-2086-5.ch017

[80] Ye, J., Lazar, N. A., & Li, Y. (2011). Sparse geostatistical analysis in clustering fMRI time series. *Journal of Neuroscience Methods*, *199*(2), 336–345. https://doi.org/10.1016/j.jneumeth.2011.05.016

[81] Ylijoki, O., & Porras, J. (2016). Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management*, *4*(1), 69–91. https://doi.org/10.24840/2183-0606_004.001_0006

[82] Zhang, Q., & Chen, Z. (2014). *A weighted kernel possibilistic c -means algorithm based on cloud*

computing for clustering big data. (September), 1378–1391. https://doi.org/10.1002/dac

[83] Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018). High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. *Information Fusion*, *39*, 72–80. https://doi.org/10.1016/j.inffus.2017.04.002

[84] Zhang, Q., Yang, L. T., Chen, Z., & Xia, F. (2015). A High-Order Possibilistic-Means Algorithm for Clustering Incomplete Multimedia Data. *IEEE Systems Journal*, *PP*(99), 1–10. https://doi.org/10.1109/JSYST.2015.2423499

[85] Zhang, Q., Zhu, C., Yang, L. T., Chen, Z., Zhao, L., & Li, P. (2017). An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things. *IEEE Transactions on Industrial Informatics*, *13*(3), 1193–1201. https://doi.org/10.1109/TII.2017.2684807

[86] Zhang, Y., Ren, S., Liu, Y., Sakao, T., & Huisingh, D. (2017). A framework for Big Data driven product lifecycle management. *Journal of Cleaner Production*, Vol. 159, pp. 229–240. https://doi.org/10.1016/j.jclepro.2017.04.172

[87] Zhao, F., Fan, J., & Liu, H. (2014). Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial information for image segmentation. *Expert Systems with Applications*, *41*(9), 4083–4093. https://doi.org/10.1016/j.eswa.2014.01.003

[88] Zhuang, W., Ye, Y., Chen, Y., & Li, T. (2012). Ensemble clustering for internet security applications. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(6), 1784–1796. https://doi.org/10.1109/TSMCC.2012.2222025

Author Profile

Nweso Emmanuel Nwogbaga received the B.Sc. degree in Computer Science from Ebonyi State University, Abakaliki, Nigeria, in 2005 and the M.Sc. degree in 2012 also from Computer Science from Ebonyi State University, Abakaliki, Nigeria. Presently this author is a PhD candidate at University Putra Malaysia, Malaysia. From 2012 to 2013, he was assistant lecturer in Computer Science Department, Ebonyi State University, Abakaliki, Nigeria. From 2013 to 2016 he was lecturer II and after his promotion, he is now Lecturer I in the same department from 2016 till date. Mr. Nwogbaga is member Computer Professionals (Registration Council of Nigeria).